Engineering and Science Mathematics 4A: Probability, Statistics and Markov Processes

Patrick D. Bangert
School of Engineering and Science
International University Bremen
p.bangert@iu-bremen.de

February 20, 2004

Contents

Acknowledgements

Most of the notes were written by me but some parts were borrowed. I thank the following people and sources for kindly lending material included in these notes:

John Lammin SRS Varadhan James W. Pitman Michael Casey Numerical Recipes in C Nanochron, LLC; 4201 Wilson Blvd., # 110-615; Arlington, VA 22203; USA. Utts, Jessica M. (1999): Seeing Through Statistics. Duxbury Press, Pacific Grove. Haigh, John (2002): Probability Models. Springer Verlag.

Preface

Probability and Statistics are topics of mathematics and of practical science. The basic concepts are very difficult as can be seen by the fact that there is still a deep controversy about a certain few concepts and how to use them. As such, much statistics that is reported is simply wrong; some of it intentionally.

As the subject is not completely agreed upon, this course will try to present all views on the topic and let the reader decide. However, a certain bias cannot be overcome. The main goal of this course is the understanding of the basic terms and concepts of probability and statistics. The concepts at issue are: Probability, Sample, Population, Representative, Significant, Correlation and Test. We will learn how to evaluate probabilities and combine them, what a population and a sample are and what makes the sample representative. Then we will learn how to test whether certain properties are correlated and hypotheses true or false and how significant the result of this test is. Finally we shall learn about how these concepts can be combined to draw useful conclusions about reality and also predict (by use of Markov Processes) future events.

There are 28 lectures, each having a chapter in this compilation. The preface is the chapter corresponding to the first lecture. Most lectures have a reading associated with them that must be read before that lecture takes place. The reading is a selection from published works that illustrates the topic at hand. Sometimes the selections are rather old and thus show more clearly what was done. Certain prejudices are apparent in these works and their writing style makes them amusing to read. They are included here because of their scientific and historical value and quality of writing as the purpose of the course is the understanding of a few very difficult concepts. The opinions stated in these writings are not to be understood to correlate at all with those of the author of this compilation.

The grading policy, homework and projects are explained in the last two chapters.

Part I

Probability

Chapter 1

The Notion of Probability

1.1 Introduction

We are now going to study the mathematics of random events. If we throw a six faced die we do not know in advance which of the numbers 1, 2, 3, 4, 5, 6 will show. If we spin a coin we do not know in advance whether it will show a head or a tail. In ordinary speech we say that it is all a matter of chance. It is possible to argue that no humanly constructible die or coin can be perfect and must show some sort of bias. This might only reveal itself after the results of thousands of trials had been subjected to statistical analysis but nevertheless it would be there. These problems are real; even attempts to lay down a precise definition of what we mean by a random sequence of numbers run into some very subtle difficulties. The so called random number generators used in computer programs are in fact anything but random; they produce repeating sequences. The sequences admittedly have very long cycles and do have some of the properties that we would like a truly random sequence to have. You can use them safely for most problems.

But we cannot spend time on these considerations interesting though they may be. As usual we will construct a mathematical model assuming that all is for the best in the best of all possible worlds. And we will then find that it does serve to solve many ”real” world problems in a satisfactory way.

But first we need some definitions.

Definition 1 *A* random event *(sometimes called a* trial*) is an event whose actual outcome is determined by chance.*

Definition 2 *The set of all possible outcomes is called the* sample space *for the event and is usually denoted by the letter S. Sometimes the sample space is referred to as the* possibility space *or* outcome space. *The term sample space is widely used in statistics. I avoid outcome space as we have already used it when we were talking about permutations and combinations. Although the sample spaces of probability theory nearly always correspond to an outcome space as previously defined I prefer to keep the nomenclature distinct.*

Thus for throwing the usual six faced die $S = \{1, 2, 3, 4, 5, 6\}$. The events of the sample space are often called *elementary events* and the word *event* itself is reserved for any non empty subset of S. Suppose the event $A = \{2, 4, 6\}$. Then we say that event A has *happened* if we get a 2, a 4 or a 6 on throwing the die.

This mathematical definition of the notion of an event is important in what follows. Try to get away from the everyday life meaning of the word which is not precise enough for us. I cannot emphasize strongly enough that an event $A = \{a, b, c, \cdots\}$ is a non empty subset of the sample space S and we say that the event A has happened if we get any one from the elementary events of A as the result of a particular trial.

We will use $n(A)$ as before to denote the number of elements in the set A. It is not the number of times the event happens in any experiment. One final note before we get down to actual problems. The sample spaces and events that we study in this chapter will all be finite sets. We will lift this restriction in a later chapter.

Consider the result of throwing a die 120 times. The results might come out like this:

Result	1	2	3	4	5	6
Frequency	18	21	20	17	24	20
Relative Frequency	$\frac{18}{120}$	$\frac{21}{120}$	$\frac{20}{120}$	$\frac{17}{120}$	$\frac{24}{120}$	$\frac{20}{120}$

The terms *result* and *frequency* used in this example are self explanatory. The term *relative frequency* is used by statisticians for the ratio of the number of times a particular event happens to the total number of trials.

We notice:

1. The relative frequencies are all approximately equal (to 1/6).

2. The sum of the relative frequencies is $120/120 = 1$. There is nothing approximate about this result.

If some enthusiast were to throw the die 1200 or even 12000 times we would expect the approximation of the relative frequencies to 1/6 to get better and better. The sum of the relative frequencies would always be 1 of course.

From considerations like these two definitions of probability can be extracted. They are called the *a priori* and *a posteriori* definitions. “Before” and “after” in English.

1.1.1 The *A Priori* Definition.

Given a sample space S composed of elementary events (e.g. $S = \{1, 2, 3, 4, 5, 6\}$ in the case of throwing a die). If there are n elementary events in S denoted by $e_1, e_2, e_3, \cdots e_n$ we assign positive fractions $p_1, p_2, p_3, \cdots p_n$ to $e_1, e_2, e_3, \cdots e_n$ respectively in such a way that $p_1 + p_2 + p_3 + \cdots + p_n = 1$.

You will think that this leaves us far too much freedom. From the example quoted above you will see that setting the $p_1, p_2, p_3, \cdots\cdots p_n$ to be positive fractions adding up to 1 is not a bad idea but you will object that this still leaves us an infinity of ways to choose the p’s.

So it does, and from any one of them we could go on and construct a probability theory. (Incidentally in certain areas of pure mathematics we do just that and it leads to a very important topic known as measure theory. If the results also correspond to probability theory the measure is called a probability measure). However in our case there is the additional restriction that we expect our assignments of values to help us solve probabilistic problems.

In actual problems then our next step is to find a method of assigning the p's in a way which both satisfies the conditions above and also makes sense in our problem.

If we decide that the elementary events are equiprobable as with the six results obtainable by throwing an unbiased die it is reasonable to take

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}. \tag{1.1}$$

This corresponds to the naive notion that a given result happens on average once in every 6 times.

If the events are not equiprobable we can still make a try at defining the probabilities in many cases. Given a fair die with one face marked 1, two faces marked 2 and 3 faces marked 6 $S = \{1, 2, 6\}$ and we would assign 1/6, 2/6, 3/6 as the corresponding probabilities.

In a difficult situation we can attempt to verify a probability assignment as follows. Given $S = \{e_1, e_2, \cdots e_n\}$. Suppose we assign probabilities $p_1, p_2, \cdots p_n$ which are positive fractions with $\sum_{i=1}^{n} p_i = 1$. We now do experiments involving a large number of trials, N say. Suppose e_1 occurs n_1 times, e_2 occurs n_2 times, $\cdots$ and e_n occurs n_n times. If our chosen p's are to make sense the relative frequencies $\frac{n_1}{N}, \frac{n_2}{N} \cdots \frac{n_n}{N}$ which certainly add up to 1 should also become very close to the p's that we have defined. It is neat and very tempting to say that we require $\lim_{N\to\infty} \frac{n_i}{N} = p_i$. The trouble with this is that we can never carry out an experiment in which $N \to \infty$. So this is not on the same footing as say $\lim_{n\to\infty} \frac{n}{n+1} = 1$ of pure math where we have techniques for calculating limits without having to carry out an (impossible) infinite number of operations.

You will now appreciate why this entire scheme is called the a priori definition of probabilities. You assign the probabilities first and verify them later (if at all). This was the method used by the early workers in this field who were developing the theory for the study of gambling games. Great use is made of this method today, but we must also consider the a posteriori definition.

1.1.2 The *A Posteriori* Definition

Suppose a manufacture of dog food wants to gather information about a new product. The company buys 100 dogs from an animal research center and administers the product to each of them. They find the following results:

1. e_1 : fed dog begs for more $n_1 = 5$.
2. e_2 : fed dog has a fit $n_2 = 10$.
3. e_3 : fed dog dies within the hour $n_3 = 85$.

and so assign the following probabilities.

$$p_1 = 5/100, \quad p_2 = 10/100, \quad p_3 = 85/100. \tag{1.2}$$

This is called the a posteriori definition because you do the experiment first and then assign the probabilities according to the results of the experiment. The devout experimental physicists among you will be lauding this eminently proper way of going about things. But it too has its problems. For example, faced with drawing up an advertising campaign to sell the product on the strength of these results the director screams that the dogs were not "fair" or "unbiased" dogs. Coming from an animal research center they probably wanted to die. He may well have a point there and certainly raises the question of how you choose your sample if you are going to establish probabilities by experimentation. In the 1960's a drug called Thalidomide was put on the market as a tranquilizer. It had of course been extensively tested on humans before it was allowed to be sold. Unfortunately the sampling had not included any pregnant women. As a result thousands of terribly deformed babies were born before the reason was found and the drug withdrawn.

The validity of sampling techniques is a very important topic in statistics. For our present task of developing the first steps of probability theory it does not matter how the probabilities have been found. Let us summarize what we really need:

Definition 3 *Given a sample space S containing a finite number of elementary events $e_1, e_2, \cdots e_n$. A positive number p_i is associated with each event e_i in such a way that*

$$\sum_{i=1}^{n} p_i = 1 \tag{1.3}$$

Then $p_1, p_2, \cdots p_i, \cdots p_n$ is called a discrete probability distribution*.*

It is called discrete because there are only a finite number of probabilities. We can now define the probability of an event $E \subseteq S$ in terms of this probability distribution.

Definition 4 *Given a sample space $S = \{e_1, e_2, \cdots e_n\}$ and a probability distribution $p_1, p_2, \cdots p_n$ for S. Suppose $E \subseteq S$ is the set $E = \{e_{r_1}, e_{r_2}, \cdots e_{r_s}\}$ where each e_{r_j} is one of the e_i and no e_i appears twice in E so that $s \leq n$. The probability of the event E happening on a random trial is denoted by $P(E|S)$ and is given by*

$$P(E|S) = p_{r_1} + p_{r_2} + \cdots + p_{r_s} \tag{1.4}$$

This definition looks complicated but it is very precise. Study it carefully. It will become crystal clear as we look at examples. Remember that an event E happens when a random trial produces a result which is in the set E.

Example 1 *Given a fair die what is the probability of throwing an even number?*

Solution.

S=\{1,2,3,4,5,6\}. $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$. $E = \{2,4,6\}$.

$$P(E|S) = p_2 + p_4 + p_6 = \frac{3}{6} = \frac{1}{2}. \tag{1.5}$$

So far as this example is concerned then our definition seems eminently reasonable.

Example 2 *If a card is drawn from a well shuffled pack what is the probability of its being an ace?*

Solution.

$S = \{$ *any one of the 52 cards* $\}$. $p_1 = p_2 = ... = p_{52} = 1/52$.

$E = \{$ *any one of the four aces* $\}$.

$$P(E|S) = \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} = \frac{1}{13}. \tag{1.6}$$

Example 3 *What is the probability of securing a hand of 13 cards all of the same suit in a game of bridge?*

Solution.

$S = \{$ *all possible different bridge hands* $\}$.

Now S has C_{13}^{52} *elements. Each one of these hands is equally likely. So the probability distribution is* $p_i = \frac{1}{C_{13}^{52}}$ *for each hand.*

$E = \{$ *the four hands containing just one suit* $\}$

$$P(E|S) = \frac{1}{C_{13}^{52}} \approx 1.6 \times 10^{-11}. \tag{1.7}$$

Example 4 *The letters of the word moon are arranged at random. What is the probability that the O's will be separated?*

Solution.

There are $\frac{{}^4P_4}{2!} = \frac{4.3.2}{2!} = 12$ *arrangements altogether. If the O's are kept together there are 6 arrangements. Hence there are also 6 arrangements in which the O's are separated.*

$S = \{$ *the 12 possible arrangements* $\}$. $p_1 = p_2 = ... = p_{12} = 1/12$

$E = \{$ *the 6 arrangements with separated O's* $\}$

$$P(E|S) = 6.\frac{1}{12} = \frac{1}{2}. \tag{1.8}$$

Example 5 *Two dice are thrown and their sums added. List the sample space and find the probability that the total score is a) 7 b)* ≥ 9.

Solution.

$S = \{2,3,4,5,6,7,8,9,10,11,12\}$. *But these are not equiprobable!*

Each die can fall in 6 ways so there are 36 different possible outcomes.

We get 2 as 1 1 1 way

3 as 1 2 2 1 2

4 1 3 3 1 2 2 3

5 1 4 4 1 2 3 3 2 4

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

7 1 6 6 1 2 5 5 2 3 4 4 3 6

8 2 6 6 2 3 5 5 3 4 4 5

9 3 6 6 3 4 5 5 4 4

10 4 6 6 4 5 5 3

11 5 6 6 5 2

12 6 6 1

Total 36 ways.

So we assign our probability distribution as:

$$\begin{array}{llllll} P(2) = \frac{1}{36} & P(3) = \frac{2}{36} & P(4) = \frac{3}{36} & P(5) = \frac{4}{36} & P(6) = \frac{5}{36} & P(7) = \frac{6}{36} \\ P(8) = \frac{5}{36} & P(9) = \frac{4}{36} & P(10) = \frac{3}{36} & P(11) = \frac{2}{36} & P(12) = \frac{1}{36} & \end{array} \tag{1.9}$$

For a) we just have an elementary event and P(7) = 1/6.

For b) E = {9,10,11,12}.

$$P(E\,|S) = \frac{4+3+2+1}{36} = \frac{5}{18}. \tag{1.10}$$

Note.

The notation P(1) = 1/36 etc. that we have used here is a useful alternative to suffices in practical calculations. We can put a detailed description of the event inside the brackets if necessary.

Example 6 *360 oysters are examined and 12 contain pearls. What is the probability that the next oyster examined contains a pearl?*

Solution.

This is an a posteriori problem. The sample space has only two elements:

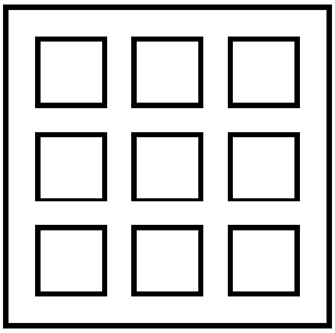
S = { oyster contains a pearl, oyster does not contain a pearl }

On the evidence given we assign :

P(oyster contains a pearl) = 12/360

P(oyster does not contain a pearl) = 338/360.

So we claim that the probability that the next oyster examined contains a pearl is 1/30. This is the best estimate we can make with the evidence we have.



The diagram shows a target made from a white square of side 20 cm. on which is painted an array of 9 red squares each of side 2 cm. A dart is thrown at random at the target. Find the probability that it hits a red square.

Example 7 *Solution.*

S = { dart hits a red square, dart hits a white area }

Since the total area is 400 cm^2 *and the red area is 9 x 4 = 36 cm*^{2} *it is reasonable to assign the probabilities:*

P(dart hits a red square) = $\frac{36}{400}$

P(dart hits a white space) = $\frac{364}{400}$

Thus the probability that it hits a red square is simply $\frac{9}{100}$.

1.1.3 Exercises

1. Make a list of all possible combinations of Boy(B) and Girl(G) for four child families. Assuming that P(B) = P(G), find the probability that in a four child family chosen at random there is (a) at least one girl, (b) at least two girls, (c) the oldest and youngest are of the same sex, (d) no girl in the family is older than a boy.

2. Four fair dice are thrown and the total score calculated. What is the probability that the total score is 23 or more?

3. What is the probability that an arrangement of the letters of the word MARROW has the R's together?

4. Three rows of cars wait to board a hover craft. The 10th car in each row is labelled A,B,C respectively. Assuming that cars are drawn at random from the front of each row until all cars are loaded find the chance that (a) A is loaded before B, (b) Both A and B are loaded before C.

5. A dart board is divided into 20 equal sectors and marked with scores from 1 to 20. If a dart hits the scoring part of the board at random, calculate the probability (a) that the score is a prime number, (b) the score will not divide by 3.

6. An urn contains equal large numbers of red and green balls. If three balls are drawn out, what is the probability that both colors are represented?

7. A biased die was thrown 100 times and gave the following results:

   Score 1 2 3 4 5 6

   Number 17 21 15 10 21 16 Total 100

   Make the best possible estimate that the sum of the scores of the next two throws will be at least 4.

8. Twelve people have their names written on pieces of paper, each of which is folded and placed in a hat. Among the twelve people there are three brothers. What is the probability that if 5 names are drawn out not all the brothers will be included?

9. Four players A,B,C and D play a game of chance three times. Find the probability that (a) A wins all the games, (b) A wins the second game only.

10. In a box there are 12 bulbs of which 3 are defective. If 2 are taken at random, what is the chance that both are defective? What is the chance that neither is defective?

11. Two dice are thrown. What is the probability that the scores differ by 2 or less?

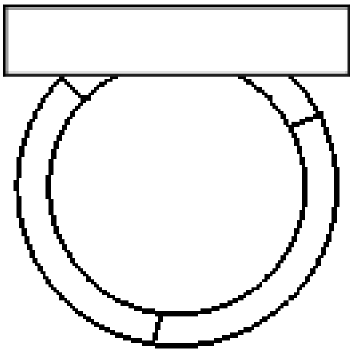
12. Trains at the same platform at Baker Street underground station go alternately on the Circle line and the Metropolitan line. A traveller who always wishes to take a Metropolitan line train, and who can be assumed to arrive on the platform at random, finds that 403 times out of 500 the next train due is a circle line train. How can this be reconciled with the fact that the trains alternate?

13. A prism has ends which are equilateral triangles and sides which are rectangular. An experiment with 1000 throws gave 882 occasions when it landed on its rectangular faces and 118 when it landed on a triangular end. The triangular ends are painted red and green and the rectangular faces red, green, and blue. Find the best estimate of the probability that for a single throw on to a flat surface (a) only red and green can be seen (b) both red faces can be seen.

14. What is the probability that if the letters of the word MANAGEMENT are arranged in random order, then the vowels will be separated.

15. An integer is chosen at random. What is the probability that it will be divisible by 2, but not by 4 or 6?

16. A white square of side 30 cm. has a black circle of radius 10 cm drawn at its center. Find the probability that a point chosen at random within the square also lies within the circle.



A cartwheel rim is made of three equal pieces joined together as shown in the diagram. The inner radius of the wheel is 74 cm and the height of the horizontal beam is 66 cm above the centre of the wheel What is the probability that when the cart stops no part of one of the joins can be seen?

17.

1.2 Reading: The Meaning of Probability by Ernest Nagel

1.2.1 Introduction to the Reading

1.2.2 The Paper

Chapter 2

Compound Probabilities

2.1 Probability of Compound Events

Up to now we have considered problems in which we have a sample space S of elementary events and a probability distribution defined on S. Given any event $E \subseteq S$ we have developed a reliable method of calculating $P\left(E|S\right)$, the probability that in a random trial we will obtain one of the elementary events belonging to E. However given a sample space S and a probability distribution defined on it we must now find out how to deal with compound events. A compound event involves two or more events E_1, E_2, $\cdots$ combined in some way. There are several ways in which this can be done and we will define the cases precisely as we come to them. We will look at a case theoretically first and then apply it.

We are first going to study $P\left(E_1 \cup E_2|S\right)$. Recall that $E_1 \cup E_2$ means the union of E_1 and E_2, i.e. the set made up of the elements of both E_1 and E_2.

$$\begin{array}{ll} [\text{ Recall: } E_1 = \{a,b,c,d\} & E_2 = \{c,d,e,f\} \\ E_1 \cup E_2 = \{a,b,c,d,e,f\} & E_1 \cap E_2 = \{c,d\}\ ] \end{array} \tag{2.1}$$

Theorem 1

$$P\left(E_1 \cup E_2\left|S\right.\right) = P\left(E_1\left|S\right.\right) + P\left(E_2\left|S\right.\right) - P\left(E_1 \cap E_2\left|S\right.\right). \tag{2.2}$$

Proof 1 *From our basic definition $P\left(E_1 \cup E_2|S\right)$ is the sum of the probabilities of the elementary events in $E_1 \cup E_2$. $P\left(E_1|S\right) + P\left(E_2|S\right)$ includes the values for any common elements of E_1 and E_2 twice. But these common elements are precisely the elements of $E_1 \cap E_2$. Hence $P\left(E_1 \cup E_2|S\right) = P\left(E_1|S\right) + P\left(E_2|S\right) - P\left(E_1 \cap E_2|S\right)$.*

Example 8 *Two dice are thrown. What is the probability of scoring either a double or a sum greater than 9?*

Solution.

We use the sample space S of Example 5.

$$\begin{array}{ll} E_1 = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\} & P\left(E_1|S\right) = \frac{6}{36} \\ E_2 = \{(4,6),(6,4),(5,5),(5,6),(6,5),(6,6)\} & P\left(E_2|S\right) = \frac{6}{36} \\ E_1 \cap E_2 = \{(5,5),(6,6)\} & P\left(E_1 \cap E_2|S\right) = \frac{2}{36} \end{array} \tag{2.3}$$

$$P\left(E_1 \cup E_2|S\right) = \frac{6}{36} + \frac{6}{36} - \frac{2}{36} = \frac{5}{18} \tag{2.4}$$

$E_1 \cup E_2$ translates to "$E_1 \cup E_2$ happens if the trial yields a result in either E_1 or E_2." We must by now be convinced of the need for the underlying sample space S. So from now on we will omit it in the notation when only one sample space is involved in a problem and no confusion can arise. Thus the result of Theorem 1 can be shortened to:

$$P\left(E_1 \cup E_2\right) = P\left(E_1\right) + P\left(E_2\right) - P\left(E_1 \cap E_2\right) \tag{2.5}$$

But don't forget the full notation. We shall need it again soon.

Definition 5 *Two events E_1 and E_2 are said to be* mutually exclusive *if $E_1 \cap E_2 = \emptyset$, the empty set.*

This is much the same as the use we made of this term when working with permutations, i.e. if two events are mutually exclusive an elementary event of E_1 cannot also be an elementary event of E_2 and vice versa. But we are now working in a more precisely defined framework.

If $E_1 \cap E_2 = \emptyset$ clearly $P\left(E_1 \cap E_2\right) = 0$ as there are no elementary events to sum over. Thus for mutually exclusive events Theorem 1 becomes

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) \tag{2.6}$$

Recall next that if we are given an event set E the *complement* of the event set denoted by E' is the set made up of all those elements of S which are not in E, i.e. if $E = \{a, d, f\}$ and $S = \{a, b, c, d, e, f, g\}$ then $E' = \{b, c, e, g\}$. We always have $E \cap E' = \emptyset$ and $E \cup E' = S$. Note that some books use E^c as their notation for the complement of E.

Theorem 2 *If E is any event $P\left(E'\right) = 1 - P\left(E\right)$.*

Proof 2

$$\begin{aligned} P\left(E \cup E'\right) &= P\left(E\right) + P\left(E'\right) - P\left(E \cup E'\right) && (2.7)\\ \Rightarrow P\left(S\right) &= P\left(E\right) + P\left(E'\right) && (2.8)\\ & && (2.9)\end{aligned}$$

But $P(S)$ involves summing over all the elementary events of the sample space and hence by the definition of a probability distribution $P(S) = 1$. $1 = P(E) + P(E')$and the theorem follows at once.

Example 9 *Two dice are thrown. What is the probability of not getting a double?*

<u>*Solution.*</u>

Let E be the event set $E = \{$ all doubles $\}$ $P(E) = \frac{6}{36} = \frac{1}{6}$.Then E' is the set of all throws which are not doubles.

$$P(E') = 1 - \frac{1}{6} = \frac{5}{6} \tag{2.10}$$

You will recognize a device which we often used in permutation and combination problems: the reverse problem is often much simpler.

We now consider a further aspect of probability theory. If you study the weather statistics for several years in Manila you will be able to calculate the (a posteriori) probability that it will rain on a day chosen at random. But if you also happen to know that the day chosen falls in the middle of the rainy season then your probability will be much higher. Probability pure and simple involves only random trials, but in applying probability theory to the "real world" we will often have partial knowledge either of the event that we are working with or of an event which can influence this. Having said this we will once again take refuge in symbols and return to the "real world" later.

We have the usual sample space S and a probability distribution defined on it. Suppose that E and F are two event subsets (and that neither E nor F is empty to avoid triviality). We want to know the probability of E given that we do know that F actually occurs.

A little thought will show you that F has in fact become our sample space. If the element of E whose probability we are looking for is not in F there is now absolutely no chance of its happening and it can be considered deleted from the sample space. Making a trial for an elementary event of E is doomed to failure if that event is not in F. So in our full notation what we must find is $P\left(E|F\right)$.

Before we rush off to the canteen to celebrate this discovery there is a remaining problem. The original probabilities defined on S will not give a probability distribution on E! Since some of the original events are missing the remaining probabilities cannot add up to 1 and that, as we all know, is an essential prerequisite for any probability distribution.

Let us consider a situation. Given: $S = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9, e_{10}\}$ with the probability distribution $p_1, p_2, \cdots p_{10}$ with $p_1 + p_2 + \cdots + p_{10} = 1$. Let $F = \{e_2, e_4, e_9\}$ so that we know that $P\left(F|S\right) = p_2 + p_4 + p_9$.

We make the hypothesis that the probability distribution for F as the new sample space is:

$$e_2 : \quad \frac{p_2}{p_2 + p_4 + p_9} \qquad e_4 : \quad \frac{p_4}{p_2 + p_4 + p_9} \qquad e_9 : \quad \frac{p_9}{p_2 + p_4 + p_9} \tag{2.11}$$

These probabilities clearly add up to 1 and theoretically satisfy the requirements for a probability distribution. They are also plausible since they multiply each of the old probabilities for e_2, e_4, e_9 by the same factor, which is greater than 1, suggesting that the probabilities have all been increased in the same ratio by our extra knowledge that one of the events of F must occur. However this remains a hypothesis. In situations where it can be experimentally tested it yields results in agreement with the experimental results. Hence it is adopted for use all the time. [The computer experts among you might like to try to design a computer simulation which would test this hypothesis.] We summarize all this in the next definition.

Definition 6 *If an event F has occurred originating from a sample space S and we wish to regard F as the new sample space, we obtain the new probability distribution of the elementary events in F by multiplying their probabilities in S by the factor $\frac{1}{P(F|S)}$. In some books these new probabilities are called* conditional probabilities *(since they only apply on condition that a previous event has already occurred) and problems involving them are called problems in conditional probability.*

From the definition we have at once the extremely important relation

$$P(E|F) = \frac{P(E \cap F|S)}{P(F|S)} \tag{2.12}$$

(i.e. we add the S probabilities of those elements of E which are also in F and divide the result by $P(F|S)$).

Do not forget the underlying meaning of $P(E|F)$. We are only considering F as the new sample space because we know that the events of F are the only ones in question.

Equation 2.12 is of vital importance in much of our more advanced work in probability and statistics. In the derivation of equation 2.12 we had to assign probabilities to the new sample space F by using the multiplying factor $\frac{1}{P(F|S)}$. This is of theoretical importance but in most practical applications of equation 2.12 it will not be necessary to actually descend to this level of detail. Also, in practical problems we shall very often be using equation 2.12 in the form

$$P(E \cap F|S) = P(E|F)\,P(F|S) \tag{2.13}$$

Study our next example carefully and you will realize that this is much simpler in practice than it may seem from the general discussion.

Example 10 *A bag contains 20 balls, 10 red, 8 white and 2 blue. The balls are indistinguishable apart from the color. Two balls are drawn in succession without replacement. What is the probability that they will both be red?*

Solution.

F is "a ball is drawn from the original 20 and is red." E is "a ball is drawn from the remaining 19 and is red." $E \cap F$ is "both balls drawn are red." Clearly the problem wants us to find $P(E \cap F|S)$. We have $P(F|S) = \frac{10}{20} = \frac{1}{2}$. $P(E|F)$ is the probability of getting a red ball from 9 red balls remaining in a bag of 19 balls i.e. $P(E|F) = \frac{9}{19}$. So using equation 2.13 $P(E \cap F|S) = \frac{1}{2} \cdot \frac{9}{19} = \frac{9}{38}$.

Example 11 *With the data of example 10 what is the probability of obtaining a blue and a white ball in either order?*

Solution.

Changing notation let:

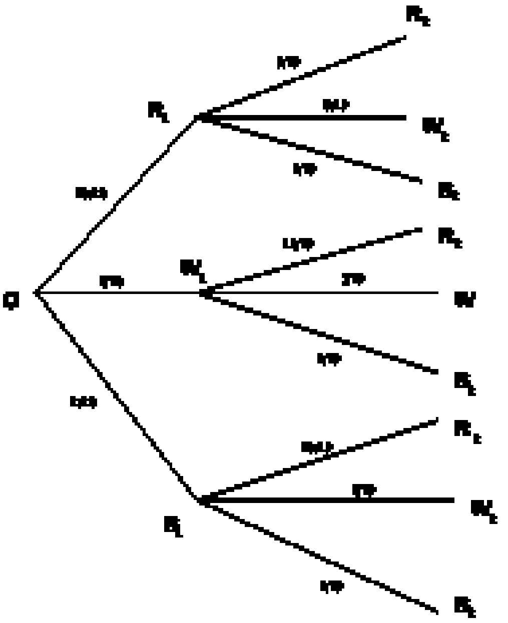
B_1 be "1st ball drawn is blue." B_2 be "2nd ball drawn is blue." W_1 be "1st ball drawn is white." W_2 be "2nd ball drawn is white".

Since the order does not matter we require $P(W_2 \cap B_1|S) + P(B_2 \cap W_1|S)$.

$$P(W_2 \cap B_1|S) = P(W_2|B_1)\,P(B_1|S) = \frac{8}{19} \cdot \frac{2}{20} = \frac{4}{95} \tag{2.14}$$

$$P(B_2 \cap W_1|S) = P(B_2|W_1)\,P(W_1|S) = \frac{2}{19} \cdot \frac{8}{20} = \frac{4}{95} \tag{2.15}$$

Hence the required probability is 8/95.



The argument used in these last two examples can be illustrated by what is called a tree diagram. In the diagram below we have added $R1$ and $R2$ to the notation already in use. The meanings need no further clarification.

The probability of getting to any "leaf" on the tree starting from O is obtained by multiplying the probabilities of the paths followed. e.g. the probability of getting two blue balls is $\frac{2}{20} \cdot \frac{1}{19} = \frac{1}{190}$ etc.

Finally for this section we discuss the related problem of independent events. Two events are independent if neither has any effect on the probability of the other. The ball drawings in the previous examples are clearly *not* independent events. To make this precise we have:

Definition 7 *Two events A and B from a given sample space and probability distribution are said to be* independent *if*

$$P\left(A \cap B|S\right) = P\left(A|S\right) P\left(B|S\right) \tag{2.16}$$

If we compare this with the conditional probability formula 2.12)

$$P\left(A \cap B|S\right) = P\left(A|B\right) \cdot P\left(B|S\right) \tag{2.17}$$

which being obtained much more generally is always true we can conclude that two events A and B are independent if $P\left(A|S\right) = P\left(A|B\right)$. And if we recall our earlier discussion this means that the occurrence of B does not affect the probability of A. And that is about as good a notion of independence as we shall get.

This definition is used both ways. If common sense tells us that A and B are independent we will use the simpler equation 2.16 to calculate probabilities. On the other hand in a more complex situation it may not be obvious whether two events are independent or not. We can find out by calculating $P\left(A \cap B|S\right)$ both using equation 2.12 and using equation 2.16. If the answers agree we conclude that the events are independent.

Example 12 *A coin is tossed and a die thrown. What is the probability of getting a head and a six?*

Solution.

The sample space is $S = \{(H,1),(T,1),(H,2),(T,2),..........(T,6)\}$

We require $P\left((H,6)\left|S\right.\right)$. *The coin tossing and die throwing are independent events so* $P\left((H,6)\left|S\right.\right) = P\left(H\left|S\right.\right)P\left(6\left|S\right.\right)$. *Here* $P\left(H\left|S\right.\right)$ *means the probability of getting a head, the die being irrelevant, and* $P\left(6\left|S\right.\right)$ *means the probability of getting a 6 the coin being irrelevant. i.e. H is the event* $H = \{(H,1),(H,2),(H,3),(H,4),(H,5),(H,6)\}$

$$P\left((H,6)\left|S\right.\right) = \frac{1}{2}\cdot\frac{1}{6} = \frac{1}{12}. \tag{2.18}$$

Example 13 *A bag contains 2 white balls and 6 red balls. A second bag contains 4 white balls and 4 red balls. what is the probability of drawing 2 white balls if:*

(i) one ball is drawn from each bag.

(ii) 2 balls are drawn from the first bag.

(iii) 2 balls are drawn from the second bag.

Solution.

(i) these are independent events so $P\left(W_1 \cap W_2\right) = \frac{2}{8}\cdot\frac{4}{8} = \frac{1}{8}$.

(ii) these are not independent events so

$$P\left(W_2 \cap W_1\left|S\right.\right) = P\left(W_2\left|W_1\right.\right)P\left(W_1\left|S\right.\right) = \frac{1}{7}\cdot\frac{2}{8} = \frac{1}{28} \tag{2.19}$$

(iii) Again $P\left(W_2 \cap W_1\left|S\right.\right) = P\left(W_2\left|W_1\right.\right)P\left(W_1\left|S\right.\right) = \frac{3}{7}\cdot\frac{4}{8} = \frac{3}{14}$.

Our final example shows how Venn diagrams can sometimes be useful.

Example 14 *A cancelled export order led to a furniture sale at which tables, sideboards and bookcases were sold. There were 200 tables, 160 sideboards and 160 bookcases and customers were limited to a maximum of one of each item. 285 people attended the sale and only 25 left without buying anything. 84 people bought all three items, 32 bought a sideboard and a table only, and 40 bought a table only. All the items were sold.*

What is the probability that a person chosen at random from the 285 sale goers bought a sideboard only?

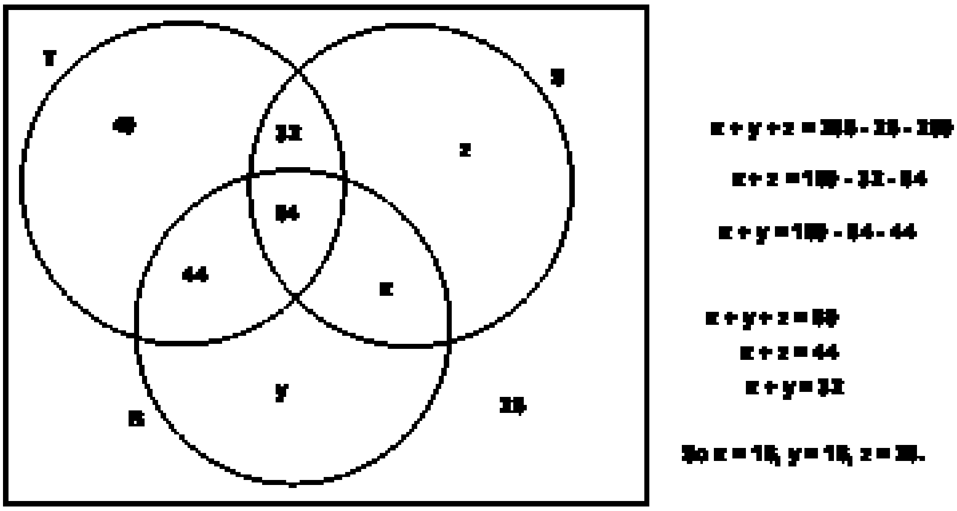
Solution.

Hence P(bought a sideboard only) $= \frac{28}{285} = 0.098$.

2.1.1 Exercises

1. Four balls are drawn at random simultaneously from an urn containing 1 purple, 2 yellow, 3 red and 6 blue balls. Find the chance that the balls are (a) all blue, (b) one of each color.

2. In a certain school, the probability that a person studying German also studies Physics is 1/4, whereas the probability that someone studying Physics also studies German is 1/5. The probability that a person chosen at random studies neither is 1/3. Calculate the probability that a person chosen at random studies both Physics and German.



Could it be said that Physics and German were independent choices at that school?

3. A man has 3 black counters in his left pocket and 3 white counters in his right pocket. He takes one counter at random from the left pocket and places it in the right. He then takes one at random from the right pocket and places it in the left. What is the chance that there are again 3 black counters in the left pocket?

4. Two bags each contain 3 red balls and 1 black ball. A ball is transferred at random from the first bag to the second. Then one is transferred at random from the second bag to the first. What is the probability that there are again 3 red balls and 1 black ball in each bag?

5. Show that if A_1 and A_2 are independent events, then (a) A_1' and A_2'

   (b) A_1' and A_2 are also independent events.

6. Show that if A_1 and A_2 are possible events which are mutually exclusive then they cannot be independent. Construct definitions for A_1 and A_2 if each represents the drawing of a specific card from a well shuffled pack.

7. On a stretch of main road there are 4 independent sets of traffic lights, each phased for 120 seconds red, 60 seconds green. What is the probability that a motorist arriving at random will have to stop at least once?

8. A poker player is dealt 5 cards. What is the chance that he receives a royal flush? (Ace, King, Queen, Jack, 10 of the same suit)

9. An urn contains 3 black and 7 white balls. Balls are drawn at random one at a time from the urn and are not replaced. Find the probability that the first black ball to be drawn is drawn on the fourth attempt.

10. Two dice are thrown. Find the probability that the product of their scores is even. If n dice are thrown, what is the probability that their product is even?

11. Three people are chosen at random. Assuming that births are equally likely throughout the week, what is the probability that they were all born on different days of the week? What is the probability that 7 people chosen at random were all born on different days of the week?

12. What is the probability of drawing first a seven and then an eight from a pack of cards?

13. In a table of random numbers, consecutive repeated digits are called doubles, triples and so on. Work out the probability that a digit chosen at random from the table will be part of (a) a double (b) a triple.

14. A car driver has four keys, only one of which will open the door. Given that the keys are otherwise indistinguishable, find the probability (before he starts trying them) that the door will open on the first, second, third and fourth attempts.

    (a) Consider two cases where (i) he discards each key which fails to open the door, (ii) he returns each key to the collection before choosing the next one at random.

    (b) Then consider the cumulative probabilities with each strategy, i.e. the probability that he will have succeeded by the first, second, third and fourth attempts.

15. Smith and Wesson fight a duel with guns. The probability that Smith kills Wesson on any shot is 1/4 and that Wesson kills Smith is 1/3. Find the probability that just one of them is killed after one shot each if (a) both fire simultaneously (b) Smith fires first.

    If both fire simultaneously, what is the probability that both are still alive after the second round?

    What is Smith's probability of survival if they duel to the death, each time firing simultaneously?

16. Two letters of the words SEEN SENSE are chosen at random without replacement. Set up a tree diagram and show that the probability that the second letter chosen is an E is the same as the probability that the first letter chosen is an E. Show that this property is also true for the letters N and S and explain why.

    Explain also why the results are the same as they would be if the selection took place with replacement.

17. One wine rack contains 3 white, 7 red and 2 *rose* bottles, and a second contains 4 white, 4 red and 1 *rose*. A rack is chosen at random and then a bottle is chosen at random from that rack. Draw a probability tree and find the probability that the bottle so chosen is red.

    Which color wine has an improved chance of selection if chosen by this method rather than by true random selection?

18. Accident casualties in a hospital are in the proportion 2/3 male 1/3 female. The hospital is large and has one large and two small mixed casualty wards. Victims of accidents are assigned alternately to large and small wards. The pattern is $LS_1LS_2LS_1$.....etc. Ward L has 7 male and 14 female nurses. Ward S_1 has 3 male and 5 female nurses. Ward S_2 has 4 male and 8 female nurses. The rota system means that the nurses on duty are effectively chosen at random.

    Find the probability that a victim chosen at random is male and is received into the ward by a male nurse.

19. A day which is fine has a probability of 3/4 of being followed by another fine day. A day which is wet has a probability of 2/3 of being followed by another wet day. Given that days are classified as either fine or wet and that June 6th is fine, set out a tree diagram for June 7th, 8th,9th. Calculate the probability that a day in June is fine.

20. A batch of fifty articles contains three which are defective. The articles are drawn in succession (without replacement) from the batch and tested. Show that the chance that the first defective met will be the rth article drawn is $\frac{(50-r)(49-r)}{39200}$.

21. With a rather gloomy weather forecast for an agricultural show the organizers decided to hold a draw for the "lucky program" holder. Free programs, eligible for the draw were given to the first 300 visitors entering the ground.

    A count was made of whether each visitor was equipped with an umbrella (U), a raincoat (R) and wellingtons (W). The results were:

    Umbrella 75 Umbrella and raincoat 40 None of these 60

    Raincoat 140 Umbrella and wellingtons 25

    Wellingtons 105 Raincoat and wellingtons 30

    Find the probability that the winner of the "lucky program" draw has an umbrella but no raincoat or wellingtons.

2.2 Reading: Concerning Probability by Pierre Simon de Laplace

2.2.1 Introduction to the Reading

2.2.2 The Paper

Chapter 3

Bernoulli's and Bayes' Theorem

3.1 Inverse Probability and Bayes Theorem.

We will start this section with an example.

Example 15 *One bag (I) contains 3 red, 2 white and 1 blue ball. A second bag (II) contains no red, 4 white and 2 blue balls.*

A die is thrown: if a one or a six appears bag I is chosen, otherwise bag II is chosen. A ball is then drawn from the selected bag. We know that the result of a certain trial was that a white ball was drawn (and that is all we know). Find the probability that it came from bag I.

Solution.

We assume that the balls are indistinguishable apart from color and that we are dealing with a fair die.

Let I = *"bag I was chosen"*

Let II = *"bag II was chosen"*

Let W = *"a white ball is drawn".*

Clearly $P(I) = \frac{1}{3}$ *and* $P(II) = \frac{2}{3}$

These are the a priori probabilities. But 1/3 is not the answer. The probability that we require is modified by our knowledge of the color of the drawn ball.

The probability of W *before the experiment is carried out is* $P(W) = P(W \cap I) + P(W \cap II)$*i.e. the sum of the probabilities of getting a white ball via bag I and via bag II.*

$$\begin{aligned} \textit{Hence } P(W) &= P(W|I)\,P(I) + P(W|II)\,P(II) \\ &= \tfrac{2}{6}\cdot\tfrac{1}{3} + \tfrac{4}{6}\cdot\tfrac{2}{3} = \tfrac{5}{9} \end{aligned} \tag{3.1}$$

But the probability that we want is $P(I|W)$*. Using equation (A) we have*

$$P(I \cap W) = P(I|W)\,P(W|S) \tag{3.2}$$

where S *is the underlying outcome space on which we calculated* $P(W) = \frac{5}{9}$

$$\begin{gathered} \textit{Hence } P(I|W) = \tfrac{P(I\cap W)}{P(W|S)} \\ = \tfrac{P(W|I)P(I)}{P(W|S)} \ \textit{(we use the well known set result } A \cap B = B \cap A\textit{)} \\ = \frac{\frac{2}{6}\cdot\frac{1}{3}}{\frac{5}{9}} = \tfrac{1}{5} \end{gathered} \tag{3.3}$$

Thus the required probability is 5/9.

The crux of this solution is the use of the set theoretic result $A \cap B = B \cap A$.

From the easily calculated $P(W|I)$ we have calculated the far from obvious $P(I|W)$. For obvious reasons this is called an inverse probability calculation.

If any of you have studied probability before at a more elementary level you will recall that it is possible to muddle ones way through problems such as Examples 11, 12 and 13 without explicit use of equation 2.12. However our use of equation 2.12 is essential in example 15. In general the method used here of basing all our calculations on the single equation 2.12 would be the preferred one at a more advanced level. Remember that equation 2.12 is not proved: it is derived from plausible arguments admittedly but the ultimate test is "does it work?". If anything went wrong and we had to revise our theory it would then only be equation 2.12 that would need examination. But if we had been solving all our problems by ad hoc methods each of these would require examination.

Let us try another example in inverse probability.

Example 16 *Three holiday caravans X,Y,Z are let to 2 adults and 4 children, 4 adults and 2 children, and 3 adults and 1 child respectively. A special holiday prize is awarded by first choosing a caravan at random and then choosing one of its occupants at random. Given the information that the winner was a child find the probability that it came from caravan X.*

Solution.

With obvious notation we require $P(X|C)$*.*

$$\begin{aligned} &P(X \cap C) = P(X|C)\,P(C) \\ &P(X|C) = \tfrac{P(X\cap C)}{P(C)} = \tfrac{P(C\cap X)}{P(C|X)+P(C|Y)+P(C|Z)} \\ &= \frac{\frac{4}{6}\cdot\frac{1}{3}}{\frac{4}{6}\cdot\frac{1}{3}+\frac{2}{6}\cdot\frac{1}{3}+\frac{1}{4}\cdot\frac{1}{3}} = \frac{8}{15} \end{aligned} \tag{3.4}$$

The required probability is thus 8/15.

Example 17 *An engineer is responsible for servicing three computer installations A,B and C. The probability that A,B,C will break down on any given day is* $\frac{1}{40}, \frac{1}{50}, \frac{1}{100}$ *respectively. The probabilities that the computers will not be in operation after 7 p.m. are* $\frac{5}{9}, \frac{2}{3}, \frac{3}{4}$*respectively.*

One day the engineer receives a message that one of the computers has broken down at 8 p.m. but the message does not say which. To which computer should the engineer go first?

Solution.

Let A = "computer A breaks down" and similarly for B, C.

Let L = "there is a breakdown after 7 p.m."

We require $P(A|L)\,, \quad P(B|L)\,, \quad P(C|L)$

$$P(L \cap A) = P(L|A)\,P(A) = P(A \cap L) = P(A|L)\,P(L) \tag{3.5}$$

Hence $P(A|L) = \frac{P(A\cap L)}{P(L)}$

$$P(A|L) = \frac{\frac{1}{40}\cdot\frac{4}{9}}{\frac{1}{40}\cdot\frac{4}{9}+\frac{1}{50}\cdot\frac{1}{3}+\frac{1}{100}\cdot\frac{1}{4}} = \frac{40}{\underline{\underline{73}}} \tag{3.6}$$

Similarly $P\left(B\left|L\right.\right)=\frac{\frac{1}{50}\cdot\frac{1}{3}}{\frac{1}{40}\cdot\frac{4}{9}+\frac{1}{50}\cdot\frac{1}{3}+\frac{1}{100}\cdot\frac{1}{4}}=\underline{\underline{\frac{24}{73}}}$
and $P\left(C\left|L\right.\right)=\underline{\underline{\frac{9}{73}}}$

We conclude that the engineer should first go to visit computer A.

The methods that we have developed in these problems on inverse probability was first used by Thomas Bayes around 1764. In later times it was crystallized into a theorem, known of course as Bayes Theorem.

The examples above, are special cases of Bayes theorem in fact and are perfectly proved starting from equation 2.12 as usual. (The general proof also starts from equation 2.12 which there appears to be no way of avoiding.).

3.1.1 Bayes Theorem

Let us consider a set A and a family of sets B_j. From conditional probability we know that

$$P(A\bigcap B_j)=P(A)P(B_j|A) \tag{3.7}$$

and also

$$P(A\bigcap B_j)=P(B_j\bigcap A)=P(B_j)P(A|B_j) \tag{3.8}$$

From these equations, we may easily see that

$$P(B_j|A)=\frac{P(B_j)P(A|B_j)}{P(A)} \tag{3.9}$$

We now consider the A_i to be events in our sample space S. Clearly we have $A_i\neq A_j$ if and only if $i\neq j$ and

$$S=\bigcup_{i=1}^{N}A_i. \tag{3.10}$$

Then we have that

$$A=A\bigcap S=A\bigcap\left(\bigcup_{i=1}^{N}A_i\right)=\bigcup_{i=1}^{N}\left(A\bigcap A_i\right) \tag{3.11}$$

and thus

$$P(A)=P\left(\bigcup_{i=1}^{N}\left(A\bigcap A_i\right)\right)=\sum_{i=1}^{N}P\left(A\bigcap A_i\right) \tag{3.12}$$

and by conditional probability, we have

$$P(A)=\sum_{i=1}^{N}P\left(A\bigcap A_i\right)=\sum_{i=1}^{N}P\left(A_i\right)P\left(A|A_i\right) \tag{3.13}$$

Finally we end up with Bayes Theorem in general,

$$P\left(A_i|A\right)=\frac{P\left(A_i\right)P\left(A|A_i\right)}{\sum\limits_{j=1}^{N}P\left(A_j\right)P\left(A|A_j\right)} \tag{3.14}$$

3.1.2 Exercises

1. One of four dice is known to be biased, showing a 6, on average, three times as often as each of the other scores. A die is chosen at random and thrown three times and three 6's appear. What is the probability that it is one of the unbiased dice?

2. A factory manufactures three different qualities of light bulb, A, B and C in the ratio 1:2:3. The bulbs are indistinguishable in external appearance but extensive tests indicate that, on the average, 1% of type A, 4% of type B and 5% of type C are below the advertised standard. A batch of 6, all of the same type is sent in error without a distinguishing label. If none of these turns out to be defective, estimate the probability that they were of type A.

3. Of a large sample of items, 40% were produced by machine A and 30% by each of machines B and C. The machinery was very unreliable, machine A producing 10% defective items, machine B 20% defective items and machine C 30%. If an item, selected at random from the sample, proves to be defective, what is the probability that it came from machine A?

4. A card is missing from a pack of 52 cards. If this is the only information you have, what is the probability that the missing card is a spade? The pack is well shuffled and the first card is removed and proves to be a spade. What would your assessment of the probability that the missing card is a spade be now? The card removed is now replaced and the pack shuffled. The top card again proves to be a spade. What is your assessment of the probability now?

5. Four bags I, II, III, IV, contain white and black balls as shown in the following table.

   Bag I II III IV

   Number of white balls 1 2 3 4

   Number of black balls 9 8 7 6

   A die is thrown; if a one appears bag I is chosen, if a 2 or 3 bag II, if a 4 or 5 bag III, and if a 6 bag IV. A ball is then drawn at random from the bag selected. If you are told that a black ball has been drawn, what should your estimate be of the probability that it came from bag I?

6. A certain rare disease from which one in ten thousand of the population suffers is diagnosed by a test which reveals the presence of the disease in 95% of the cases of those tested who actually have the disease. However, it also incorrectly yields a positive reaction in 1% of the cases of those who are not suffering from the disease. If a person selected at random from the population shows a positive reaction what is the probability that he is actually suffering from the disease?

7. One of four pennies is known to be double headed, the other three being normal. One of the pennies is selected at random and tossed three times. If the result is three heads, what is the probability that the coin tossed is the double headed one?

8. A farmer grows three varieties A, B, C of strawberries in the ratio of 50:30:20. The proportion of fruits classified as ”large” is 30% for variety A, 40% for B and 60% for C. Bird damage accounts for 5% of all fruits of variety A, 10% of B and 20% of C.

   (a) What proportion of large and small strawberries are produced on the farm?

   (b) What proportion of bird damaged strawberries are large and small?

   (c) What proportion of large and small strawberries are bird damaged?

   (d) Is there evidence that birds attack large and small strawberries unequally?

9. In a hotel 70% of the guests are male and 30% female. Of the male guests the proportions who are travelling on business and on holiday are 80% and 20%. For female guests the proportions are 10% and 90%. The hotel gives a choice of continental or English breakfasts. The proportion of male and female guests on holiday who take the English breakfast are 80% and 50% respectively. The proportion of guests on business taking English breakfast is the same for men and women at 70%. Calculate:

   (a) the proportion of guests in the hotel who are on holiday.

   (b) the proportion of continental breakfasts served.

   (c) the probability that a guest who has a continental breakfast is on business.

   (d) the probability that a guest who has an English breakfast is a woman on holiday.

10. One urn contained 3 white and 3 black balls. A second urn contained 4 white and 3 black balls. A ball is transferred unseen from the first to the second urn. Then a ball is drawn at random from the second urn. Calculate:

    (a) the probability that the ball which is drawn is white.

    (b) the probability that if the ball drawn is white, then the transferred ball was also white.

3.2 Bernoulli's Theorem

Bernoulli's theorem is one of those instances in probability theory where an extremely simple sounding statement is very deep indeed and as such usually neglected, misunderstood or deliberately misused. Bernoulli himself said that it took him twenty years of thought before he was confident enough to write it down as a theorem. Without further ado we shall state it and then discuss it. For a more complete discussion, see the reading.

Theorem 3 (Bernoulli's theorem or the Law of Large Numbers) *We are concerned with finding the probability of a certain event. Let the real probability of this event be p. If a number of independent trials are carried out under identical conditions, then the most probable ratio of successful trials to total trials is p and in addition the probability that the computed ratio will differ from p by less than a given tolerance level will increase with increasing number of trials.*

This means that the ratio interpretation of probability is asymptotically correct. It is this statement that gives the key to experimental statistics. All we must do to assess a probability is carry out a certain number of experiments and compute a ratio. As stated above, the theorem is a deep philosophical statement that can be thought about for a long time. Later we will state it again in different language but then in a manner that can be proven mathematically. Of course, this proof, as all proofs in mathematics, depend crucially on a set of assumptions that may or may not hold in reality. The judgement of whether these assumption do or do not hold in reality is not the concern of mathematics but of applied science.

The law of large numbers not only says that the ratio interpretation is the best one but that it will become more and more correct as more trials are made. Intuitively this makes sense as more information should means more certain conclusions. The statement of asymptotic correctness begs the question of a formula for the error. This is a totally non-trivial question going beyond these notes.

3.3 Reading: The Law of Large Numbers by Jacob Bernoulli

3.3.1 Introduction to the Reading

3.3.2 The Paper

Chapter 4

The Notion of Random Variables

4.1 Variables

From basic high school algebra we know variables as “place holders” for numbers that we do not yet know or that we want to hide for some reason. We then learn to “solve” equations of the type $x+1 = 2$ for the variable x. The crucial understanding is that x, by and of itself, can take on any value whatsoever, it need not be 1. If the equation above is *in addition* supposed to be true, then the variable x is forced to take on the value 1.

We note that x in the above equation is not really variable in the sense that its value may change. It is really an unknown constant that we have labelled x for lack of a better name. We have, after some work, found a much better name for it, namely 1. Thus, x is *not* a variable in our sense of the word. It is a constant!

This is the understanding that allows us to then learn about straight lines, i.e. the equation $y = mx+b$. The colloquial agreement is that x is called the *independent variable* and y the *dependent variable*. This may sound silly but the reason is very deep. Before we explain it, we note that both m and b are constants in the above sense, i.e. they are not meant to change for a particular line and take on numerical values.

Both x and y are variables in the sense that they may change. For a fixed m and b (and they are, of course, always fixed for any particular line), any particular value for x will allow precisely one value for y. Thus these two variables are *dependent* on each other. Fixing one fixes the other. It does not matter which one we vary and which one we “solve for” by our high school algebra from above. It is however important that we decide this in a systematic fashion. This is why the equation is usually written $y = mx + b$. You are meant to pick values for x at your leisure and then work out the value for y by simple arithmetic. This procedure makes x dependent only on your whim, i.e. independent of the problem at hand and y dependent on x. Hence the names.

We then agree further that x shall be graphed on the horizontal axis and y on the vertical axis. The result is then a picture of several points which we have worked out so far. We may join them by a line and we obtain a picture of a straight line segment in between the smallest and largest values of x we have happened to choose. We can never, of course, draw the whole line as it is infinitely long.

Based on these two fundamental choices, we shall agree on something quite important: The horizontal axis will nearly always hold a piece of information that has been selected by a human being for some reasons. These reasons may or may not be logical, fair, unbiased. In fact, in popular statistics that one may read in the newspaper, these choices are usually very carefully made so that the final picture looks most like what the author needs for the particular policy he or she wishes to propound. The vertical axis will consequently show the datum that has been arrived at either by calculation or experimentation. Again, frequently the experimentation is biased or faulty but this is a topic for later.

To give an example, let us consider that we have carried out a experiment to ascertain the heights of a large number of men. We wish to display this information. Generally this is done by drawing a histogram. The basic choice here is that of bins. We choose to start with 100 cm and go to 250 cm in steps of 10 cm and count how many men are in each interval. Presumably, all the men asked will fall into one of these intervals and we may draw the result. Note that what is variable here is our choice of bin size only. The largest and smallest height and the number of men of any height is fixed by experimentation. Thus the only choice we have is the choice of bins. Having all bins of equal size is considered good practise and it is the experience of many that the distribution looks like a bell curve. This will soon be called a normal distribution as we will learn later on.

What does one do if one wants to establish the thesis that there are a large number of short men? An easy remedy for this is to introduce a single bin from 100 cm to 150 cm and then count in 10 cm steps. This will increase the count in the first bin tremendously. This is known as a biased plot. All you then have to do is to print the legend in a small font and in an inconvenient place using strange words and this fact will completely go by most readers of the newspaper you work for. Unfortunately this little joke and its many variants is practised widely. So watch out for it!

4.2 What is Random?

Something is called *deterministic* when we can predict what its future values are going to be. It is called *indeterministic* we can not do this. Out inability to predict does however not mean that it is fundamentally so. We may simply not have hit upon the underlying regularity yet. It is a popular view that everything we have not predictively understood is given to chance or to a higher power.

Mathematicians agree that there a few good sources of random data: the weather and the stock market. It is important to understand that this data is not truly random because both respond to pressures and both are governed by known laws. They are however indeterministic in the sense that we can not yet predict them and so we view them as random. Whether there is anything truly random in the universe is a matter of philosophical debate that we will not go into here.

For our purposes, random will be taken to mean:

Definition 8 *A* random event *is an event whose occurrence is influenced by so many factors and subject to so many conditions that are not under our control or not subject to sufficiently accurate measurement that the accurate prediction of this event is not possible given reasonable means.*

Please be careful to note all the qualifying words in that definition!! For a gambler, the roll of a die is a random event even though we have enough technology available to us to measure the strength and direction of the throw, the properties of the environment, etc. to be able to predict the numerical outcome of the throw. In fact, a simple gravitational model of the game of roulette combined with educated guesses as to the angle of the table etc. made a group of mathematicians millions of dollars a few decades ago. Thus the degree of randomness of an event is something relative to the experiment. Fundamentally random event are those that we can not, even theoretically, predict with any amount of presently available technology. Examples include the exact moment of decay of a radioactive atomic nucleus or the exact path taken by a tornado.

4.3 Random Variables

Definition 9 *A* random variable *is a variable that takes random values.*

For the gambler, the variable x defined as the outcome of a single throw of a single die is a random variable. He can say that x will definitely be one of 1, 2, 3, 4, 5 or 6 and that, to the best of his knowledge they are all equally likely to occur, but he can not say anymore than that. In fact, the occurrences of the six possibilities in practise are *never* truly equally likely because of small imperfections in the physical makeup of the system under investigation. For a die, for example, a small chip on one side causes an imbalance that will show up a non-uniformity in the throwing frequency over a long time. This will not, of course, matter for a game that will have a hundred throws or so unless the die is badly damaged and as such is of no interest to the gambler. However, this should be borne in mind.

We are concerned with obtaining a sequence of random numbers for many practical applications. We simply note here that there exist algorithms for the computer that are perfectly deterministic in nature if examined line by line but that produce a sequence of numbers without apparent regularity. Philosophically speaking, the regularity in this sequence of numbers is so complex that the likelihood of arriving at it given the sequence is virtually zero. Thus, by the above definition, the numbers are random events. But we do know, in fact, the method by which they are obtained and so we call them *pseudo-random numbers*. It is a complicated matter to study and compare these methods profitably and we shall not need to do so. It suffices to say that they have achieved a quality that they can be considered (for all practical purposes) random. Furthermore, these methods are included in all standard data processing computer packages.

4.4 Reading: The Red and the Black by Charles Sanders Peirce

4.4.1 Introduction to the Reading

4.4.2 The Paper

Chapter 5

Expectation

Probability theory and statistics are very closely linked. In this chapter our aim is to develop the ideas upon which this linkage is based. This is done through the concepts of random variables, mean, variance, standard deviation and expectation value.

To begin with let us return to the simple probability problem of throwing an unbiased 6 faced die. The outcome space is $S = \{1, 2, 3, 4, 5, 6\}$, and as we know, each of the elementary events has probability 1/6. For this example the set of integers $\{1, 2, 3, 4, 5, 6\}$ is the set of values of the random variable that this experiment determines. It is called random since before the die is thrown we have no idea which one of the 6 values we will get. It is called a variable since from trial to trial we will get different values. In this case the experiment determines the random variable in a natural manner.

But now consider a different problem in probability. We have a bag containing 2 red balls, 1 blue ball and 3 green balls. We make many experiments of drawing 1 ball with the ball drawn being replaced before each new trial. We know that:

$$\begin{array}{l}\text{P(red ball drawn)} = \frac{2}{6}\\ \text{P(blue ball drawn)} = \frac{1}{6}\\ \text{P(green ball drawn)} = \frac{3}{6}\end{array} \tag{5.1}$$

There is now no obvious way of assigning numbers (random variables) to each of these events. In fact it would be rather pointless.

But suppose we turn this bag of colored balls into a rather rudimentary gambling game. A player simply draws a ball from the bag and replaces it after each draw. If he draws a red ball he wins 5 peso. If he draws the blue ball he wins 10 peso and if he draws a green ball he has to pay 8 peso. We then naturally associate the numbers 5, 10, -8 with the drawing of a red, blue or green ball respectively.

Clearly it would be possible to arbitrarily assign a random variable to each of the elementary events of any outcome space. However it will not be done (in this course at any rate) unless the values of the random variables have some significance as is the case in the examples that we have so far discussed.

Definition 10 *Suppose we have a random experiment whose outcome space S consists of n elementary events i.e. $S = \{e_1, e_2, \cdots e_n\}$. We will have a probability p_i associated with each event e_i for $1 \leq i \leq n$. A set of real numbers $X = \{x_1, x_2, \cdots x_n\}$*

where the number x_i is associated with the event e_i is called a random variable *defined on the outcome space. x_i is called the* value *of the random variable for (or at) the event e_i. We can write $X(e_i) = x_i$ as in function theory.*

Review the rules for assigning probabilities recalling that:

1. $p_i > 0$ for $1 \leq i \leq n$
2. $\sum\limits_{i=1}^{n} p_i = 1$
3. The set $\{p_1, p_2, \cdots p_n\}$ is called a *finite probability distribution.*

For our simple game with the colored balls we can set : e_1 : a red ball is drawn e_2: a blue ball is drawn e_3 : a green ball is drawn. x_1 = 5, x_2 = 10, x_3 = -8. $X = \{5, 10, -8\}$

$$X(e_1) = 5 \qquad X(e_2) = 10 \qquad X(e_3) = -8. \tag{5.2}$$

You will have noticed by now that we could have defined a random variable as a function whose domain is the outcome space and whose range is the set of real numbers. This is strongly brought out by the notation $X(e_i) = x_i$. Remember that X is the random variable (or function) and $x_i$is a particular value that it takes.

Example 18 *Player A pays 5 peso for the privilege of tossing two unbiased coins. If he gets two heads player B pays him 5 peso and returns his stake money. If A gets a head and a tail (order immaterial) B only returns his stake money. If A gets two tails B keeps the stake money.*

From player A's point of view we could assign a random variable A as: A(HH) = 5, A(HT) = 0, A(TH) = 0, A(TT) = -5. From player B's point of view we could assign a random variable B as: B(HH) = -5, B(HT) = 0, B(TH) = 0, B(TT) = 5. Be aware that in solving problems there may be several ways to assign a random variable. Make a choice and stick to it.

Example 19 *Player A pays 10 peso to throw two dice. For any double he receives his stake money back together with a prize of 100 peso for a double 6 and 20 peso for any other double. Otherwise he looses his stake. Set up a suitable outcome space, probability distribution and random variable for this game.*

Solution.

Let e_1 be a double 6. Let e_2 be any other double. Let e_3 be anything else. Then $S = \{e_1, e_2, e_3\}$ $\quad p_1 = \frac{1}{36}, \quad p_2 = \frac{5}{36}, \quad p_3 = \frac{30}{36}$ *and we take* $x_1 = 110, \quad x_2 = 30, \quad x_3 = -10$.

Example 20 *A man insures his life for one year for \$100000 paying a premium of \$P. Discuss.*

Solution.

The outcome space is just: e_1 :the man survives the year e_2 :the man dies during the year. We don't know the probabilities. Actuaries try to estimate them using age, occupation, smoking habits etc. etc. We could assign $x_1 = -P$ as the premium is certainly not returned. $x_2 = 100000 - P$since premiums are not treated as stake money. We also note that such calculations may be of great interest to the relatives and the insurance company but are of no interest at all to the patient.

5.1 Expectation Value.

Closely connected with the notion of a random variable defined, together with a probability distribution on an outcome space is the idea of expectation value. (It is often referred to simply as expectation). To start let us return to our simple gambling game with the colored balls. Recall that we have:

$$\begin{array}{l} e_1 \text{ a red ball is drawn } p_1 = \frac{2}{6}, \quad x_1 = 5 \\ e_2 \text{ a blue ball is drawn } p_2 = \frac{1}{6}, \quad x_2 = 10 \\ e_3 \text{ a green ball is drawn } p_3 = \frac{3}{6}, \quad x_3 = -8 \end{array} \tag{5.3}$$

Suppose a compulsive gambler plays a very large number N of games. We would expect him to get a red ball on approximately $\frac{2N}{6}$ occasions thereby winning $\frac{5N}{3}$ peso. We would expect him to draw a blue ball on approximately $\frac{N}{6}$ occasions thereby winning another $\frac{5N}{3}$ peso. We would expect him to draw a green ball on approximately $\frac{N}{2}$ occasions thereby loosing $4N$ peso. His net gain is thus about $\frac{5N}{3} + \frac{5N}{3} - 4N = -\frac{2N}{3}$ peso (i.e. a loss). His average loss per game is thus about 2/3 peso.

As N gets greater and greater we would expect the actual average calculated from the actual results to get closer and closer to this value of 2/3. It is then this theoretically calculated average which is called the *expectation value* of the random variable. Do not confuse it with the mean (average) which must be computed from experiments actually carried out. The term expectation value is used because we can say that if someone plays this game a large number of times he can *expect* his losses to be on average 2/3 peso per game.

The expectation value need not be (and in this example certainly is not) an element in the range set of the random variable. Having illustrated the concept of expectation value with this simple example let us try to develop a general definition.

Given an outcome space $S = \{e_1, e_2, \cdots e_n\}$ with a probability distribution $\{p_1, p_2, \cdots p_n\}$ and a random variable $X : e_i \mapsto x_i$ for $1 \leq i \leq n$. Suppose that we carry out a very large number of experiments N. We would expect to get the value x_i approximately $p_i N$ times and so the average value should be about

$$\frac{p_1 x_1 N + p_2 x_2 N + \cdots\cdots + p_n x_n N}{N} = \underline{\underline{\sum_{i=1}^{n} p_i x_i}} \tag{5.4}$$

The theoretical value $E(X) = \sum\limits_{i=1}^{n} p_i x_i$ is called the *expectation value* of the variable X.

Example 21 *An unbiased die is thrown. What is the expected score? (i.e. what is the expectation value of the score?)*

Solution.

Here $E(X) = \sum\limits_{i=1}^{6} p_i x_i = \sum\limits_{i=1}^{6} \frac{i}{6} = \frac{1}{6}(1+2+3+4+5+6) = 3.5$

Example 22 *The inveterate gamblers A and B play a game with three coins. A pays B a stake of 10 peso and tosses the three coins in turn. If he gets 3 heads B* *returns his stake plus a prize of 30 peso. For 2 consecutive heads B returns the stake plus a prize of 10 peso. In all other cases B keeps the stake and there is no prize. Is this a fair game?*

Solution.

[In game theory a game is said to be fair if the expectation value of gain is zero for both players.]

Here $S = \{(HHH), (HHT), (THH), (others)\}$ *P(HHH)* $= \frac{1}{8}$, *P(HHT) = P(THH)* $= \frac{1}{8}$, *P(others)* $= 1 - \frac{1}{8} - \frac{1}{4} = \frac{5}{8}$. *For random variable we take player A's net gain: X(HHH) =30, X(HHT) = X(THH) = 10, X(others) = -10. So A's expectation of net gain is* $\frac{30}{8} + \frac{10}{8} + \frac{10}{8} - \frac{5.10}{8} = 0$. *In a two person game one players gain is the others loss so we need not calculate separately for player B. The game is fair. Over a long period of time neither player should win or lose any money.*

Example 23 *A game can result in three possible outcomes whose probabilities are* $\frac{1}{4}, \quad \frac{1}{2}, \quad \frac{1}{4}$.*The players net gain is respectively* $x^2, \quad -x, \quad 1$ *where* $x > 0$. *(a) Find the players expectation of net gain. (b) Is it possible to choose x so that the game is fair?*

Solution.

(a) $E = \frac{x^2}{4} - \frac{x}{2} + \frac{1}{4} = \frac{x^2-2x+1}{4} = \underline{\underline{\frac{(x-1)^2}{4}}}$

(b) Yes. Simply set x = 1.

5.1.1 Exercises

1. The values of a random variable, together with their associated probabilities for four different experiments, are given in the tables below. Calculate E(x) in the four cases.

$$\text{(i)} \begin{array}{ccccccc} x_i & 0 & 1 & 2 & 3 & 4 & 5 \\ p_i & \frac{1}{15} & \frac{2}{15} & \frac{1}{3} & \frac{1}{5} & \frac{2}{15} & \frac{2}{15} \end{array} \tag{5.5}$$

$$\text{(ii)} \begin{array}{cccccc} x_i & -2 & -1 & 0 & 1 & 2 \\ p_i & \frac{1}{10} & \frac{2}{5} & \frac{3}{10} & \frac{1}{10} & \frac{1}{10} \end{array} \tag{5.6}$$

$$\text{(iii)} \begin{array}{ccccccc} x_i & 1 & 2 & 3 & 4 & 5 & 6 \\ p_i & \frac{1}{20} & \frac{3}{20} & \frac{1}{5} & \frac{2}{5} & \frac{1}{10} & \frac{1}{10} \end{array} \tag{5.7}$$

$$\text{(iv)} \begin{array}{ccccccccccc} x_i & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ p_i & 0 & \frac{1}{8} & 0 & \frac{3}{8} & 0 & \frac{1}{4} & 0 & \frac{1}{8} & 0 & \frac{1}{8} \end{array} \tag{5.8}$$

2. In Question 1. a new random variable Y is constructed so that $Y = 2X - 1$. If the probability distributions remain the same, calculate E(Y) in the four cases. Can you draw a general conclusion from these results?

3. A player pays a certain sum of money to spin two coins. For two heads he receives back 10 peso, for two tails he receives 2 peso, for a head and a tail he receives nothing. In all four cases he forfeits his stake money. What should the stake money be for the game to be fair?

4. Two dice are thrown. Find the expectation of the higher score showing (or the score of one of them if they fall alike).

5. If the probability that a man aged 60 will survive another year is 0.9, what premium should he be charged for a life insurance policy of $1000? (If he survives the year he receives no money back.)

6. X_1 and X_2 are two random variables, each with values 0,1,2,3,...,9, and each possessing a uniform probability distribution. Evaluate:

$$\text{(i) } E(X_1 - X_2) \qquad \text{(ii) } E\left(|X_1 - X_2|\right). \tag{5.9}$$

7. Two bags each contain ten colored discs as shown below.

	Red	Green	Blue
Bag I	4	3	3
Bag II	5	3	2

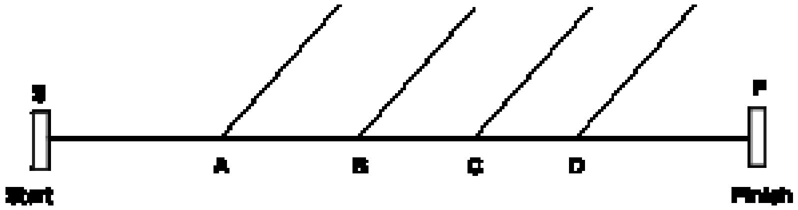
(5.10)

A player stakes a certain sum of money for the privilege of drawing two discs, one from each bag. For two discs of the same color his stake is returned, and in addition he is awarded a prize of 10 peso for two reds, 20 peso for two greens, and 25 peso for two blues. For two discs of different colors he loses his stake.

Show that if the stake money is 8 peso he can anticipate gaining in the long run, but that with the stake at 9 peso he should expect to lose.

8. The game of Question 7 is repeated, but the player now tosses a coin to decide which bag he must choose from: if he tosses a head, he chooses bag I, if a tail, bag II; he then draws a disc at random from the chosen bag, notes its color and replaces the disc before the bag is well rattled and he is allowed to draw a second disc. He is then awarded prizes based on the colors of the discs according to the scheme of Question 7. Determine the minimum stake (to the nearest peso) required to ensure that the player will show a loss in the long run.

9. The path in the figure below represents a simple maze along which a rat is made to run. It starts at S and has to finish at F. If it makes a mistake at A by turning along AA' it will return to A and be forced by the construction of the maze, to turn towards F, and similarly at each of the other junctions. The probability of taking either of the two paths available at each junction is 1/2. Find the expected number of mistakes the rat will make in running from A to F.



The Rat Race.

10. Two dice are thrown in one ”turn”, each turn costing 5 peso. If a prize of 40 peso is given for a double 6, and a prize of 20 peso for any other double (together in both cases with the stake money), determine the expected loss to a player who plays the game 100 times.

11. A man put three \$5 notes into one envelope and three \$1 notes into a similar envelope. Each year at Christmas, he chooses one envelope at random and gives his nephew a note from it. As soon as either envelope is emptied by his taking the last note from it the process ends.

    a. State the different totals that the nephew may have received when the process ends.

    b. For each of these totals calculate the chance of its occurrence.

    c. Show that the nephew's expectation of gain is \$12.375.

5.2 Reading: The Theory of Economic Behavior by Leonid Hurwicz

5.2.1 Introduction to the Reading

5.2.2 The Paper

Chapter 6

Mean, Variance and Standard Deviation

6.1 Standard deviation.

In the previous section we emphasized that the expectation value of a random variable was its theoretical average or mean value. We expect the actual average to approach this value as the number of trials or repetitions of the experiment get greater. The usual letter for averages actually obtained from experiments is μ (or μ_x if we want to refer to the random variable whose mean is being calculated). In practice these letters are also frequently used for expectation values as well, particularly in the context of the general theory which we will soon be developing. Be careful to study the context of a situation so that you know precisely what is involved. People who work with statistics cannot be expected to have the precision of mathematicians or scientists.

Neither expectation values nor actual experimental averages give us any information about the way the values are spread about the mean. In a class of 20 students if 10 students got 100% and 10 students got 0% we would have an average of 50%. This innocent looking average tells us nothing about the extremely strange distribution of marks for this class. In an accurate physics experiment to measure say the mass of an electron we would expect many repetitions of the experiment to yield values all very close to the average value, some a little more, some a little less. But if we take the age of a person selected at random from the population of a large city we have all ages from that of a new born baby to 100+ representing large fluctuations about whatever the mean may be.

Statistics finds it useful to establish a mathematical measure of the distribution of results about a mean. We will first examine the theoretical case in which we perform the calculations directly from the theoretical probability distribution as we did for expectation values.

Suppose that we have two unbiased dice. The first has its face numbered in the usual way from 1 to 6. Denoting its expectation value by E_1 we know that $E_1 = 3.5$. The second die has three faces showing 1 dot and three faces showing 6 dots. If its expectation value is E_2 then clearly $E_2 = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 6 = 3.5$ as before. But the spread of the results about 3.5 is obviously radically different for the two dice. Suppose we define a new random variable for the first die by $Y_1 = X_1 - 3.5$, i.e. $\begin{array}{l} Y_1(1) = 1 - 3.5 = -2.5 \\ Y_1(2) = 2 - 3.5 = -1.5 \\ Y_1(3) = 3 - 3.5 = -0.5 \\ Y_1(4) = 4 - 3.5 = 0.5 \\ Y_1(5) = 5 - 3.5 = 1.5 \\ Y_1(6) = 6 - 3.5 = 2.5 \end{array}$ The values of Y_1 are called the *deviations* from the expectation value (or simply the *deviations from the mean* in both the theoretical and experimental cases). We now calculate the expectation value of these deviations from the mean.

$$E(Y_1) = \frac{1}{2}\left[-2.5 - 1.5 - 0.5 + 0.5 + 1.5 + 2.5\right] = \underline{\underline{0}} \tag{6.1}$$

This does not seem to be a useful measure of the deviation since the deviation is certainly not zero.

To convince ourselves that this interesting method is in fact useless let us carry out the same calculation for the second die. We have:

$$Y_2(1 \text{ dot}) = 1 - 3.5 = -2.5 \text{ and } Y_2(6 \text{ dots}) = 6 - 3.5 = 2.5. \tag{6.2}$$

$$E(Y_2) = \frac{1}{2}\left[-2.5 + 2.5\right] = 0 \tag{6.3}$$

So apart from the inappropriateness of the value 0 we get the same result for both distributions.

[It is interesting to note that the value of zero is inevitable using this method no matter what the situation. For an arbitrary probability distribution and random variable we have:

$$E(X) = \sum_{i=1}^{n} p_i x_i \qquad Y(x_r) = x_r - \sum_{i=1}^{n} p_i x_i \tag{6.4}$$

$$E(Y) = \sum_{r=1}^{n} p_r \left(x_r - \sum_{i=1}^{n} p_i x_i \right) = \sum_{r=1}^{n} p_r x_r - \sum_{r=1}^{n} p_r \left(\sum_{i=1}^{n} p_i x_i \right) \tag{6.5}$$

$$= E(X) - 1 \cdot \sum_{i=1}^{n} p_i x_i = E(X) - E(X) = 0 \tag{6.6}$$

]

In the actual examples it seems to be the cancelling out of the positive and negative deviations which causes the trouble. To avoid this let us try the effect of squaring the deviations. i.e. we have for die 1.

$$\begin{array}{l} Y_1(1) = (1 - 3.5)^2 = 6.25 = Y_1(6) \\ Y_1(2) = (2 - 3.5)^2 = 2.25 = Y_1(5) \\ Y_1(3) = (3 - 3.5)^2 = 0.25 = Y_1(4) \end{array} \tag{6.7}$$

and $E(Y_1) = \frac{1}{3}\left[6.25 + 2.25 + 0.25\right] \underline{\underline{= 2.92}}$

For die 2: $Y_2(1) = (1 - 3.5)^2 = 6.25$ $\quad Y_2(6) = (6 - 3.5)^2 = 6.25$ and $E(Y_2) = \frac{1}{2}\left[6.25 + 6.25\right] = 6.25$.

This makes more sense. The larger value obtained for the second die corresponding to the fact that the values obtained with this die are on average further away from the mean than those obtained with the normal die.

The expectation value of the squared deviations of the random variable X is called the *variance* of X and it is denoted by σ_x^2 . i.e.

$$\begin{aligned}\sigma_x^2 &= E\left[(X-E(X))^2\right] \\ &= \sum_{i=1}^{n} p_i\left[x_i - E(X)\right]^2 = \sum_{i=1}^{n} p_i\left[x_i-\mu_x\right]^2\end{aligned} \tag{6.8}$$

Definition 11 *The positive square root of the variance of X is denoted by σ_x and is called the* standard deviation *of X.*

[Let us outline the calculation for an experimental situation since up to now we have concentrated on the theory. We will not have much occasion for this in this course but it will be of interest for us to see how the statisticians go about it.

Suppose that there are n possible outcomes to an experiment $\{x_1, x_2, \cdots, x_n\}$ and that a large number N of trials are carried out. The trick is to use the a posteriori probabilities $p_i = \frac{n_i}{N}$ for $1 \le i \le n$ where event x_i occurs n_i times. Statisticians are fond of calling these a posteriori probabilities *relative frequencies* by the way.

The mean is then $\frac{\sum_{i=1}^{n} n_i x_i}{N} = \sum_{i=1}^{n} p_i x_i$ This is formally the same as the theoretical value but uses the a posteriori probabilities.

Using this value as μ_x we calculate the experimentally observed variance exactly as before. e.g. $\sigma_x^2 = \sum^{p_i[x_i-\mu_x]^2}$. Thus the formulae are exactly the same. It is only necessary to remember that in the statistical case the p_i are the a posteriori probabilities and that these are often called relative frequencies.]

Example 24 *Find the mean and variance of the random variable X whose probability distribution is:*

$$\begin{array}{cccccc} X & 0 & 1 & 2 & 3 & 4 \\ p & \frac{1}{8} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{8} \end{array} \tag{6.9}$$

<u>*Solution.*</u>

$$E(X) = \mu_x = \sum_{i=1}^{n} p_i x_i = \frac{1}{8}\cdot 0 + \frac{1}{4} + \frac{2}{4} + \frac{3}{4} + \frac{4}{8} = 2 \tag{6.10}$$

$$\sigma_x^2 = \sum^{p_i(x_i-\mu_x)^2 = \frac{4}{8}+\frac{1}{8}+\frac{0}{8}+\frac{1}{8}+\frac{4}{8}=\frac{3}{2}} \tag{6.11}$$

Example 25 *Find the variance for the number of heads showing if three unbiased coins are tossed simultaneously.*

<u>*Solution.*</u>

The sample space, the probabilities and the random variable (1 for a head 0 for a tail) are as shown below.

3H 1/8 3
2H 1T 3/8 2
1H 2T 3/8 1

3T 1/8 0

$$E(X) = 3 \cdot \frac{1}{8} + 2 \cdot \frac{3}{8} + 1 \cdot \frac{3}{8} + 0 \cdot \frac{1}{8} \underline{\underline{= \frac{3}{2}}} \tag{6.12}$$

The squared deviations Y are:

$$Y(3H) = \left(3 - \frac{3}{2}\right)^2 \quad Y(2H) = \left(2 - \frac{3}{2}\right)^2 \quad Y(1H) = \left(1 - \frac{3}{2}\right)^2 \quad Y(0H) = \left(0 - \frac{3}{2}\right)^2 \tag{6.13}$$

i.e. $Y(3H) = \frac{9}{4} \qquad Y(2H) = \frac{1}{4} \qquad Y(1H) = \frac{1}{4} \qquad Y(0H) = \frac{9}{4}$

$$\sigma_x^2 = \sum_{i=1}^{4} p_i y_i^2 = \frac{1}{8} \cdot \frac{9}{4} + \frac{3}{8} \cdot \frac{1}{4} + \frac{3}{8} \cdot \frac{1}{4} + \frac{1}{8} \cdot \frac{9}{4} \underline{\underline{= \frac{3}{4}}} \tag{6.14}$$

Thus the mean or expectation value is 3/2 and the variance is 3/4.

Example 26 *A random variable has mean μ and variance σ^2. If a constant c is added to each value of the random variable calculate the new mean and the new variance.*

<u>Solution.</u>

Let there be n events e_i with probabilities p_1 and associated values of random variable x_1 for $1 \leq i \leq n$. Then $\mu = \sum\limits_{i=1}^{n} p_i x_i \qquad \sigma^2 = \sum\limits^{p_i(x_i-\mu)^2}$. The new value of the random variable are $x_i + c$ for $1 \leq i \leq n$. Let the new values of the mean and variance be μ_c and σ_c^2 respectively.

$$\mu_c = \sum_{i=1}^{n} p_i \left(x_i + c\right) = \sum_{i=1}^{n} p_i x_i + \sum_{i=1}^{n} p_i c = \mu + c \tag{6.15}$$

$$\sigma_c^2 = \sum_{i=1}^{n} p_i \left(x_i + c - \mu_c\right)^2 = \sum^{p_i(x_i+c-\mu-c)^2=\sigma^2} \tag{6.16}$$

Thus, as we might have expected, the mean is simply shifted by the same amount c and the spread of the distribution is unaltered.

The results of this example give us a clue to the simplification of some numerical work. We note the properties of what we can call a translated random variable:

$$\mu_c = \mu + c \qquad \sigma_c = \sigma \tag{6.17}$$

Example 27 *The numbers $1, 2, 3, \cdots 20$ are assigned as values of a random variable to 20 equiprobable events. Calculate the mean and variance of the distribution.*

<u>Solution.</u>

Let the original random variable be X. Guess a mean (in this case say 10). Define a new random variable from the guessed mean as Y = X - 10. Then $\mu_y = \sum\limits_{i=1}^{20} \frac{1}{20} y_i = \frac{1}{20}\left[-9 - 8 - ... - 1 + 0 + 1 + 2 + ... + 10\right] = \frac{1}{2}$

$$c = -10 \textit{ here so } \mu_y = \mu_x + c \Rightarrow \mu_x = \frac{1}{2} + 10 = \underline{\underline{10.}}5 \tag{6.18}$$

$$\sigma_x^2 = \sigma_y^2 = \sum_{i=1}^{20} p_i \left(y_i - \mu_y\right)^2 = \sum_{i=1}^{20} \left(p_i y_i^2 - 2p_i\mu_y y_i + p_i\mu_y^2\right) \tag{6.19}$$

$$\begin{aligned} &= \sum_{i=1}^{20} p_i y_i^2 - 2\mu_y \sum_{i=1}^{20} p_i y_i + \mu_y^2 \sum_{i=1}^{20} p_i \\ &= \sum_{i=1}^{20} p_i y_i^2 - 2\mu_y^2 + \mu_y^2 = \sum_{i=1}^{20} p_i y_i^2 - \mu_y^2 \end{aligned} \tag{6.20}$$

Hence $\sigma_x^2 = \frac{1}{20}\left[2\left(1^2 + 2^2 + \cdots + 9^2\right) + 0^2 + 10^2\right] - \frac{1}{4} = 33.5 - 0.25 = 33.25$

The results of these examples can be generalized into a useful theorem as follows:

Theorem 4 *Given a sample space* $S = \{e_1, e_2, \cdots, e_n\}$ *and a probability distribution* $P = \{p_1, p_2, \cdots, p_n\}$ *and a random variable* X *with values* $\{x_1, x_2, \cdots, x_n\}$*. The expectation value of this system is denoted by* μ *and its variance by* σ^2*. Suppose that* a *is any real constant. Then:*

1. $E\left[(X - a)\right] = \mu - a$

2. $E\left[(X - a)^2\right] = \sigma^2 + (\mu - a)^2$

3. $\sigma^2 = E\left[X^2\right] - \mu^2$

Proof 3 *1. This is exactly the same as the first part of Example 37.9 with c replaced by -a.*

2.

$$\begin{aligned} E\left[(X - a)^2\right] &= \sum_{i=1}^{n} p_i \left(x_i - a\right)^2 \\ &= \sum_{i=1}^{n} p_i \left[(x_i - \mu) + (\mu - a)\right]^2 \end{aligned}$$

$$= \sum^{p_i(x_i-\mu)^2+2(\mu-a)\sum_{i=1}^{n} p_i(x_i-\mu)+(\mu-a)^2\sum_{i=1}^{n} p_i}$$

So noting that $\sum_{i=1}^{n} p_i = 1$ *and* $\sum_{i=1}^{n} p_i\left(x_i - \mu\right) = \sum_{i=1}^{n} p_i x_i - \mu \sum_{i=1}^{n} p_i = \mu - \mu = 0$ *we have* $\underline{\underline{E\left[(X - a)^2\right] = \sigma^2 + (\mu - a)^2}}$

3. This follows at once from part (ii) on setting a = 0. This last result is really only a corollary to the theorem but I want to stress its importance. It is the preferred method for calculating the variance.

Theorem 5 *If we have* $S, P, X,\quad \mu$ *and* σ *defined as in the previous theorem and if* λ *and* a *are any real numbers and we define the new random variable* $Y = \lambda X + a$ *then:*

1. $\mu_y = \lambda\mu + a$

2. $\sigma_y = \lambda\sigma$

(Note that the results of Example 9 follow at once from this theorem if we set $\lambda = 1$*.)*

Proof 4 *1.* $E(\lambda X + a) = \sum^{p_i(\lambda x_i + a)}$

$$= \lambda \sum_{i=1}^{n} p_i x_i + a \sum_{i=1}^{n} p_i = \lambda E(X) + a = \underline{\underline{\lambda\mu + a = \mu_y}} \tag{6.21}$$

2. $E\left[(Y - \mu_y)^2\right] = \sum_{i=1}^{n} p_i \left[\lambda x_i + a - \lambda\mu - a\right]^2$

$$\sum_{i=1}^{n} p_i \left[x_i - \mu\right]^2 \lambda^2 = \lambda^2 \sum_{i=1}^{n} p_i \left[x_i - \mu\right]^2 = \lambda^2 \sigma^2 \tag{6.22}$$

$$\sigma_y = \lambda\sigma \tag{6.23}$$

Example 28 *In an experiment the values of a quantity x and the number of times f that it was observed are given in the following table:*

$$\begin{array}{cllllll} x & 10 & 20 & 30 & 40 & 50 & 60 \\ f & 15 & 13 & 12 & 10 & 9 & 11 \end{array} \tag{6.24}$$

Calculate the mean and standard deviation.

[This example illustrates how the concepts that we are discussing are usually handled in statistics.]

Solution.

The a posteriori probabilities are $\frac{15}{70}, \frac{13}{70}, \frac{12}{70}, \frac{10}{70}, \frac{9}{70}, \frac{11}{70}$.

Let us use $y = \frac{x}{10} - 3$ *as a new variable. (i.e. we guess a mean of 3 for the old variable divided by 10. The point of guessing a mean is that it ensures that the numbers will be kept as small as possible. This can become very important if the data covers thousands of samples as it may well do in real situations.)*

$$\begin{array}{ccccccc} y & -2 & -1 & 0 & 1 & 2 & 3 \\ p & \frac{15}{70} & \frac{13}{70} & \frac{12}{70} & \frac{10}{70} & \frac{9}{70} & \frac{11}{70} \end{array} \tag{6.25}$$

$$\mu_y = \frac{1}{70}\left[-1(15) + (-1)(13) + 0(12) + 1(10) + 2((9) + 3(11)\right] \tag{6.26}$$

$$= \frac{1}{70}\left[-43 + 61\right] = \frac{18}{70} = \frac{9}{35} \tag{6.27}$$

But $\mu_y = \lambda\mu + a \Rightarrow \frac{9}{35} = \frac{1}{10}\mu - 3 \Rightarrow \mu = 10\left[3 + \frac{9}{35}\right] = \frac{1140}{35} = 32.57$

$$\sigma_y^2 = E(Y^2) - \mu_y^2 = \frac{1}{70}\left[15(4) + 13(1) + 12(0) + 10(1) + 9(4) + 9(11)\right] - \left(\frac{9}{35}\right)^2 \tag{6.28}$$

$$\frac{218}{70} - (32.57)^2 = 3.048 \tag{6.29}$$

Thus $\sigma_y = 1.746$ *and* $\sigma_y = \lambda\sigma \Rightarrow \sigma = 17.46$

In the age of scientific calculators and computers such a sophisticated method may seem pointless as you could get the answers at least as quickly using the raw data. However this is a good illustration of important theoretical results.

Example 29 *Two unbiased die are thrown. Find the expectation value of the total score and its standard deviation.*

Solution.

The possible scores with their probabilities are listed below.

$$\begin{array}{ccccccccccc} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ \frac{1}{36} & \frac{2}{36} & \frac{3}{36} & \frac{4}{36} & \frac{5}{36} & \frac{6}{36} & \frac{5}{36} & \frac{4}{36} & \frac{3}{36} & \frac{2}{36} & \frac{1}{36} \end{array} \tag{6.30}$$

Let $Y = X$ *- 7*

This leads to a new table as below:

$$\begin{array}{ccccccccccc} -5 & -4 & -3 & -2 & -1 & 0 & 1 & 2 & 3 & 4 & 5 \\ \frac{1}{36} & \frac{2}{36} & \frac{3}{36} & \frac{4}{36} & \frac{5}{36} & \frac{6}{36} & \frac{5}{36} & \frac{4}{36} & \frac{3}{36} & \frac{2}{36} & \frac{1}{36} \end{array} \tag{6.31}$$

$$\mu_y = \frac{1}{36}\left[-5-8-9-8-5+5+8+9+8+5\right] = 0 \tag{6.32}$$

$$\mu_y = \mu - 7 \Rightarrow \underline{\underline{\mu = 7}} \tag{6.33}$$

$$\sigma_y^2 = E(Y^2) - \mu_y^2 \tag{6.34}$$

$$= \frac{1}{36}\left[25+32+27+16+5+5+16+27+32+25\right] = \frac{210}{36} \tag{6.35}$$

But $\sigma_y = \lambda\sigma = \sigma$ *here* $\Rightarrow \sigma^2 = \frac{35}{6} \Rightarrow \sigma = 2.42$

Example 30 *Suppose an unbiased coin is tossed until a head appears. We could get 0,1,2,.... tails before the first head appears. Calculate the expectation value of the number of tails before we get the first head.*

Solution.

Theoretically we could be spinning the coin for an arbitrarily large number of trials before we got the head. So our outcome space and the number of values of the random variable is infinite. e.g.

H TH TTH TTTH ....

0 1 2 3 ....

p $\frac{1}{2} \quad \left(\frac{1}{2}\right)^2 \quad \left(\frac{1}{2}\right)^3 \quad \left(\frac{1}{2}\right)^4$ ....

Note that $\sum\limits_{i=1}^{\infty} p_i = \frac{a}{1-r} = \frac{1/2}{1-1/2} = 1$

E(No. of tails)$= 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right)^2 + 2\left(\frac{1}{2}\right)^3 + 3\left(\frac{1}{2}\right)^4 \cdots\cdots = \sum\limits_{r=1}^{\infty} r \cdot \frac{1}{2^{r+1}}$

To sum this is interesting.

For $|x| < 1$ *we have that*

$$(1-x)^{-2} = 1 + (-2)(-x) + \frac{(-2)(-3)(-x)^2}{2!} + \frac{(-2)(-3)(-4)(-x)^3}{3!} + \cdots \tag{6.36}$$

$$= 1 + 2x + 3x^2 + 4x^3 + \cdots \tag{6.37}$$

Set x = 1/2 $\begin{array}{l} \left(1-\frac{1}{2}\right)^{-2} = 1 + 2\cdot\frac{1}{2} + 3\cdot\left(\frac{1}{2}\right)^2 + 4\cdot\left(\frac{1}{2}\right)^3 + \cdots \\ \left(\frac{1}{2}\right)^2\left(1-\frac{1}{2}\right)^{-2} = \left(\frac{1}{2}\right)^2 + 2\cdot\left(\frac{1}{2}\right)^3 + 3\cdot\left(\frac{1}{2}\right)^4 + \cdots \end{array}$ *=E(No. of tails)*

$$\textit{E(No. of tails)} = \frac{1}{4}\left(\frac{1}{2}\right)^{-2} = 1 \tag{6.38}$$

Example 31 *A and B alternately throw a die. The game ends in a win for A if he throws a 1 or a 6 or in a win for B if he throws a 2, 3, 4, 5. Find the probability that A wins and given that he wins find the average number of throws that he takes. It is assumed that A throws first.*

Solution.

This game could theoretically go on for ever.

Clearly P(A wins on a particular throw) = 1/3.

P(A wins 1st time) $= \frac{1}{3}$

P(A wins 2nd time) $= \frac{2}{3}\frac{1}{3}\frac{1}{3}$

P(A wins 3rd time) $= \frac{2}{3}\frac{1}{3}\frac{2}{3}\frac{1}{3}\frac{1}{3}$

P(A wins on nth attempt) $= \left(\frac{2}{3}\cdot\frac{1}{3}\right)^{n-1}\cdot\left(\frac{1}{3}\right)$

P(A wins) $= \frac{1}{3}\left[1+\left(\frac{2}{9}\right)+\left(\frac{2}{9}\right)^2+\cdots\cdots\right] = \frac{1}{3}\cdot\frac{1}{1-2/9} = \underline{\underline{\frac{3}{7}}}$

We now need the expected number of throws of A given that A wins. Since it is now given that he wins the probabilities are now modified. e.g.

$$P(X = r \mid A\ wins) = \frac{P(X = r \cap \ A\ wins)}{P(Awins)}. \tag{6.39}$$

This is the probability that A wins on the r th try given that we know that he did win at some point.

Thus all the previous probabilities are multiplied by $\frac{1}{P(A\ wins)} = \frac{1}{3/7} = \frac{7}{3}$.

E(No. of throws given that A wins) $= 1\cdot\frac{1}{3}\cdot\frac{7}{3} + 2\left(\frac{2}{9}\right)\left(\frac{1}{3}\right)\left(\frac{7}{3}\right) + 3\left(\frac{2}{9}\right)^2\left(\frac{1}{3}\right)\left(\frac{7}{3}\right) + \cdots\cdots$

$$= \frac{7}{9}\left[1 + 2\left(\frac{2}{9}\right) + 3\left(\frac{2}{9}\right)^2 + \cdots\cdots\right] \tag{6.40}$$

But from Example 37.13.

$$\begin{array}{l} (1-x)^{-2} = 1 + 2x + 3x^2 + \cdots\cdots \\ 1 + 2\left(\frac{2}{9}\right) + 3\left(\frac{2}{9}\right)^2 + \cdots\cdots = \left(1-\frac{2}{9}\right)^{-2} = \frac{81}{49} \end{array} \tag{6.41}$$

$$\underline{\underline{E(No.\ of\ throws) = \frac{7}{9}\cdot\frac{81}{49} = \frac{9}{7}}} \tag{6.42}$$

6.1.1 Exercises

1. The following table gives the number of children in each of 360 families.

No. of children	0	1	2	3	4	5	6	7	8
No. of families	38	91	108	76	39	5	2	0	1

(6.43)

   Calculate the mean and standard deviation of the number of children per family.

2. X is a random variable with mean μ and standard deviation σ. Find the mean and standard deviation of each of the following random variables: a. -X; b. X + 1; c. 3X-1; d. (X - μ)/σ

3. Calculate the mean and variance for each of the following distributions:

(a)

X	0	1	2	3	4	5
p	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{3}{8}$

(b)

X	1	2	3	4	5	6
p	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$

(c)

X	-3	-2	-1	0	1	2	3	4
p	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{1}{4}$	$\frac{3}{20}$	$\frac{3}{20}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{20}$

(d)

X	50	100	150	200	250
p	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{1}{10}$

4. A cubical die is so weighted that the probability of obtaining any face is proportional to the score showing on that face. Find the mean and variance of the score obtained.

5. If the random variable X can take the value 1,2,3,...,n all values being equally likely calculate the mean and variance of X.

6. A, B and C repeatedly throw a die, A starting, then B, then C then A again and so on. The winner is the first to throw a six. What are their respective chances of winning the game?

7. A throws a pair of unbiased dice, B a pair of dice of which one is unbiased and the other is such that the probability of a six is p. If they throw in turn and the winner is the first to throw a double 6, find p, given that when A has the first throw the game is fair.

8. Two dice are thrown together and the scores added. What is the chance that the total score exceeds 8? Find the mean and standard deviation of the total score. What is the standard deviation of the score for a single die?

9. A card is drawn at random from a standard pack and scores the face value of the card (with ace as 1 and picture cards as 10 each). Find the mean and variance of the score.

   If the card is replaced, the pack well shuffled and a second card drawn, find the probability that the total score for both draws is 12.

10. Two bags contain red and white discs as shown in the table below:

	Red	*White*
BagI	5	15
BagII	10	10

(6.44)

   One of the bags is selected at random and a disc drawn from it proves to be red. If the red discs are now valued at $1 each and the white discs are valueless, what would be a fair price to pay for the remaining discs in the selected bag?

11. (This problem is known as the St. Petersburg Paradox). A coin is spun. If a head is obtained first time you are paid $1; If you get a tail followed by a head you receive $2; for two tails followed by a head $4, the next prize being

$8 and so on. Show that, however much you are prepared to pay to play the game your expected profit will be positive.

Criticize any assumptions you have made and indicate what further knowledge you would require before offering a more realistic "fair price" for the game. If the banker against whom you are playing starts with a capital of $100, what would be a fair price for you to offer him before playing the game?

6.2 Moments

When a set of values has a sufficiently strong central tendency, that is, a tendency to cluster around some particular value, then it may be useful to characterize the set by a few numbers that are related to its *moments*, the sums of integer powers of the values.

Best known is the *mean* of the values x_1, x_2, $\cdots$, x_N,

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{N} x_j \tag{6.45}$$

which estimates the value around which central clustering occurs. Note the use of an overbar to denote the mean; angle brackets are an equally common notation, e.g., $\langle x \rangle$. You should be aware that the mean is not the only available estimator of this quantity, nor is it necessarily the best one. For values drawn from a probability distribution with very broad "tails," the mean may converge poorly, or not at all, as the number of sampled points is increased. Alternative estimators, the *median* and the *mode*, are mentioned at the end of this section.

Having characterized a distribution's central value, one conventionally next characterizes its "width" or "variability" around that value. Here again, more than one measure is available. Most common is the *variance*,

$$Var\left(x_1, x_2, \cdots, x_N\right) = \frac{1}{N-1}\sum_{j=1}^{N}\left(x_j - \overline{x}\right)^2 \tag{6.46}$$

or its square root, the *standard deviation*,

$$\sigma\left(x_1, x_2, \cdots, x_N\right) = \sqrt{Var\left(x_1, x_2, \cdots, x_N\right)} \tag{6.47}$$

Equation 6.46 estimates the mean squared deviation of x from its mean value. There is a long story about why the denominator of 6.46 is $N-1$ instead of N. If you have never heard that story, you may consult any good statistics text. Here we will be content to note that the $N-1$ *should* be changed to N if you are ever in the situation of measuring the variance of a distribution whose mean x is known *a priori* rather than being estimated from the data. (We might also comment that if the difference between N and $N-1$ ever matters to you, then you are probably up to no good anyway - e.g., trying to substantiate a questionable hypothesis with marginal data.)

As the mean depends on the first moment of the data, so do the variance and standard deviation depend on the second moment. It is not uncommon, in real life, to be dealing with a distribution whose second moment does not exist (i.e., is infinite). In this case, the variance or standard deviation is useless as a measure of the data's width around its central value: The values obtained from equations 6.46 or 6.47 will not converge with increased numbers of points, nor show any consistency from data set to data set drawn from the same distribution. This can occur even when the width of the peak looks, by eye, perfectly finite. A more robust estimator of the width is the *average deviation* or *mean absolute deviation*, defined by

$$ADev\left(x_1, x_2, \cdots, x_N\right) = \frac{1}{N}\sum_{j=1}^{N}\left|x_j - \overline{x}\right| \tag{6.48}$$

One often substitutes the sample median x_{med} for x in equation 6.48. For any fixed sample, the median in fact minimizes the mean absolute deviation. Statisticians have historically sniffed at the use of 6.48 instead of 6.46, since the absolute value brackets in 6.48 are ”nonanalytic” and make theorem- proving difficult. In recent years, however, the fashion has changed, and the subject of *robust estimation* (meaning, estimation for broad distributions with significant numbers of ”outlier” points) has become a popular and important one. Higher moments, or statistics involving higher powers of the input data, are almost always less robust than lower moments or statistics that involve only linear sums or (the lowest moment of all) counting.

Figure 6.1: Distributions whose third and fourth moments are significantly different from a normal (Gaussian) distribution. (a) Skewness or third moment. (b) Kurtosis or fourth moment.

That being the case, the *skewness* or *third moment*, and the *kurtosis* or *fourth moment* should be used with caution or, better yet, not at all.

The skewness characterizes the degree of asymmetry of a distribution around its mean. While the mean, standard deviation, and average deviation are *dimensional* quantities, that is, have the same units as the measured quantities x_j, the skewness is conventionally defined in such a way as to make it *nondimensional.* It is a pure number that characterizes only the shape of the distribution. The usual definition is

$$Skew\left(x_1, x_2, \cdots, x_N\right) = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{x_j - \overline{x}}{\sigma}\right]^3 \tag{6.49}$$

where $\sigma = \sigma\left(x_1, x_2, \cdots, x_N\right)$ is the distribution's standard deviation 6.47. A positive value of skewness signifies a distribution with an asymmetric tail extending out towards more positive x; a negative value signifies a distribution whose tail extends out towards more negative x (see Figure 6.1).

Of course, any set of N measured values is likely to give a nonzero value for 6.49, even if the underlying distribution is in fact symmetrical (has zero skewness). For 6.49 to be meaningful, we need to have some idea of *its* standard deviation as an estimator of the skewness of the underlying distribution. Unfortunately, that depends on the shape of the underlying distribution, and rather critically on its tails! For the idealized case of a normal (Gaussian) distribution, the standard deviation of 6.49 is approximately $\sqrt{15/N}$ when $\overline{x}$ is the true mean and $\sqrt{6/N}$ when it is estimated by the sample mean, 6.45. In real life it is good practice to believe in skewnesses only when they are several or many times as large as this.

The kurtosis is also a nondimensional quantity. It measures the relative peakedness or flatness of a distribution. Relative to what? A normal distribution, what else! A distribution with positive kurtosis is termed *leptokurtic*; the outline of the Matterhorn is an example. A distribution with negative kurtosis is termed *platykurtic*; the outline of a loaf of bread is an example. (See Figure 6.1) And, as you no doubt expect, an in-between distribution is termed *mesokurtic.*

The conventional definition of the kurtosis is

$$Kurt\,(x_1, x_2, \cdots, x_N) = \left(\frac{1}{N} \sum_{j=1}^{N} \left[\frac{x_j - \overline{x}}{\sigma} \right]^4 \right) - 3 \tag{6.50}$$

where the -3 term makes the value zero for a normal distribution.

The standard deviation of 6.50 as an estimator of the kurtosis of an underlying normal distribution is $\sqrt{96/N}$ when σ is the true standard deviation, and $\sqrt{24/N}$ when it is the sample estimate 6.47. However, the kurtosis depends on such a high moment that there are many real-life distributions for which the standard deviation of 6.50 as an estimator is effectively infinite.

Calculation of the quantities defined in this section is perfectly straightforward. Many textbooks use the binomial theorem to expand out the definitions into sums of various powers of the data, e.g., the familiar

$$Var\,(x_1, x_2, \cdots, x_N) = \frac{1}{N-1} \left[\left(\sum_{j=1}^{N} x_j^2 \right) - N\overline{x}^2 \right] \approx \overline{x^2} - \overline{x}^2 \tag{6.51}$$

but this can magnify the roundoff error by a large factor and is generally unjustifiable in terms of computing speed. A clever way to minimize roundoff error, especially for large samples, is to use the *corrected two-pass algorithm*: First calculate $\overline{x}$, then calculate $Var\,(x_1, x_2, \cdots, x_N)$ by

$$Var\,(x_1, x_2, \cdots, x_N) = \frac{1}{N-1} \left[\sum_{j=1}^{N} (x_j - \overline{x})^2 - \frac{1}{N} \left(\sum_{j=1}^{N} (x_j - \overline{x}) \right)^2 \right] \tag{6.52}$$

The second sum would be zero if $\overline{x}$ were exact, but otherwise it does a good job of correcting the roundoff error in the first term.

6.2.1 Semi-Invariants

The mean and variance of independent random variables are additive: If x and y are drawn independently from two, possibly different, probability distributions, then

$$\overline{(x+y)} = \overline{x} + \overline{y}, \qquad Var(x+y) = Var(x) + Var(y) \tag{6.53}$$

Higher moments are not, in general, additive. However, certain combinations of them, called *semi-invariants*, are in fact additive. If the centered moments of a distribution are denoted M_k,

$$M_k = \left\langle (x_i - \overline{x})^k \right\rangle \tag{6.54}$$

so that, e.g., $M_2 = Var(x)$, then the first few semi-invariants, denoted I_k are given by

$$I_2 = M_2, \quad I_3 = M_3, \quad I_4 = M_4 - 3M_2^2, \quad I_5 = M_5 - 10M_2M_3, \quad I_6 = M_6 - 15M_2M_4 - 10M_3^2 + 30M_2^3 \tag{6.55}$$

Notice that the skewness and kurtosis, equations 6.49 and 6.50 are simple powers of the semi-invariants,

$$Skew(x) = I_3/I_2^{3/2}, \quad Kurt(x) = I_4/I_2^2 \tag{6.56}$$

A Gaussian distribution has all its semi-invariants higher than I_2 equal to zero. A Poisson distribution has all of its semi-invariants equal to its mean.

6.2.2 Median and Mode

The median of a probability distribution function $p(x)$ is the value x_{med} for which larger and smaller values of x are equally probable:

$$\int\limits_{-\infty}^{x_{med}} p(x)dx = \frac{1}{2} = \int\limits_{x_{med}}^{\infty} p(x)dx \tag{6.57}$$

The median of a distribution is estimated from a sample of values $(x_1, x_2, \cdots, x_N)$ by finding that value x_i which has equal numbers of values above it and below it. Of course, this is not possible when N is even. In that case it is conventional to estimate the median as the mean of the unique *two* central values. If the values $(x_1, x_2, \cdots, x_N)$ are sorted into ascending (or, for that matter, descending) order, then the formula for the median is

$$x_{med} = \begin{cases} x_{(N+1)/2} & N \text{ odd} \\ \frac{1}{2}\left(x_{N/2} + x_{N/2+1}\right) & N \text{ even} \end{cases} \tag{6.58}$$

If a distribution has a strong central tendency, so that most of its area is under a single peak, then the median is an estimator of the central value. It is a more robust estimator than the mean is: The median fails as an estimator only if the area in the tails is large, while the mean fails if the first moment of the tails is large; it is easy to construct examples where the first moment of the tails is large even though their area is negligible.

The *mode* of a probability distribution function $p(x)$ is the value of x where it takes on a maximum value. The mode is useful primarily when there is a single, sharp maximum, in which case it estimates the central value. Occasionally, a distribution will be *bimodal*, with two relative maxima; then one may wish to know the two modes individually. Note that, in such cases, the mean and median are not very useful, since they will give only a "compromise" value between the two peaks.

6.2.3 Summing Up

The mean and variance are two examples of moments of distributions.

Definition 12 *The n^{th}* raw moment *(the moment about zero) of a distribution $P(x)$ is defined by*

$$\mu'_n = \langle x^n \rangle \tag{6.59}$$

where

$$\langle f(x) \rangle = \begin{cases} \sum f(x)P(x) \textit{discrete distribution} \\ \int f(x)P(x)dx \textit{continuous distribution} \end{cases} \tag{6.60}$$

The mean $\mu = \mu_1$. If the moment is taken about a point a, then we have

$$\mu_n(a) = \langle (x-a)^n \rangle = \sum (x-a)^n P(x) \tag{6.61}$$

We note that a distribution is not uniquely specified by its moments. Most commonly the moments are taken about the mean.

Definition 13 *The* central moments *are the moments taken about the mean,*

$$\mu_n = \langle (x-\mu)^n \rangle \tag{6.62}$$

$$= \int (x-a)^n P(x)dx \tag{6.63}$$

Clearly the first central moment is zero. The second central moment is the variance,

$$\mu_2 = \sigma^2 \tag{6.64}$$

6.3 Reading: Chance by Henri Poincaré

6.3.1 Introduction to the Reading

6.3.2 The Paper

Chapter 7

Probability Distributions: The Binomial Distribution

We now look at some special probability distributions to which we will apply our general theory.

7.1 Uniform Probability Distributions.

This is the name given to a probability distribution when all the events of the outcome space are equiprobable. We have already seen examples of this as for instance when an unbiased die is thrown the probability of each face appearing is 1/6.

In general suppose we have an outcome space consisting of n equiprobable events: S $= \{e_1, e_2, \cdots e_n\}$ each with probability $p_i = \frac{1}{n}$ for $1 \leq i \leq n$. We will assign the random variable X to this distribution in the "natural" way i.e. $x_i = i$ for $1 \leq i \leq n$.

The expectation value is: $\mu = E(X) = \sum\limits_{i=1}^{n} p_i x_i = \sum\limits_{i=1}^{n} \frac{1}{n} i = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$

The variance $\sigma^2 = E\left[(X-\mu)^2\right] = E(X^2) - \mu^2$

$$= \overset{i^2\frac{1}{n} - \frac{(n+1)^2}{4} = \frac{1}{n}\cdot\frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}}{\sum} \tag{7.1}$$

We handled many problems on uniform distributions in the last chapter. It is not necessary to use the general theory in such a simple case and we will not do so. This discussion has served to show how we can in fact study the expectation value and variance in a general way when the distribution has some particular property. The next example is more important (and a little more complicated).

7.2 The Binomial Distribution.

We consider situations where the event space has only two possible outcomes. The names Bernoulli trials or binomial experiments are given to such situations. At first this will strike you as even more trivial than the last situation but we are not going to study individual trials but the probability distribution that arises when a large number of trials is carried out. Before giving a formal definition of the binomial distribution we will look at a numerical example.

Consider the families in a large community which have exactly four children. We consider the birth of each child as a single trial resulting in two possible outcomes, a boy or a girl. The birth of all 4 children to make up a family forms a set of four repetitions of the same trial. We will ignore cases of twins triplets and quads. Suppose that for this population we have P(boy born) = 0.51; P(girl born) = 0.49.

Now 4 child families can have 0,1,2,3,4 boys. Consider a family with 3 boys and a girl. This can happen in 4 ways with respect to the order in which the children are born e. g. GBBB, BGBB, BBGB, BBBG. Thus P(3 boy family) $= 4 \times (0.51)^3 \times 0.49$.

If we now use the random variable X for the number of boys in a family we can tabulate the probabilities as follows.

$$\begin{array}{llllll} x & 0 & 1 & 2 & 3 & 4 \\ P(X=x) & (0.49)^4 & 4\,(0.49)^3\,(0.51) & 6\,(0.49)^2\,(0.51)^2 & 4\,(0.49)\,(0.51)^3 & (0.51)^4 \end{array} \tag{7.2}$$

We now note that if we used the binomial theorem to evaluate $(0.49+0.51)^4$ we would get

$$(0.49+0.51)^4 = (0.49)^4 + 4\,(0.49)^3\,(0.51) + 6\,(0.49)^2\,(0.51)^2 + 4\,(0.49)\,(0.51)^3 + (0.51)^4 \tag{7.3}$$

This is where the name binomial distribution really comes from. Also note that $(0.49+0.51)^4 = 1^4 = 1$confirming without detailed calculation that all the probabilities for the various possible numbers of boys in the family add up to 1 as they should. A first obvious advantage of this is that we do not have to actually list he probabilities any more. Using our knowledge of binomial expansions we can write

$$P(X=x) = \left(\begin{array}{c} 4 \\ x \end{array} \right) (0.49)^x\,(0.51)^{4-x} \qquad x = 0,1,2,3,4 \tag{7.4}$$

We will allow the results of this example to serve as a model for our general definition of a binomial distribution.

Definition 14 *A discrete random variable X is said to have the* binomial distribution *if*

$$P(X=x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x\,(1-p)^{n-x} \qquad x = 0,1,2,3\cdots\cdots, n \tag{7.5}$$

where p represents the probability of “success” in a single Bernoulli trial and n is the total number of trials carried out.

It is implicit in this definition that we are dealing with trials which can only have one of two possible outcomes. The terms “success” and “failure” are just convenient to use in a general discussion. The experiment consists of n repetitions of the Bernoulli trial and $P(X=x)$ is the probability of getting x successes. Once the values of n and of p are specified so is the binomial distribution. Thus the notation $B(n,p)$ is often used for a particular binomial distribution.

Example 32 *A fair coin is tossed 5 times. Find the probability of getting three or more heads.*

Solution.

P(H) = P(T) = 1/2. We are working with B(5, 1/2).

We require P(3) + P(4) + P(5)

$$= \begin{pmatrix} 5 \\ 3 \end{pmatrix} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 + \begin{pmatrix} 5 \\ 4 \end{pmatrix} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right) + \begin{pmatrix} 5 \\ 5 \end{pmatrix} \left(\frac{1}{2}\right)^5 \tag{7.6}$$

$$= \frac{1}{2}. \tag{7.7}$$

Of course we have solved problems like this before but we now have some machinery for handling all problems of this type systematically.

Example 33 *7 seeds each with a probability of germinating of 0.2 are planted in each of 80 pots. How many of these pots may be expected to have two or less seedlings?*

Solution.

P(germination) = 0.2, P(non-germination) = 0.8

For a particular pot we are dealing with B(7,0.2) and require P(0) + P(1) + P(2)

$$= (0.8)^7 + \begin{pmatrix} 7 \\ 1 \end{pmatrix} (0.8)^6 (0.2) + \begin{pmatrix} 7 \\ 2 \end{pmatrix} (0.8)^5 (0.2)^2 = \underline{\underline{0.852}} \tag{7.8}$$

This is the probability that a particular pot chosen at random will have two or fewer seedlings. Hence out of the 80 pots we would expect about 80 x 0.852 = 68.16 or say about 68 pots to have fewer than 2 seedlings.

Example 34 *A fair coin is tossed n times.. Given that n is even what is the probability that the coin will fall heads exactly half the number of times. Do the numerical calculation for n = 10, 20, 50, 100.*

Solution.

We are dealing with B(n,1/2). $P\left(\frac{n}{2}\right) = \begin{pmatrix} n \\ \frac{n}{2} \end{pmatrix} \left(\frac{1}{2}\right)^{\frac{n}{2}} \left(\frac{1}{2}\right)^{n-\frac{n}{2}} = \underline{\underline{\begin{pmatrix} n \\ \frac{n}{2} \end{pmatrix} \left(\frac{1}{2}\right)^n}}$

For n = 10 $P(5) = \begin{pmatrix} 10 \\ 5 \end{pmatrix} \left(\frac{1}{2}\right)^{10} = \underline{\underline{0.246}}$

For n = 20 $P(10) = \begin{pmatrix} 20 \\ 10 \end{pmatrix} \left(\frac{1}{2}\right)^{20} = \underline{\underline{0.176}}$

For n = 50 $P(25) = \begin{pmatrix} 50 \\ 25 \end{pmatrix} \left(\frac{1}{2}\right)^{50} = \underline{\underline{0.112}}$

For n = 100 $P(50) = \begin{pmatrix} 100 \\ 50 \end{pmatrix} \left(\frac{1}{2}\right)^{100} = \underline{\underline{0.08}}$

This example shows the need for precision. As n increases the proportion of heads gets closer and closer to 0.5 but the probability that exactly half of the tosses will result in a head gets smaller and smaller.

7.2.1 Exercises

1. A coin is tossed 6 times. What is the probability of obtaining fewer than three heads?
2. In a family of 5 children, what is the probability that there are (a) 2 boys and 3 girls, (b) at least 2 boys. Assume P(boy) = 0.51.
3. 5 seeds with probability of germination 0.3 are planted in each of 60 pots. How many pots can be expected to have exactly three seedlings?
4. Calculate the probability of getting two sixes if four dice are thrown.
5. What is the probability of obtaining two clubs if a card is taken from each of 6 well shuffled packs? Compare this with the probability of two clubs if all six cards are taken from the same pack.
6. What is the probability of obtaining fewer than three ones if twelve dice are tossed.
7. The probability of a shot at goal actually going in is 1/5. What is the probability of two goals being scored in a match in which there are 14 shots at goal?
8. Racing cars are notoriously prone to break down. The probability of a car of a certain team finishing is 0.6. What is the probability that three out of four cars in this one team will not finish.
9. A gun has probability 0.3 of hitting a target and it takes at least three direct hits to destroy it. What is the probability that the target is destroyed if 5 shots are fired?
10. A tetrahedral die has its faces colored red, green, yellow and blue. If a group of 8 such dice are thrown, calculate the probabilities of 0, 1, 2, 3, 4 red faces being seen. How many times would you expect to see only 2 red faces if the experiment of throwing the 8 dice was repeated 200 times?

7.3 Mean and Variance of a Binomial Distribution.

It is of course quite possible to calculate the mean and variance of any particular binomial distribution by the methods of the last chapter. But this would obviously lead to very cumbersome calculations for situations where an experiment was repeated a very large number of times. Even a number as small as 10 would involve quite a lot of button pushing on the calculator. It turns out that if we grasp the nettle of attempting the calculation for the general distribution B(n,p) we will be able to obtain simple formulae which reduce the calculation of the mean and variance of any binomial distribution to triviality. Admittedly we will have to work a little to get these formulae.

Theorem 6 *For any binomial distribution B(n.p) we can calculate the expectation value μ from the formula $\mu = E(X) = np$.*

Proof 5

$$E(X) = 0.P(0) + 1.P(1) + 2.P(2) + \cdots\cdots + n.P(n) \tag{7.9}$$

$$= \sum_{i=0}^{n} i \left(\begin{array}{c} n \\ i \end{array} \right) p^i \left(1-p\right)^{n-i} = \sum_{i=0}^{n} \frac{in!}{i!\left(n-i\right)!} p^i \left(1-p\right)^{n-i} \tag{7.10}$$

$$= \sum_{i=1}^{n} \frac{n!}{\left(i-1\right)!\left(n-i\right)!} p^i q^{n-i} \text{ where } q = 1-p \text{ for convenience} \tag{7.11}$$

$$= np \sum^{\frac{(n-1)!}{(i-1)!(n-i)!} p^{i-1} q^{n-i}} \tag{7.12}$$

$$= np \left[q^{n-1} + \left(\begin{array}{c} n-1 \\ 1 \end{array} \right) q^{n-2} p + \left(\begin{array}{c} n-1 \\ 2 \end{array} \right) q^{n-3} p^2 + \cdots\cdots + \left(\begin{array}{c} n-1 \\ n-1 \end{array} \right) p^{n-1} \right] \tag{7.13}$$

$$= np \left(q+p\right)^{n-1} = np \tag{7.14}$$

$$\mu = E(X) = np \tag{7.15}$$

Theorem 7 *The variance σ^2 of the general binomial distribution B(n,p) is given by the formula $\sigma^2 = np\left(1-p\right)$.*

Proof 6 *Again let 1 - p = q for convenience.*

$$\sigma^2(X) = E(X^2) - \mu^2 \tag{7.16}$$

$$= E\left[X(X-1)\right] + E(X) - \mu^2 = E\left[X(X-1)\right] + \mu - \mu^2 \tag{7.17}$$

Since μ is known from the preceding theorem we only have to find E[X(X - 1)].

$$E\left[X(X-1)\right] = 0(-1)P(0) + 1.0P(1) + 2.1P(2) + \cdots\cdots + n(n-1)P(n) \tag{7.18}$$

$$= \sum_{i=0}^{n} i\left(i-1\right) \left(\begin{array}{c} n \\ i \end{array} \right) p^i q^{n-i} = \sum_{i=2}^{n} \frac{i\left(i-1\right)n!}{i!\left(n-i\right)!} p^i q^{n-i} \tag{7.19}$$

$$= n\left(n-1\right)p^2 \sum_{i=2}^{n} \frac{\left(n-2\right)!}{\left(i-2\right)!\left(n-i\right)!} p^{i-2} q^{(n-2)-(i-2)} \tag{7.20}$$

$$= n\left(n-1\right)p^2\left(p+q\right)^{n-2} = n\left(n-1\right)p^2 \tag{7.21}$$

$$\sigma^2 = n\left(n-1\right)p^2 + np - n^2p^2 = np\left(1-p\right) \tag{7.22}$$

From this we have at once that the standard deviation of the binomial distribution B(n,p) is $\sqrt{np\left(1-p\right)}$.

Example 35 *A machine produces resistors and 2% of them are defective. They are packed into boxes of 24. Find the mean and standard deviation of the number of defectives per box.*

Solution.

[You may be interested to note that in statistics this situation would be handled by calling a defective resistor a "success".]

For us P(defective) = 0.02 = p

P(perfect) = 0.98 = 1 - p on trying any resistor at random. Since there are only the two possible outcomes perfect and defective we have a binomial distribution for the multiple trials (i.e. 24 per box). i.e. we have a case of B(24, 0.02).

Thus the mean number of defectives in a box of 24 is np = 24 x 0.02 = 0.48.

The standard deviation $= \sqrt{np\,(1-p)} = \sqrt{24\,(0.02)\,(0.98)} = 0.686$

[There is an interesting rule for using the standard deviation of a binomial distribution with a large number of trials and not to small a probability p. To be more precise we require np ¿ 10. It runs as follows.

68% of the trials will lie within 1 standard deviation of the mean.

95% of the trials will lie within 2 standard deviations of the mean.

Only on very rare occasions will we get a result more than 3 standard deviations from the mean. This result can in fact be proved although we will not attempt to do that here.

The condition np ¿ 10 is imperative. For Example 38.4 with np = 24 x 0.02 = 0.48 ¡¡ 10 this result would not apply.

But suppose we toss a fair coin 100 times and consider getting a head a success. Here np = 100 x 0.5 = 50 ¿ 10, $\mu = 50$ and $\sigma = 5$.

Thus we can expect that if we have the patience to toss a coin 100 times on a large number of occasions about 68% of the experiments should produce a number of heads between 45 and 55 and 95% of the experiments will produce a result with between 40 and 60 heads. It will be a rarity (in practice never) to get values less than 35 or greater than 65.]

Example 36 *A mass produced piece of plastic is such that 4% of any production run is defective. Discuss the approximate distribution of defectives to be found in boxes of 500.*

Solution.

We are dealing with a case of B(500, 0.04).

$$\begin{aligned} \mu &= np = 500(0.04) = 20. \\ \sigma &= \sqrt{np\,(1-p)} = \sqrt{19.2} = 4.4 \end{aligned} \tag{7.23}$$

We can treat 68%, 95% etc. as probabilities i.e.

$$P(15.6 < x < 24.6) \approx 0.68 \textit{ (say 16 - 24)} \tag{7.24}$$

$$P(11.2 < x < 29) \approx 0.95 \textit{ (say 12 - 29)} \tag{7.25}$$

$$P(6.6 < x < 33.4) \approx 1 \textit{ (say 7 - 33)} \tag{7.26}$$

From this we can see that only about 1 box in 20 would have more than 29 defectives. A manufacturer studying this data might decide that it would be cheaper to throw in *an extra 30 items per box and sell on the basis that each box will contain at least 500 perfects rather than to set up elaborate testing procedures to find and replace the defectives in the output.*

Example 37 *A very large number of balls are in a bag one eighth of them being black and the rest white. Twelve balls are drawn at random. Find*

1. *P(3B and 9W)*
2. *P(at least 3B)*
3. *E(number of black balls)*
4. *the most likely number of black balls in the sample.*

Solution.

We neglect the change in the probabilities as the balls are drawn since there are a very large number of balls in the bag. We consider a black ball a success. P(drawing a black ball) = 1/8 and we consider B(12, 1/8)

1. *P(3B and 9W)* $= \left(\begin{array}{c} 12 \\ 3 \end{array} \right) \left(\frac{1}{8}\right)^3 \left(\frac{7}{8}\right)^7 \approx \underline{\underline{0.129}}$

2. *P(at least 3B) = 1 - P(0B) - P(1B) - P(2B)*

$$= 1 - \left(\frac{7}{8}\right)^{12} - \left(\begin{array}{c} 12 \\ 1 \end{array} \right) \left(\frac{1}{8}\right) \left(\frac{7}{8}\right)^{11} - \left(\begin{array}{c} 12 \\ 2 \end{array} \right) \left(\frac{1}{8}\right)^2 \left(\frac{7}{8}\right)^{10} = \underline{\underline{0.182}} \quad (7.27)$$

3. *E(number of black balls) = np = 12 x (1/8) = 1.5*

4. *P(r black balls)* $= \left(\begin{array}{c} 12 \\ r \end{array} \right) \left(\frac{1}{8}\right)^r \left(\frac{7}{8}\right)^{12-r}$

   P(r + 1 black balls) $= \left(\begin{array}{c} 12 \\ r+1 \end{array} \right) \left(\frac{1}{8}\right)^{r+1} \left(\frac{7}{8}\right)^{12-r-1}$

$$\frac{\textit{P(r + 1 black)}}{\textit{P(r black)}} = \frac{\frac{12!}{(r+1)!(11-r)!} \left(\frac{1}{8}\right)^{r+1} \left(\frac{7}{8}\right)^{11-r}}{\frac{12!}{r!(12-r)!} \left(\frac{1}{8}\right)^{r} \left(\frac{7}{8}\right)^{12-r}} \quad (7.28)$$

   $= \frac{12-r}{r+1} \left(\frac{1}{8}\right) \left(\frac{8}{7}\right) = \frac{12-r}{7(r+1)}$

   From this we see that P(r + 1) ¿ P(r) provided

$$\frac{12-r}{7r+7} > 1 \Rightarrow 12 - r > 7r + 7 \Rightarrow r < \frac{5}{8} \quad (7.29)$$

   i.e. P(1) ¿ P(0) but after that the probabilities decrease. Thus the most likely number of black balls is 1.

7.3.1 Exercises

1. What is the mean and standard deviation of the number of ones if 3 dice are thrown repeatedly?

2. What is the mean and standard deviation of the number of heads if 5 coins are tossed repeatedly?

3. Find the mean and standard deviation of a binomial distribution in which (a) n = 20, p = 0.4 (b) n = 50, p = 0.7

4. A binomial distribution has mean 6 and standard deviation 2. Calculate n and p.

5. A binomial distribution has mean 18 and standard deviation 3. Calculate the probability of 18 successes.

6. In the manufacture of silicon chips P(success) = 0.4. If the chips are made in batches of 80, what is the expected number of good chips per batch? What is the lowest number of defectives per batch likely in thousands of batches?

7. A binomial distribution has n = 36. If its variance is 8 calculate the possible values of its mean.

8. A manufacturing process makes components with a 94% success rate. The components are packed 400 in each box. What is the mean of the number of defectives in each box and within what range would you expect 95% of the results to fall?

9. A multiple choice examination paper gives 5 possible answers to each of 60 questions. What is the likely highest score among candidates who just guess?

10. A population of bacteria have a feature Φ which occurs in 2% of individuals. A series of equally sized samples is taken and the average number with property Φ in each sample is 30. Calculate the approximate sample size. Is it likely that any sample will contain fewer than 5 bacteria with this property?

11. In a multiple choice examination there are 4 possible answers to each of 60 questions. Is it likely that a candidate who just guesses could score over 30%?

12. In a quality control laboratory, samples of size 60 were examined and the number of defectives counted. Over thousands of trials, the number of defectives never exceeded 14 nor was less than 2. Assuming unchanged probability over the testing period, what was the approximate percentage defective?

13. A coin is tossed 4 times. Find the probability that heads appear

    (a) at the first two tosses followed by two tails.

    (b) just twice in the four throws.

    (c) at least twice in the four throws.

14. From a packet containing a large number of seeds, 40% of which are advertised to give red flowers and the others white, 10 plants are produced. What is the probability

    (a) that all the plants have red flowers

    (b) that all the plants have white flowers

    (c) that half the plants have red flowers and half white?

15. (a) In a trial eight coins are tossed together. In 100 such trials how many times should one expect to get three heads and five tails?

    (b) If 8% of articles in a large consignment are defective what is the chance that a sample of 30 articles will contain fewer than three defectives?

16. Nine unbiased dice are thrown. Find P(r) the probability that r sixes appear, and hence determine the value of P(r + 1)/P(r). Find

    (a) the expected number of sixes

    (b) the most likely number of sixes

    (c) the probability of getting more than one six.

17. Playing a certain "one armed bandit" which is advertised "to increase your money tenfold" costs 5 pesos a turn. The player is returned 50 pesos if more than eight balls out of a total of ten drop in a specified slot. The chance of any one ball dropping in this slot is p. Determine the chance of winning in a given turn and for p = 0.65 calculate the mean profit made by the machine on 500 turns. Evaluate the proportion of losing turns in which the player comes within on or two balls of winning for the case of p = 0.65.

7.4 Reading: Mathematics of Population and Food by Thomas Robert Malthus

7.4.1 Introduction to the Reading

7.4.2 The Paper

Chapter 8

Probability Distributions: The Poisson Distribution

8.1 The Poisson Distribution

In our work on normal distributions we were at pains to demonstrate real situations which led to a normal distribution. We also described how Gauss developed the mathematical pdf appropriate to a normal distribution by finding the correct type of function to fit experimental data. Of course not all situations will lead to a normal distribution. Not all situations will fit into any predetermined pattern and will have to be dealt with on their merits by constructing a pdf from the experimental data. There is however one other type of situation, called the Poisson distribution, that arises often enough to have been studied in detail by Simeon Denis Poisson (1781 - 1840). [He was a theoretical physicist and has a much greater claim to fame in the partial differential equation bearing his name.]

Definition 15 *Given an infinite, but discrete outcome space* $\{e_0, e_1, \cdots e_n, \cdots\}$ *with corresponding random variable* $X = \{0, 1, 2, \cdots n, \cdots\}$ *so that* $X_i = i$ *for* $0 \leq i < \infty$ *the* Poisson probability distribution *is defined by* $P(e_r) = e^{-a}\frac{a^r}{r!}$ *where* $a > 0$ *is a constant.*

This is a discrete distribution and in a course devoted entirely to probability and statistics would be discussed immediately after the binomial distribution. Each $a > 0$ gives a different Poisson distribution. The precise significance of the a will appear shortly. This really is a probability distribution since

$$\sum_{r=0}^{\infty} P(e_r) = \sum^{e^{-a}\frac{a^r}{r!}=e^{-a}\sum^{\frac{a^r}{r!}=e^{-a}e^{a}=1}} \tag{8.1}$$

The usual kinds of example used to introduce and justify the Poisson distribution run as follows. Consider a long distance telephone operator in a small town. It is found by observation over a three month period that the average number of long distance calls that she has to put through in the hour from 2 am to 3 am is 3 calls. We will assume that the time needed to make a connection is negligible. If we divide the hour up into a large number of equal short time periods, let us say 720, each of duration 5 seconds for definiteness and also assume that the requests for such calls arrive randomly then there is a probability of 3/720 = 1/240 of a connection being made in any one of the time intervals. [What further assumption(s) is implicit in this modelling?] We now pose the problem: on a given morning what is the probability of the operator handling 0 calls, 1 call, 2 calls etc. in the hour from 2 am to 3 am.?

[The number of calls is artificially small you may think. This helps us visualize what is going on in the argument that follows. Also larger numbers introduce some problems of computation which have nothing to do with the theory underpinning the Poisson distribution and which are best discussed after we know what the distribution is all about.]

We can get an approximate solution (and a very good one) by applying the binomial distribution as follows. This will give us some confidence when we try to use the Poisson distribution.

Since we have divided the one hour interval into 720 subintervals of 5 sec each we can consider the problem as making 720 successive Bernoulli trials of the form "a call arrives or does not arrive in the 5 sec period". The probability of a call arriving in such a period is 1/240 and so we are studying $B\left(720, \frac{1}{240}\right)$.

Using this model:

$$P(0) = \binom{720}{0}\left(\frac{1}{240}\right)^0\left(1-\frac{1}{240}\right)^{720} = 0.04948 \tag{8.2}$$

$$P(1) = \binom{720}{1}\left(\frac{1}{240}\right)^1\left(1-\frac{1}{240}\right)^{719} = 0.1484 \tag{8.3}$$

$$P(2) = \binom{720}{2}\left(\frac{1}{240}\right)^2\left(1-\frac{1}{240}\right)^{718} = 0.2242 \tag{8.4}$$

$$P(3) = \binom{720}{3}\left(\frac{1}{240}\right)^3\left(1-\frac{1}{240}\right)^{717} = 0.2245 \tag{8.5}$$

$$P(4) = \binom{720}{4}\left(\frac{1}{240}\right)^4\left(1-\frac{1}{240}\right)^{716} = 0.1684 \tag{8.6}$$

(Note that before the age of electronic calculators these calculations would have been somewhat tiresome.)

Now suppose we set $a = 3$ (which happens to be the mean number of calls received during this hour!!) in the Poisson distribution and calculate the corresponding probabilities.

$$P(r) = e^{-3}\frac{3^r}{r!} \tag{8.7}$$

$$P(0) = e^{-3} = 0.04979 \tag{8.8}$$

$$P(1) = e^{-3}\frac{3}{1} = 0.1494 \tag{8.9}$$

$$P(2) = e^{-3}\frac{3^2}{2!} = 0.2246 \tag{8.10}$$

$$P(3) = e^{-3}\frac{3^3}{3!} = 0.2240 \tag{8.11}$$

$$P(4) = e^{-3}\frac{3^4}{4!} = 0.1680 \tag{8.12}$$

Note how the results obtained compare very closely with the ones obtained using $B\left(720, \frac{1}{240}\right)$ (and with much simpler calculation.)

It can be proved theoretically that as $n \to \infty, \quad p \to 0$ in such a way that $np = a$ (i.e. the a = 3 of our example) then the correspondence between the binomial distribution $B(n, p))$ and the Poisson distribution $e^{-a}\frac{a^r}{r!}$ gets better and better, the discrepancies tending to zero. [This is a result of pure math concerning the limits of the algebraic expressions. It does not give evidence for or against either distribution being a good model for a particular real situation. That must be decided by other means]

However in our example the lengths of the time subintervals into which our hour is divided soon becomes so small that we feel justified in saying $B\left(10000, \frac{3}{10000}\right)$ will be an almost perfect model for our problem. Moreover the results obtained from $B\left(10000, \frac{3}{10000}\right)$will be for all practical purposes indistinguishable from those obtained by using the much easier to calculate $e^{-a}\frac{a^r}{r!}$. For problems of this nature we will from now on use the Poisson distribution without further ado.

Example 38 *Traffic accidents are reported in a certain town at an average rate of 4 per week. Estimate the probabilities:*

1. *of a given week being accident free*
2. *of there being 3 or fewer accidents in a given week.*
3. *of there being more than 4 accidents in a given week.*

Solution.

Since we could clearly split the week into a very large number of subintervals with a very low probability of an accident happening in any one subinterval a Poisson distribution with a = 4 seems reasonable.

$$P(0) = e^{-4}\frac{4^0}{0!} = \underline{\underline{0.018}} \tag{8.13}$$

$$P(r \le 3) = e^{-4}\left(1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!}\right) = 0.433 \tag{8.14}$$

$$P(r > 4) = 1 - e^{-4}\left(1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!} + \frac{4^4}{4!}\right) = 0.371 \tag{8.15}$$

Theorem 8 *The mean and variance of the Poisson distribution $e^{-a}\frac{a^r}{r!}$ are both a.*

Proof 7

$$\mu = E(X) = \sum_{r=0}^{\infty} r\frac{e^{-a}a^r}{r!} = ae^{-a}\sum_{r=1}^{\infty}\frac{a^{r-1}}{(r-1)!} \tag{8.16}$$

$$= ae^{-a}\left[1 + a + \frac{a^2}{2!} + \cdots\cdots\right] = ae^{-a}e^{a} = \underline{\underline{a}} \tag{8.17}$$

$$\sigma^2 = E\left(X^2\right) - \mu^2 = \sum_{r=0}^{\infty} r^2 \frac{e^{-a}a^r}{r!} - a^2 = e^{-a}a \sum_{r=1}^{\infty} \frac{ra^{r-1}}{(r-1)!} - a^2 \tag{8.18}$$

$$= e^{-a}a\left[1 + 2a + \frac{3a^2}{2!} + \frac{4a^3}{3!} + \cdots\cdots\right] - a^2 \tag{8.19}$$

$$= e^{-a}a\left[\frac{d}{da}\left\{a + a^2 + \frac{a^3}{2!} + \frac{a^4}{3!} + \cdots\cdots\right\}\right] - a^2 \tag{8.20}$$

$$= e^{-a}a\left[\frac{d}{da}a\left\{1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \cdots\cdots\right\}\right] - a^2 \tag{8.21}$$

$$= e^{-a}a\left(\frac{d}{da}ae^a\right) - a^2 = e^{-a}a\left[e^a + ae^a\right] - a^2 = \underline{\underline{a}} \tag{8.22}$$

Thus $\underline{\underline{\mu = \sigma^2 = a}}$

Example 39 *Cars on a highway pass under a bridge at an average rate of 280 per hour. What number is expected to pass under the bridge in 1.5 minutes? What is the probability that this expected number does in fact pass under the bridge in a 1.5 minute period?*

Solution.

We would assume from the given quantities that on average $280 \div \frac{1.5}{60} = 7$ *cars pass under the bridge in a 1.5 minute period. Hence the number of cars actually passing under the bridge in a 1.5 minute period should be a Poisson distribution with mean 7 i.e.* $P(r) = \frac{e^{-7}7^r}{r!}$ *Pr(7 cars pass)* $= \frac{e^{-7}7^7}{7!} \approx 0.149$.

Example 40 *The proportion of car headlight bulbs which are faulty is 0.5%. If they are packed in boxes of 250 what is the probability that the box will contain at least one defective bulb?*

Solution.

This is actually a binomial distribution with mean or expected value of 1.25 defective bulbs. However by analogy with our first example, splitting the box 250 "bulb intervals" tells us that $P(r) = \frac{e^{1.25}(1.25)^r}{r!}$ *ought to be a good approximation to the actual distribution. [Good because 250 is already a relatively large number and* $p = \frac{1}{200}$ *is already small. But it is only an approximation since we cannot go to the limit. We can divide time intervals as finely as we like but not light bulbs.] Using this approximation* $P(0) = e^{-1.25}$*and Pr(at least one defective)* $= 1 - e^{-1.25} = 0.713$

This technique is often used on binomial distributions with large numbers of trials and small probabilities.

Example 41 *Calls come into a telephone exchange at an average rate of 2.5 per minute. Only 4 calls per minute can be handled. Find the probability that a particular caller will have to be put on hold.*

Solution.

The number of calls arriving per minute form a Poisson distribution with mean 2.5. A caller cannot be serviced whenever 5 or more calls arrive in a minute. So we require

$$P(5) + P(6) + P(7) + \cdots\cdots = 1 - P(0) - P(1) - P(2) - P(3) - P(4) \tag{8.23}$$

Using $P(r) = \frac{e^{-2.5}(2.5)^r}{r!}$ *we find Pr(caller put on hold) =0.109*

Let us now look at an example arising out of an experimental situation.

Example 42 *Quetelet and von Bortkiewicz collected data on the number of men in the Prussian army who were killed by being kicked by their horses whilst they were grooming them. These two devoted savants conducted this experiment over a 20 year period in the late 19th C. They studied 10 separate cavalry corps (each containing the same number of men and horses) over this period. They tabulated deaths from horse kicks per corp per year. With 10 corps for 20 years this gave 200 observations. They tabulated these as follows:*

Deaths in a corp in a year	0	1	2	3	4
Frequency	109	65	22	3	1

From this data we find a total of 65 + 2x22 + 3x3 + 4 = 122 deaths.

Mean number of deaths/corp/year = 122/200 = 0.61

$$\sigma^2 = E\left(X^2\right) - (0.61)^2 = \frac{65}{200}(1)^2 + \frac{22}{200}(2)^2 + \frac{3}{200}(3)^2 + \frac{1}{200}(4)^2 - (0.61)^2 = \underline{\underline{0.607}} \qquad (8.24)$$

Recalling that for a Poisson distribution$\sigma^2 = \mu$ *we are tempted to test the data for the Poisson distribution* $P(r) = \frac{e^{-0.61}(0.61)^r}{r!}$ $P(0) = e^{-0.61} = 0.5434$.*For the 200 corp years this gives an expected number of corp years with zero deaths of 200 x 0.5434 = 108.7 (109 in the table)*

$$P(1) = \frac{e^{-0.61}(0.61)}{1} = 0.3314 \Rightarrow 200 \times 0.3314 = 66.2 \qquad (8.25)$$

$$P(2) = \frac{e^{-0.61}(0.61)^2}{2} = 0.1011 \Rightarrow 200 \times 0.1011 = 20.2 \qquad (8.26)$$

$$P(3) = \frac{e^{-0.61}(0.61)^3}{3!} = 0.02056 \Rightarrow 200 \times 0.02056 = 4 \qquad (8.27)$$

$P(4) = \frac{e^{-0.61}(0.61)^4}{4!} = 0.003135 \Rightarrow 200 \times 0.003135 = 0.6$*(1 in the table).*

The agreement is quite remarkable. Ill disposed people have been known to assert that Q and von B bribed the horses with sugar lumps.

There is probably a Ph.D. in statistics awaiting the person who carries out a similar piece of research on the probability of a lion tamer being eaten by his client.

It is certainly common in experimental statistics to try to use a Poisson distribution whenever $\mu \approx \sigma^2$.

Example 43 *In a large unnamed capital city there are on average 1000 traffic accidents per week. What is the probability:*

(i) that there are 800 accidents in a given week.

(ii) that there are 1500 accidents in a give week.

Solution.

We can argue that the Poisson distribution is still valid for the time interval and hence the p in B(n,p) can be made as small as we please.

So $P\left(800\right) = \frac{e^{-1000}(1000)^{800}}{800!}$ *which is just fine until we try to calculate it!!*
It can be done as follows:

$$\ln P(800) = \ln\left(e^{-1000}\right) + \ln 1000^{800} - \ln 800! \tag{8.28}$$

$$= -1000 + 5526.20422 - \ln 800! \tag{8.29}$$

Fortunately at this point we recall Stirling's asymptotic formula for n!
i.e. $n! \approx n^{n+\frac{1}{2}}\sqrt{2\pi}e^{-n}$

[For n = 10 this gives 3.599×10^6 *and in fact 10!* $= 3.629\times10^6$*. By the time we get to 800! the approximation can be taken as perfect for normal computational purposes. To prove this result is quite advanced. A new continuous function is constructed which has the factorial values for positive integer values of the independent variable. This is called the gamma function. It is defined by an integral and is continuous and differentiable as many times as we please. This means that all the techniques of real and complex variable analysis can be applied to it.]*

Using Stirling's formula we find $\ln 800! = 800.5 \ln 800 + 0.5 \ln 2\pi - 800 = 4550.1127$. *Thus* $\ln P(800) = -23.91$ *which give a totally negligible probability. You can use the same techniques to find* $P(1000) = 0.0126157$ *which although small is not negligible. These figures are small because we are asking for the probability of a single precise value. It would make more sense to ask for the probability that there are between 900 and 1100 accidents in a week. This would best be handled by building a cumulative distribution table as is done for the normal distribution.*

8.1.1 Exercises

1. An electrical component is packed in boxes of 100. If the proportion of defectives is 2%, calculate the proportion of boxes with 2 or less defectives in them.

2. If telephone calls come into an exchange at an average rate of 70 per hour, find the probability of there being 0,1,2 calls in a period of 2 minutes. What is the probability that there are no calls in a period of 7 minutes.

3. If computer malfunctions occur on average once in every 30 hours of operation, calculate the probability that there will be 2 malfunctions in an 8 hour shift.

4. A mass-produced article is packed in boxes of 50. If the expected proportion of defectives produced by the machine is 3% find the most likely number of defective in each box.

5. A small car hire firm has 6 cars which it hires out by the day. The demand for these cars follows a Poisson distribution with mean 3.5. Calculate the expected number of days per year when (a) no cars are required, (b) all 6 are in use but no customer is turned away, (c) some customers are turned away.

6. The average number of cars crossing a narrow bridge is 130 per hour. If more than 5 cars try to cross within 1.5 minutes a queue forms. What is the probability that a queue forms.

7. A car manufacturer gives a guarantee of 1 year and finds that on average 1 car in 14 needs attention during this period. What is the probability that a dealer for these cars will have to give attention to exactly 4 out of the 75 cars that he has sold? What is the probability that more than 4 will need attention?

8. A quantity of radioactive material emits 400 particles per hour. Calculate the expected number of 10 second intervals in a period of 10 hours when 0,1,2,3,4 particles are emitted.

9. In a warehouse, 6 spare parts of a particular type are issued each week on average. Find the number of spare parts which need to be kept in stock so that there is a greater than 95% chance that all demands in a week can be met.

10. A time traveller watching the World Cup soccer tournament in 2020 counted the total goals scored in each match. i.e. if England lost 10-1 to Andorra he counted 11. His total data is collected in the table below. Fit a theoretical Poisson distribution to this information and comment on the closeness of the fit.

Total goals	0	1	2	3	4	5
Number of matches	24	26	11	11	1	1

11. After World War II, an area of South London was divided into 576 squares each of 0.25 square mile area. The number of flying bomb hits in each square was recorded. The results are given in the table below. Compare this with a Poisson model with the same mean.

Number of hits	0	1	2	3	4	$\geq$ 5
Number of squares	229	211	93	35	7	1

12. A book has errors on 403 out of 480 pages. Assuming that the errors follow a Poisson distribution, find the expected number of pages with just one error.

13. Over a series of fixed time intervals a quantity of radioactive material was found to emit one particle in 95 intervals and 2 in 141 intervals. Assuming a Poisson distribution, calculate the average number of particles emitted per interval and the total number of intervals. How many intervals would have had no particles recorded?

14. Calculate the possible means of a Poisson distribution where P(1), P(2), P(3) are in arithmetic progression.

15. Show that no Poisson distribution exists in which P(0), P(1) and P(2) are in geometric progression. Show that in general no three consecutive terms of a Poisson distribution are in a geometric progression.

16. Find the possible values for the mean and standard deviation of a Poisson distribution in which P(2) = 0.25.

17. The road accidents in a certain area occur at an average rate of one per two days. Calculate the probability of 0,1,2,3,4,5,6 accidents per week in this area. What is the most likely number of accidents per week? How many days in a week are expected to be free of accidents?

18. Explain briefly what is meant by a Poisson distribution and show that for such a distribution the mean is equal to the variance. In a bakery 3600 cherries are added to a mixture which is later divided up to make 1200 small cakes.

    (a) Find the average number of cherries per cake.

    (b) Assuming that the number of cherries in each cake follows a Poisson distribution estimate the number of cakes which will be without a cherry and the number with 5 or more cherries.

19. The average proportion of bad eggs in an egg packing station is 1 in 2000. The eggs are packed in boxes containing 6 eggs each.

    (a) Evaluate the probability that a box contains exactly one bad egg.

    (b) A housewife complains if she obtains two or more boxes with one bad egg each per 100 boxes bought. What is the probability that she complains?

8.2 Reading: Statistics of Deadly Quarrels by Lewis Fry Richardson

8.2.1 Introduction to the Reading

8.2.2 The Paper

Chapter 9

Probability Distributions: The Normal Distribution 1

9.1 Probability Distribution Functions

Up to now our discussion of probability has been confined to problems involving discrete outcome spaces. This means that the possible elementary events of our outcome spaces are clearly separated from each other. It is easy to see whether a coin falls heads or tails and which numbered face of a die is uppermost. With the die we cannot get readings of 1.1, 2.345 etc.

However, it is possible to have sample spaces where this is not the case. For example, if we have a line AB 10cm long and choose a point P on it at random, we can take the distance AP as a random variable and all real numbers between 0 and 10 inclusive are possible. At least theoretically and for the purpose of constructing mathematical models. In actual practice we can only measure the distance AP to a given accuracy and so all we can really do is to split the line into a very large number (several million perhaps if the super accurate methods available in modern technology are used) of very small intervals. We could then say with certainty into which one of these intervals our random point fell but to say whereabouts in that interval it is to be found is impossible as it would mean achieving a higher standard of precision than our equipment is capable of. Similar arguments apply to all "real" world situations involving continuous variables so at first sight it would appear that studying a probability distribution for a random variable such as

$$X = \{x \in R : 0 \leq x \leq 10\} \tag{9.1}$$

is only an interesting exercise in pure mathematics. This is not so however. When very large numbers of very close values are involved the methods of integral calculus can be used if the variable is assumed to be continuous and these methods are much easier than handling very large discrete summations. Moreover the results will be perfectly accurate in the sense that using integration rather than summation will not make our errors any bigger than they would be anyway because of the limitations of our measuring apparatus. This statement needs more justification than we can give it in this course, but if you recall that integrals are defined as limits of finite sums, you will see that it is at least plausible that treating random variables with a very large number of closely packed values as continuous and using the methods of integration will give excellent approximations to the real world answers. Let us now put these ideas into a more precise mathematical form.

To do this we will continue with our analysis of choosing a point P at random on the line AB which is 10cm long and using this distance AP as the random variable x.

It is reasonable to assume that all points on the line are equiprobable in this case. Suppose we assign the value ε ¿ 0 to the probability of a particular precise point (say the midpoint of AB for definiteness) being chosen. We know that the probability of some point on the line being chosen is 1. But if we attempt to calculate this probability from the above assumptions we arrive at the result $1 = \varepsilon$ x (the number of points on the line)!! Since this is obvious nonsense we are forced to conclude that the attempt to assign probabilities in the manner that we have used so far is futile.

For the rest of this discussion refer to Figure 39.1.

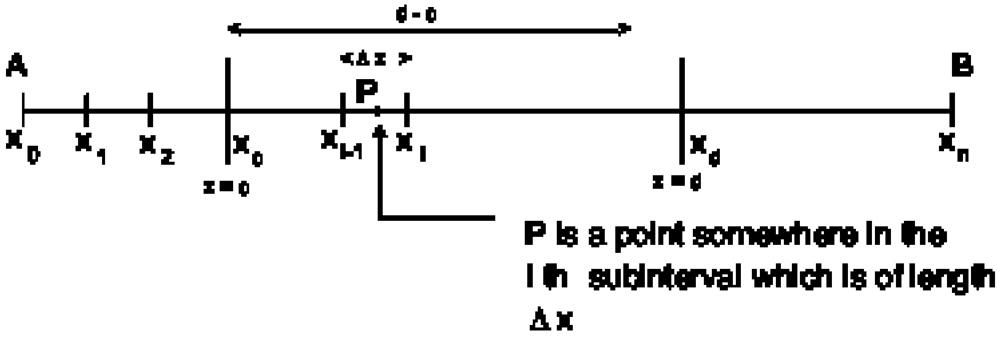

d - c
<Δx >
A
P
B
x_0
x_1
x_2
x_c
x_{i-1}
x_i
x_d
x_n
x = c
x = d
P is a point somewhere in the i th subinterval which is of length Δx

Figure 39.1

We have divided the line into n equal subintervals each of length $\Delta x = 10/n$. We have shown the point P as lying somewhere in the i th subinterval. We can denote the probability of it falling in the i^{th} subinterval as $p_i = f(x_{i-1})\Delta x = \frac{\Delta x}{10}$ where we have introduced a function f (x) which in this example is constant over AB but which we identify by its value at the beginning of each of its intervals. This is not done to make life complicated but so as to be able to extend the method we are developing to cases which do not have equiprobable distributions. In this example $f(x)$ is just the constant function 1/10.

If now we wish to assign a probability to the point P being somewhere in the interval [c,d] where for the moment we assume that c and d fall respectively at the beginning and the end of their particular subdivisions it is only necessary to add up the probabilities for the subdivisions involved. i.e. $P(c \leq x \leq d) = \sum\limits_{i=m}^{n} f(x_{i-1})\Delta x$where m and n are the numbers of the first and last subintervals involved.

The next step is to allow the number of subdivisions to increase without limit. Then Δx tends to 0 and it becomes immaterial whether or not c and d are actually at the beginning or ends of their respective intervals. Even if f(x) where not constant it would also become immaterial whether or not we chose values at the endpoints or from within the intervals so long as f was at least a continuous function. You will recognize here the outlines of the process which leads to the Riemann integral of a continuous function on an interval [c,d].

Without more ado we will define the probability that the point falls somewhere in the interval [c,d] as $P(c \leq x \leq d) = \int_c^d f\,(x)\,dx$ where in this example f(x) = 1/10. Evaluating this for our example we have $P(c \leq x \leq d) = \int_c^d \frac{1}{10}dx = \frac{d-c}{10}$ which coincides with our common sense expectation of what that probability should be. Moreover since according to this $P(0 \leq x \leq 10) = \int_0^{10} \frac{1}{10}dx = 1$ we have also ensured that the probability of the point being somewhere on the line is 1. As we know from our earlier work this is a vital prerequisite for a probability distribution.

Before we give a formal definition note that the function we have defined here could be redefined for $-\infty < x < \infty$ in the following way:

$$f : x \mapsto f(x) \qquad \begin{cases} f(x) = 0 & -\infty < x < 0 \\ f(x) = \frac{1}{10} & 0 \leq x \leq 10 \\ f(x) = 0 & 10 < x < \infty \end{cases} \tag{9.2}$$

This enables us to tidy things up by always using ±∞ as our extreme limits of integration. Our essential condition on f can then be rewritten as $\int_{-\infty}^{\infty} f(x)dx = 1$.

We can also define a new function F(x) to represent the probability of finding the point in the interval from -∞ to x as $F(x) = \int_{-\infty}^{x} f(t)dt$. Note that the name of the function variable inside the integral sign has to be changed as we are now using x as a limit.

It is easy to see that in our example F(x) = 0 for $-\infty < x < 0$. F(x) = x/10 for 0≤ x ≤ 10. F(x) = 1 for x ¿ 10. *F* is called the cumulative distribution function (cdf.) corresponding to the probability distribution function (pdf.) f. From all this we can extract a rather lengthy set of definitions as a summing up.

Definition 16 *Given a continuous random variable X which may be defined either on the whole of R or on a subinterval only of R. A function f defined on all of R but having the value 0 outside the interval on which X is defined will be called a* probability distribution function (pdf.) *for X if*

1. *f is continuous on R.*
2. $f(x) \geq 0 \forall x \in R$.
3. $\int_{-\infty}^{\infty} f(x)dx = 1$.

Now given a subinterval E = [x_1, x_2] of the real line R we define the probability of finding the value of the random variable X in E to be:

$$P(x \in E) = \int_{x_1}^{x_2} f(x)dx. \tag{9.3}$$

Finally we define the cumulative distribution function F (cdf.) corresponding to f by $F(x) = \int_{-\infty}^{x} f(t)dt$. Note that F(x) is the probability of finding X somewhere between the left end of the distribution (whether or not this is actually -∞) and the value x.

Any function *f* satisfying the definition above is a possible pdf. It is of course our first task in any problem to find the correct pdf. It is often useful to think of the pdf. as defining an area under a curve and above the x axis which determines the probability of finding X on that particular portion of the x axis. This interpretation follows at once of course from our earlier work on the relationship between integrals and areas. eg. for our numerical example we have the situation of

The cdf. has the following important properties:

1. $\lim\limits_{x \to -\infty} F(x) = 0$
2. $\lim\limits_{x \to \infty} F(x) = 1$
3. $\frac{dF}{dx} = f(x)$

The properties a. and b. are obvious from the definitions. Property c. follows from the definition of F(x) as an integral and the fundamental theorem of calculus.

The cdf. is also useful in finding probabilities on intervals e.g.

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x)dx = \int_{-\infty}^{x_2} f(x)dx - \int_{-\infty}^{x_1} f(x)dx \tag{9.4}$$

$$= F(x_2) - F(x_1) \tag{9.5}$$

This last property of the cdf. will be very useful when we discuss the normal distribution.

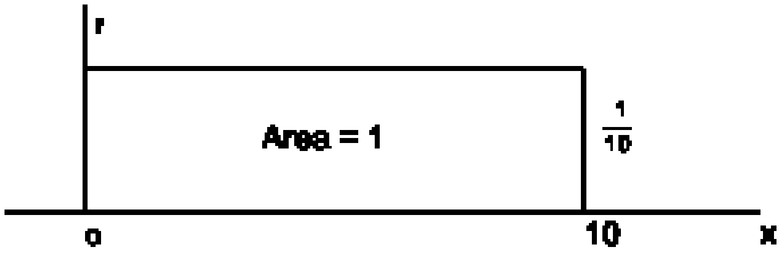


Figure 39.2

Let us stay for the moment with the problem of choosing a point at random on the line segment AB of length 10cm which we take to stretch from $x = 0$ to $x = 10$ on the real axis. However we will now suppose that the choice is not equiprobable but that the person making the choice has a subconscious preference for points near the center of the segment. We might postulate a pdf. for the situation now as follows: $f(x) = \begin{cases} 0 & -\infty < x < 0 \\ kx & 0 \leq x < 5 \\ k(10-x) & 5 \leq x \leq 10 \\ 0 & 10 < x < \infty \end{cases}$ where k is a positive constant.

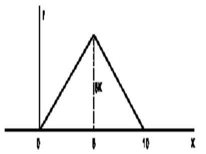
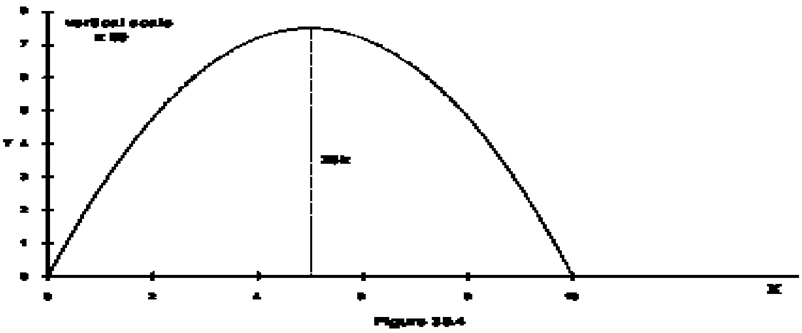


Figure 39.3

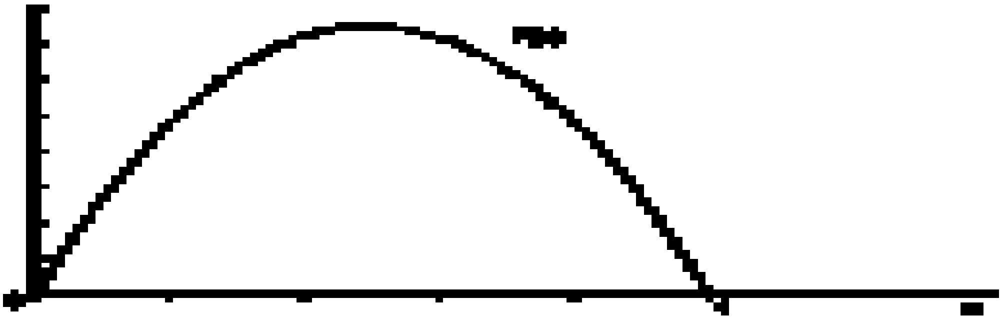
This is illustrated in Figure 39.3. The constant k is not of course arbitrary. We must have $\int_{-\infty}^{\infty} f(x)dx = \int_0^{10} f(x)dx = 1$. Here this is simply the condition $\frac{1}{2} \cdot 10 \cdot 5k =$



$\underline{\underline{1 \Rightarrow k = \frac{1}{25}}}$. We used simple geometry to evaluate the area here but to illustrate the integral we have: $P(3 \le x \le 6) = \int_3^5 \frac{x}{25} dx + \int_5^6 \frac{10-x}{25} dx = \left[\frac{x^2}{50}\right]_3^5 + \left[\frac{-(10-x)^2}{50}\right]_5^6$

$$\frac{16}{50} + \frac{9}{50} \underline{\underline{= \frac{1}{2}}} \quad . \tag{9.6}$$

Another variation of the same problem would assign a pdf. as follows: $f(x) = \begin{cases} 0 & -\infty < x < 0 \\ kx(10-x) & 0 \le x \le 10 \text{ where k is a positive constant.} \\ 0 & 10 < x < \infty \end{cases}$



This pdf. concentrates the probability still more about the center of the segment. We require $\int_0^{10} kx(10-x)dx = 1 \Rightarrow \left[5kx^2 - \frac{kx^3}{3}\right]_0^{10} = 1$

$$\Rightarrow 500k - \frac{1000k}{3} = 1 \Rightarrow k = \frac{3}{500} \tag{9.7}$$

We also have $P(3 \le x \le 6) = \int_3^6 \frac{3}{500}\left(10x - x^2\right)dx = \frac{3}{500}\left[5x^2 - \frac{x^3}{3}\right]_3^6 = \frac{93}{250}$

Note that this is less than the previous result yet the probability is more centrally concentrated. Explain?

Definition 17 *A* mode *of a random variable X with pdf. $f(x)$ is a value of the random variable for which $f(x)$ has a local maximum.*

Definition 18 *The* median *m of a random variable X with pdf. $f(x)$ is that value of the random variable defined by the equation $\int_{-\infty}^{m} f(x)dx = \frac{1}{2}$.*

Definition 19 *The* lower *and* upper quartile values *L and Q respectively of a random variable X with pdf. $f(x)$ are defined by the equations:*

$$\int_{-\infty}^{L} f(x)dx = \frac{1}{4}, \qquad \int_{-\infty}^{Q} f(x)dx = \frac{3}{4}. \tag{9.8}$$

Example 44 *Given a pdf. $f(x)$ for a random variable X on the interval [0,1] defined as $f(x) = \begin{cases} 0 & -\infty < x < 0 \\ kx(1-x) & 0 \le x \le 1 \\ 0 & 1 < x < \infty \end{cases}$ where k is a positive constant. Find*

1. *The value of k.*
2. *The cdf. F(x)*
3. *The pdf. g and cdf. G of the random variable W where $w = x^2$ on $[0,1]$. Sketch graphs of f and F.*

Solution.

1. *We need $\int_{-\infty}^{\infty} f(x)dx = 1 \Rightarrow \int_0^1 kx(1-x)dx = 1 \Rightarrow \underline{\underline{k = 6}}$*
2. *If $x < 0$ then $F(x) = 0$ and if $x > 1$ $f(x) = 1$. If $0 \le x \le 1$ $F(x) = \int_0^x 6t\,(1-t)\,dt = [3t^2 - 2t^3]_0^x = x^2\,(3-2x)$*

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ x^2\,(3-2x) & 0 \le x \le 1 \\ 1 & 1 < x < \infty \end{cases} \tag{9.9}$$

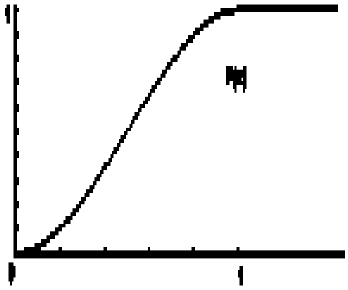
3. *It is easier to calculate $G(w)$ first. Since $w = x^2$ we have $x = \sqrt{w}$ (no sign problem since $0 \le x \le 1$) $G(w) = w\,(3 - 2\sqrt{w})$since the probability that w is in the interval [0, w] is clearly that of x being in the interval $[0, \sqrt{w}]$ since $w = x^2$.*

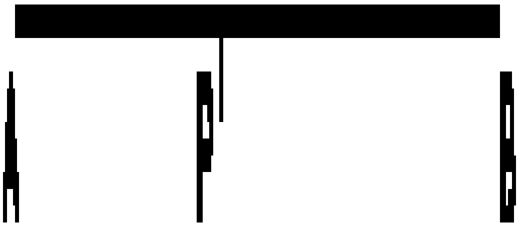
So $G(w) = \begin{cases} 0 & -\infty < w < 0 \\ 3w - 2w^{\frac{3}{2}} & 0 \le w \le 1 \\ 1 & 1 < w < \infty \end{cases}$

We now use the relation $g = \frac{dG}{dx}$ to get

$$g(w) = \begin{cases} 0 & -\infty < w < 0 \\ 3\,(1-\sqrt{w}) & 0 \le w \le 1 \\ 0 & 1 < w < \infty \end{cases} \tag{9.10}$$

The graphs of $f(x)$ and $F(x)$ are shown above. The vertical scale of the graph of $f(x)$ is magnified.





We must now investigate the expectation value and variance of a continuous probability distribution. First we recall that for a discrete random variable we have

$$E(x) = \sum_{i=1}^{n} p_i x_i. \tag{9.11}$$

The expectation value of a continuous random variable is defined analogously using integration instead of summation as

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx \tag{9.12}$$

where $f(x)$ is the pdf. for the distribution. As in the discrete case $E(x)$ is often referred to as the mean μ.

Example 45 *Find the mean of the random variable X whose density function is defined by* $f(x) = \begin{cases} 0 & -\infty < x < 0 \\ 4x\left(1 - x^2\right) & 0 \leq x \leq 1 \\ 0 & 1 < x < \infty \end{cases}$

<u>*Solution.*</u>

$$\mu = E(x) = \int_{-\infty}^{\infty} x f(x) dx = \int_{0}^{1} 4x^2 \left(1 - x^2\right) dx = \frac{8}{\underline{\underline{15}}} \tag{9.13}$$

Similarly, but much more generally, if we have a new function h(x) of the original random variable x, with the pdf. f(x) being unaltered, we can define the expectation value of h(x) as $E[h(x)] = \int_{-\infty}^{\infty} h(x) f(x) dx$.

If $h(x) = (x - \mu)^2$ *which you will recognize as the squared deviation from the mean, then again by analogy with the discrete case we shall define the variance of* X *as*

$$E[h(x)] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \sigma_x^2 = V(X). \tag{9.14}$$

Now $\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - 2\mu\int_{-\infty}^{\infty} xf(x)dx + \mu^2\int_{-\infty}^{\infty} f(x)dx$

$$= \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 \tag{9.15}$$

$\sigma^2 + \mu^2 = E\left(x^2\right)$ *and we note that this result is exactly the same as that obtained in the discrete case.*

Example 46 *Find the variance of the random variable* X *of Example 39.2.*

Solution.

Using the results of Example 39.2 we have $\sigma^2 + \left(\frac{8}{15}\right)^2 = \int_0^1 x^2 \cdot 4x\left(1-x^2\right) dx and$ *from this we easily find* $\underline{\underline{\sigma^2 = \frac{11}{225}}}$

9.1.1 Exercise

1. $f(x) = kx^{-2}$ if $x \geq 1$, $\quad f(x) = 0$ if $x < 1$; find the value of k and then find $P(x < 2)$

2. $f(x) = ke^{-2x}$ if $x \geq 0$, $\quad f(x) = 0$ if $x < 0$; find k and hence find $P(1 < x < 2)$.

3. $f(x) = k\sin\pi x$ if $0 \leq x \leq 1$ and is zero otherwise. Find k and hence find $P(x > 1/3)$.

4. Find the (cumulative) distribution function for each of the density functions in Nos. 1-3.

5. Find the mean, median, mode and variance for the random variable with the pdf.

$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{3}{4}x\left(2-x\right) & 0 \leq x \leq 2 \\ 0 & x > 2 \end{cases} \tag{9.16}$$

6. Find the mean, median, mode and variance for the random variable with pdf.

$$f(x) = \begin{cases} 0 & x < -\frac{1}{2} \\ \frac{1}{2}\pi\cos\pi x & -\frac{1}{2} \leq x \leq \frac{1}{2} \\ 0 & x > \frac{1}{2} \end{cases} \tag{9.17}$$

7. If the density function f, of a distribution is given by

$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2}x & 0 \leq x \leq 1 \\ \frac{1}{2} & 1 < x \leq 2 \\ \frac{1}{2}\left(3-x\right) & 2 < x \leq 3 \\ 0 & x > 3 \end{cases} \tag{9.18}$$

Find the cumulative distribution function F, and sketch the graphs of f and F. Find the density and distribution functions, g and G, of the new random variable Y where y = x^2 and sketch the graphs of g and G.

8. A probability density function of a random variable X is defined as follows: $f(x) = \begin{cases} x(x-1)(x-2) & 0 \le x < 1 \\ \lambda & 1 \le x \le 3 \\ 0 & \text{otherwise} \end{cases}$ where λ is a suitable constant.

   Calculate the expectation value μ of x. What is the probability that x $\le$ μ ?

9. The probability density function p(t) of the length of life, t hours, of a certain electronic component is given by $p(t) = ke^{-kt}$ $(0 \le t < \infty)$, where k is a positive constant. Show that the mean and standard deviation of this distribution are both equal to 1/k. Find the probability that the life of a component will be at least t$_0$ hours. Given that a particular component is already t$_1$ hours old and has not failed, show that the probability that it will last at least a further t$_0$ hours is e^{-kt_0}.

   A computer contains three components of this type and the failure of any one may be assumed to be independent of the failure of the others. Find the probability that:

   (a) none of the three will have failed at t$_0$ hours.

   (b) exactly one will fail in the first t$_0$ hours, another in the next t$_0$ hours and a third after more than 2t$_0$ hours.

9.2 The Uniform Continuous Probability Distribution.

We have studied equiprobable distributions in the discrete case. Obvious examples are the tossing of a coin or a die. The uniform distribution is a generalization of this kind of distribution to the continuous case.

Definition 20 *If a continuous random variable X can assume all values between $x = a$ and $x = b$ $(b > a)$ and if the pdf. f, is independent of position within this interval then the random variable X is said to possess a uniform distribution.*

To make this work we must obviously set $f(x) = \begin{cases} 0 & -\infty < x < a \\ \frac{1}{b-a} & a \le x \le b \\ 0 & b < x < -\infty \end{cases}$ so that $\int_{-\infty}^{\infty} f(x)dx = 1$. Then $F(x) = \begin{cases} \int_{-\infty}^{x} f(t)dt = 0 \text{ if } x < a \\ \int_{-\infty}^{x} \frac{dt}{b-a} = \frac{x-a}{b-a} \text{ if } a \le x \le b \\ 1 \text{ if } x > b \end{cases}$ The mean $\mu = \int_{-\infty}^{\infty} xf(x)dx = \int_a^b \frac{x}{b-a}dx = \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{b+a}{2}$ and $\sigma^2 = E\left(X^2\right) - \mu^2 = \int_a^b \frac{x^2}{b-a}dx - \frac{(b+a)^2}{4} = \left[\frac{x^3}{3(b-a)}\right]_a^b - \frac{(b+a)^2}{4}$

$$= \frac{b^2 + ab + a^2}{3} - \frac{b^2 + 2ab + a^2}{4} = \frac{(b-a)^2}{12} \tag{9.19}$$

We summarize these results as

Theorem 9 *The mean μ and the variance σ^2 of the uniform distribution of of the above definition are given by $\mu = \frac{a+b}{2}$, and $\sigma^2 = \frac{(a-b)^2}{12}$.*

Proof 8 *As given above.*

Example 47 *The random variable X has uniform distribution in the interval $0 \leq x \leq 10$. Find $P(X = x)$ given that $x^2 - 5x + 4 > 0$)*

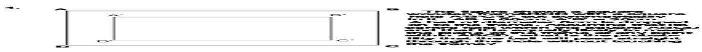
Solution.

$$x^2 - 5x + 4 > 0 \Rightarrow (x-1)(x-4) > 0 \qquad x > 4 \text{ or } x < 1 \tag{9.20}$$

For our problem this means $0 \leq x < 1$ or $4 < x \leq 10$ since f(x) = 0 outside 0 ≤ x ≤ 10. So $P(X : x^2 - 5x + 4 > 0) = \int_0^1 \frac{dx}{10} + \int_4^{10} \frac{dx}{10} = \frac{7}{10}$

Example 48 *The line AB has length 10 cm. A point P is taken at random on AB all points being equally likely. Find the probability that the area of the circle of radius AP will exceed 10 cm^2.*

Solution.



If $AP = x$ the area of the circle is πx^2. $\pi x^2 > 10 \Rightarrow x > \sqrt{\frac{10}{\pi}}$

$$P(\mathit{required}) = \int_{\sqrt{\frac{10}{\pi}}}^{10} \frac{dx}{10} = 1 - \frac{1}{\sqrt{10\pi}}. \tag{9.21}$$

9.2.1 Exercises

1. The line AB has length 10cm. An interval of length 2cm is marked at random on the line, the positions of the interval being uniformly distributed. What is the probability that the interval will contain the midpoint of AB?

2. A circular disc of radius 10cm is placed on a table. Another disc, of radius 3cm, is now placed on the table so that it is at least partially in contact with the first disc. On the assumption that the permissible positions of the smaller disc are uniformly distributed, what is the probability that it covers the center of the larger disc?

3. A point P is taken at random on the side AB (and between A and B) of the square ABCD, the positions of P being uniformly distributed. If PC cuts BD at X find the probability (a) that BX ¡ (1/2)BD; (b) that BX ¡ (1/4)BD

4. A point A is marked on the circumference of a circle of radius r and a chord AP is drawn at random, the positions of the point P on the circumference being uniformly distributed. Find the expected length of the chord.

5. Determine the variance of the length AP of Question 5.

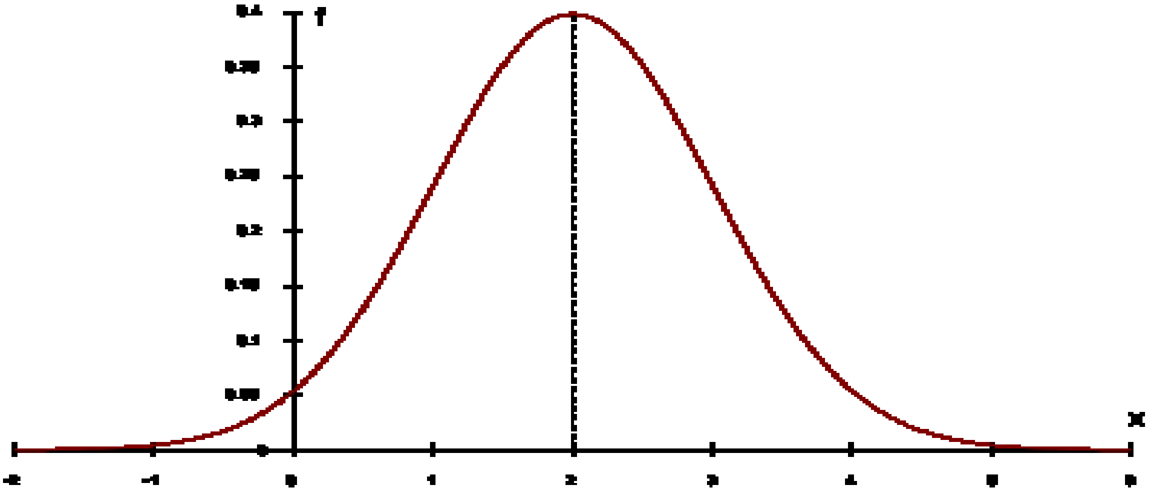


Figure 39.3

6. A point P is marked on the side AB of the square ABCD, the points within AB being uniformly distributed. Find the mean and variance of the area of the triangle APD.

7. ABC is an isosceles triangle right angled at B. Through A a line is drawn at random to cut BC at P, the angle BAP being uniformly distributed between 0 and $\pi/4$. If AB = a, show that the expected area of the triangle ABP is $\frac{a^2 \ln 2}{\pi}$ and find its variance.

9.3 Reading: Classification of Men According to their Natural Gifts by Sir Francis Galton

9.3.1 Introduction to the Reading

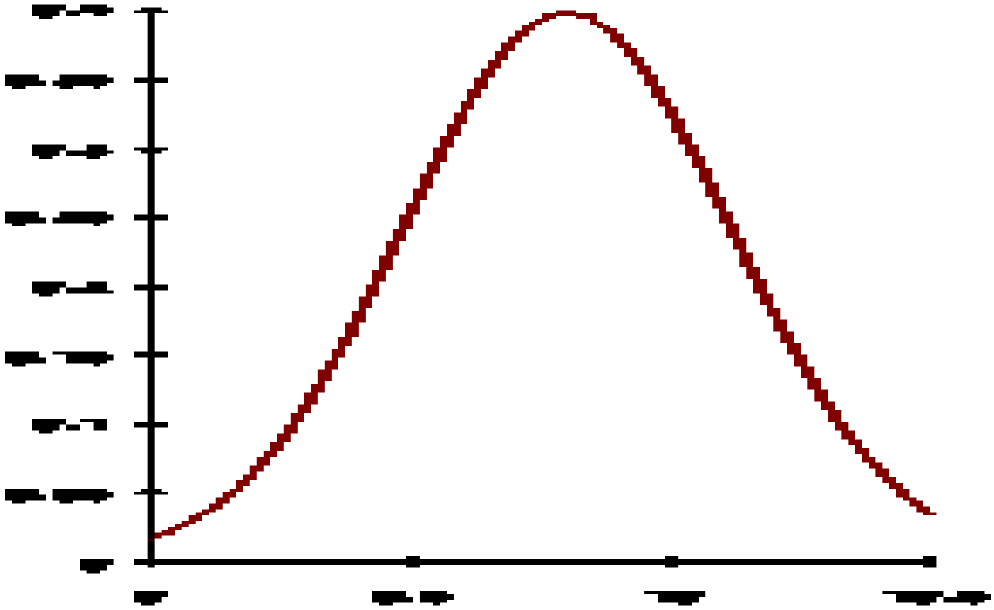
9.3.2 The Paper

Chapter 10

Probability Distributions: The Normal Distribution 2

10.1 The Normal Distribution.

The normal distribution is probably he single most widely used distribution in statistics. It arises from the observation that in practice many statistical variables are rather closely packed around a mean value and tail off fairly rapidly on either side of this mean. [This statement is deliberately vague. We will soon have as much precision as we require.] The pdf. of such a distribution would have the general appearance of Figure 39.3.



This is the famous bell shaped curve here centered on a mean of 2. Let us now discuss without going into too much detail some of the experiments which can lead to normal distributions.

Experiment 1.

If we measure the IQ. of a very large section of the population it is found that the mean is about 100, and that there will be very few incidences below 60 or above 140. If the probabilities are worked out as relative frequencies i.e.

$$P(\text{IQ.} = 100) = \frac{\text{number of cases of IQ} = 100}{\text{total tested}} \tag{10.1}$$

and then plotted against the IQ values we will get a bell shaped curve. [This is really discrete variable, as in fact are most of the examples occurring in practice but for large samples the approximation involved in joining up the points by the best fitting curve and so getting our bell shaped curve is justified.]

Experiment 2.

If a physics experiment to measure say g the acceleration due to gravity at the earth's surface is performed a large number of times the results will have a range of values depending on the accuracy of the experimental techniques used. If we plot the histogram for the results. [The results will be divided into equally spaced groups e.g. results falling in the ranges 9.60 - 9.65, 9.65 - 9.70, 9.70 - 9.75 etc. will be counted and their relative frequencies plotted. The size of the interval determining a group will depend on the accuracy of the experiment getting smaller and smaller as the experimental technique is refined.] This histogram will give us a bell shaped curve. There is not just one bell shaped curve involved in this experiment. A super accurate experiment would have an extremely sharply peaked bell shaped curve centred around a value near 9.8. as shown below. A crude experiment would produce a much "better" bell shaped curve.

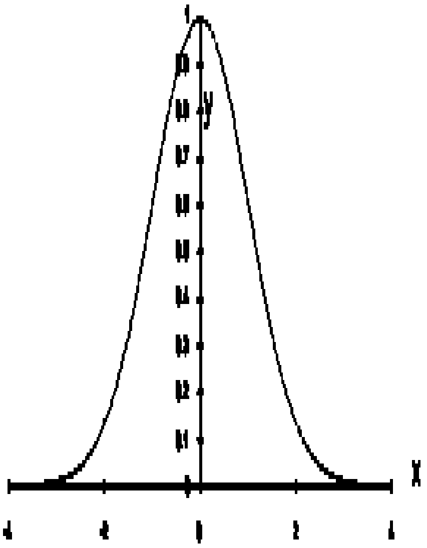


Figure 39.4

It is found then that many situations arising in the analysis of numerical data of various kinds produce these bell shaped curves. If we are to work with this kind of distribution mathematically we need a more powerful tool than a well drawn graph however pretty this may be. We need a mathematical model for what we will call a normal probability distribution. [For the moment this is defined unsatisfactorily as the kind of distribution that the above experiments and others like them lead to.]

The problem of constructing a mathematical model was tackled first by Gauss in the middle of the last century. He found that functions of the form $\frac{a}{b+cx^2}$ or $\frac{a}{(b+cx^2)^2}$ gave approximately bell shaped curves but for any particular real situation it was never possible to choose the constants a,b and c so as to get a good fit of the mathematical curve to the observed data over the whole of its range.

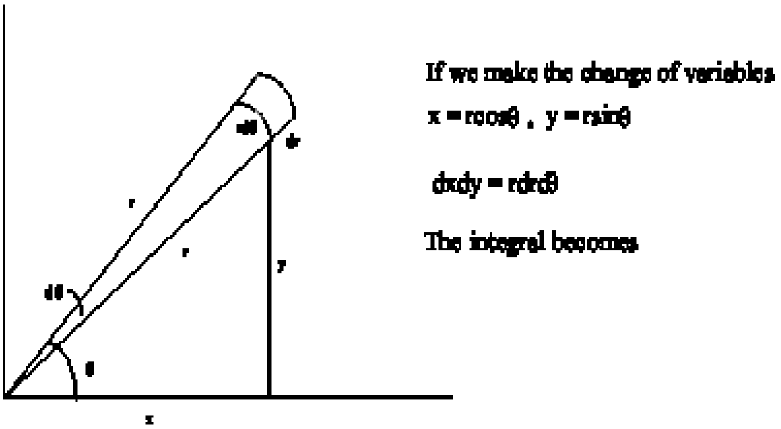
So Gauss now turned to the curve $y = Ae^{-\lambda x^2} \quad \lambda > 0, \quad A > 0$. Clearly $y(0) = A$ and $y \to 0$ very rapidly for $|x| > 1$. Also $\frac{dy}{dx} = -2\lambda Axe^{-\lambda x^2}$ and $\frac{dy}{dx} = 0 \Rightarrow x = 0$.

$$\frac{d^2y}{dx^2} = -2\lambda Ae^{-\lambda x^2} + 4\lambda^2 Ax^2 e^{-\lambda x^2} \text{ so } \left(\frac{d^2y}{dx^2}\right)_{x=0} = -2\lambda A < 0 \qquad (10.2)$$

Thus the curve has a maximum at (0,A).

$$\frac{d^2y}{dx^2} = 0 \Rightarrow -2\lambda A + 4\lambda^2 Ax^2 = 0 \Rightarrow x = \pm\frac{1}{\sqrt{2\lambda}}. \qquad (10.3)$$

and the curve has points of inflexion when $x = \pm\frac{1}{\sqrt{2\lambda}}$. Analytically this looks to be a good candidate for a BSC. and we confirm this by the plot of $y = e^{-\frac{x^2}{2}}$ as in Figure 39.4.



If we make the change of variables

$x = r\cos\theta$, $y = r\sin\theta$

$dxdy = rdrd\theta$

The integral becomes

Gauss then found that by suitable choices of A and λ and possibly by translations so that the mean need not be at 0 he was able to get an extremely good fit to observed distributions. He then chose $\lambda = \frac{1}{2}$ and $A = \frac{1}{\sqrt{2\pi}}$ for special examination.

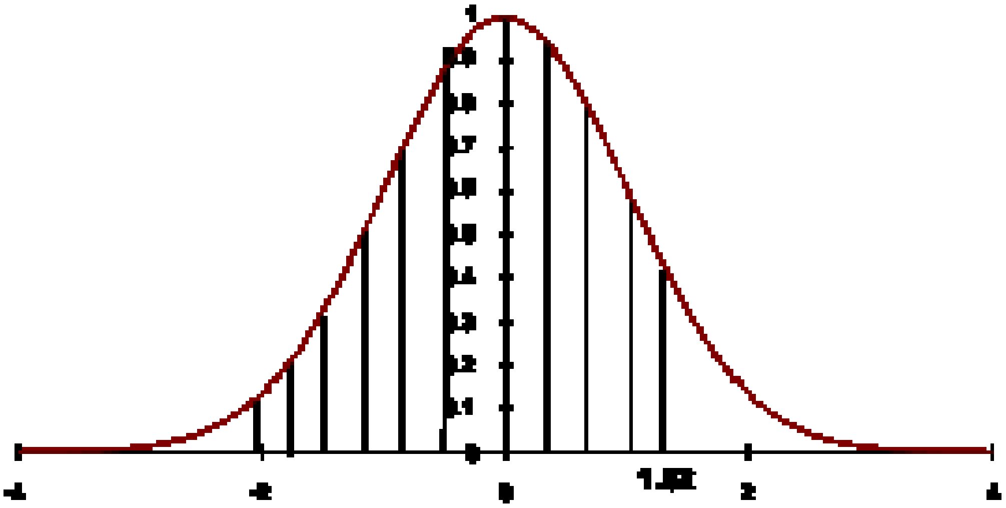
Definition 21 *The random variable X is said to have a* standard normal distribution *if it has the pdf.* $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$.

Note that the small ϕ is the usual notation for the standard normal distribution and is used in all the literature rather than f. For this to be a genuine pdf. we must have $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx = 1$. To check this we need to be able to evaluate $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx$. This is quite a tricky integral and for our purposes we are allowed to assume that $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx = \sqrt{2\pi}$ which gives us the *normalization* required.

[For the curious here is a sketch of one way to find this integral.

$$\text{Let } I = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx. \qquad (10.4)$$

Consider $I_1 = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy$ where the x and y are the usual x and y of the cartesian plane. We have $I_1 = I^2$. Assuming the integrals can be combined (and this would have to be proved) we have $I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{\left(x^2+y^2\right)}{2}} dxdy$



$$I^2 = \int_0^{\infty} \int_0^{2\pi} e^{-\frac{r^2}{2}} rdrd\theta = 2\pi \int_0^{\infty} re^{-\frac{r^2}{2}} dr = 2\pi \left[-e^{-\frac{r^2}{2}}\right]_0^{\infty} = 2\pi \tag{10.5}$$

and $I = \sqrt{2\pi}$ as required.]

It is perhaps worth noting that there is a well developed theory of integrals of this kind. They are called Gaussian Integrals and arise in many applications such as the kinetic theory of gasses and quantum theory.

The qualifying word *standard* in the definition is important. This is obviously a special case but by using the word *standard* we are implying that we hope to be able to obtain a model for any normal distribution by suitably modifying this *standard* distribution. This hope will be realized shortly.

Theorem 10 *The mean μ and variance σ^2 of the standard normal distribution $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ are given by $\mu = 0$ and $\sigma^2 = 1$ respectively.*

Proof 9

$$\mu = E(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} xe^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \left[-e^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty} = 0. \tag{10.6}$$

{ *Strictly, these integrals involving infinite limits have not so far been defined in our course. They are certainly not Reiman integrals. The following method of handling them is plausible:*

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} xe^{-\frac{x^2}{2}} dx = \lim_{M\to\infty} \int_{-M}^{M} \frac{1}{\sqrt{2\pi}} xe^{-\frac{x^2}{2}} dx \tag{10.7}$$

$= \lim_{M\to\infty} \frac{1}{\sqrt{2\pi}} \left[-e^{-\frac{M^2}{2}} + e^{-\frac{M^2}{2}}\right] = 0\}$

$$\sigma^2 = E(X^2) - \mu^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}} dx \tag{10.8}$$

$$\textit{Let } x = u \qquad xe^{-\frac{x^2}{2}} dx = dv \qquad \textit{so } dx = du \textit{ and } v = -e^{-\frac{x^2}{2}} \tag{10.9}$$

$$\textit{Then } \sqrt{2\pi}\sigma^2 = \left[-xe^{-\frac{x^2}{2}}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = 0 + \sqrt{2\pi} \tag{10.10}$$

$$\textit{And hence } \sigma^2 = 1. \tag{10.11}$$

Gauss chose $e^{-\frac{x^2}{2}}$ *rather than* e^{-x^2} *itself for the standard normal distribution because of this property that* $\sigma^2 = 1$.

The cumulative standard normal distribution function is denoted always by $\Phi(x)$ and is of course, defined by $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = \int_{-\infty}^{x} \phi\left(t\right) dt$.

Now this integral really is difficult. In fact there is no way that we can get an exact answer as we could albeit with difficulty, for $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx$. There is no formula for this integral. This is not because it is too advanced for this course; no formula exists or, what amounts to the same thing this integral defines a completely new function just as $\int_1^x \frac{1}{t} dt$ defines the function $\ln x$ in elementary work. Every value such as say $\Phi\left(1.2\right) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$has to be worked out to the required number of decimal places by approximate methods. Because of this it is usual to use tables of $\Phi(x)$ in statistical work and we shall follow this practice.

The usual table stops at x = 4 since Φ(4) = P(- ∞ ¡ x ≤ 4) = 0.99997 and further tabulation is obviously pointless.

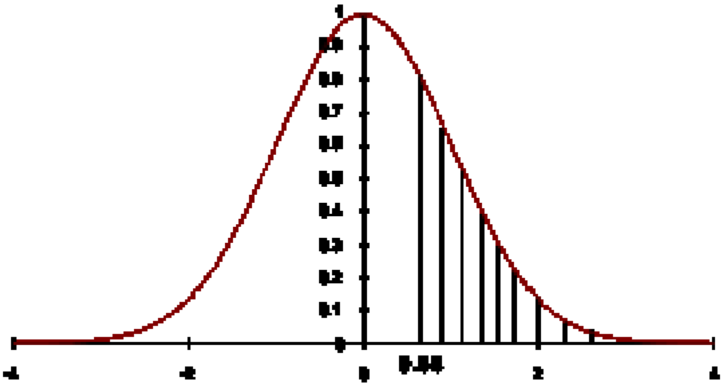
Note that the IB. tables which we use and in fact most of the tables available take advantage of the symmetrical nature of the distribution about its mean of 0 in order to save space. This means that a little ingenuity in the use of the table is sometimes required. For example if we want to know the probability of getting a value of x in the interval (1.2,2) we require $\int_{1.2}^{2} \phi(x) dx$ which cannot be found directly from the tables. However $\int_{1.2}^{2} \phi(x) dx = \int_{-\infty}^{2} \phi(x) dx - \int_{-\infty}^{1.2} \phi(x) dx = \Phi(2) - \Phi(1.2)$. The last two values can be found in the table and we have $\int_{1.2}^{2} \phi(x) dx = 0.9772 = 0.8849 = 0.0923$.

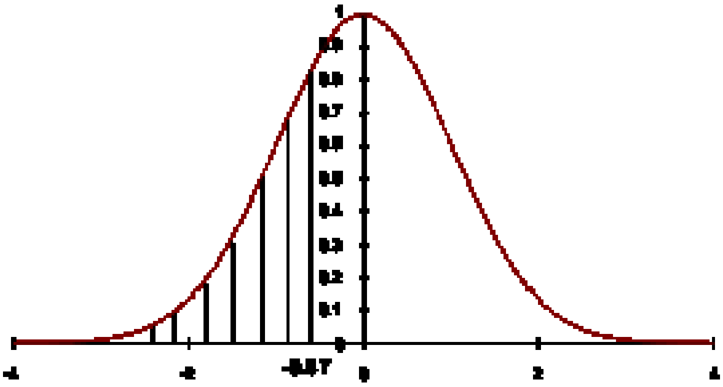
Before we develop the theory for any normal distribution no matter what its mean and standard deviation it will be a good idea to practice the use of these tables on the standard normal distribution. From now on we shall refer to normal distributions briefly using the notation N(μ,σ), where μ and σ are of course the mean and standard deviation. Thus the standard normal distribution is simply N(0,1). Here are some examples of the use of the tables for N(0,1).

Example 49 *For N(0,1) find P(X ≤ 1.62)*

Solution.

Do not lose sight of the fact that these probabilities are represented by areas under the pdf. curve. This helps our understanding of some of the gymnastics we must undertake in order to get by with just this one table. In this simple first example the answer is just $\Phi\left(1.62\right) = 0.9474$ *which we read straight from the table.*





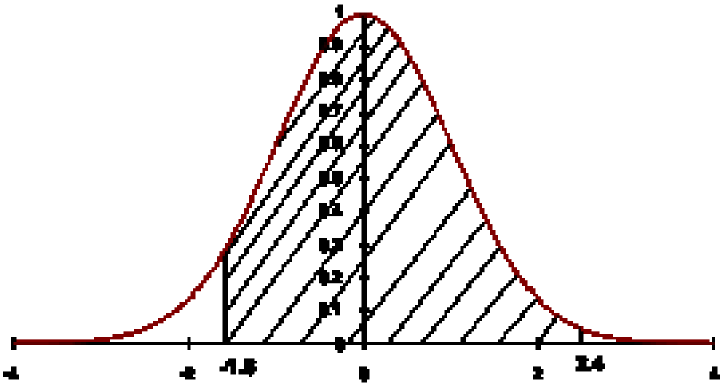
Example 50 *For N(0,1) find P(x* $\geq$ *0.58)*

Solution.

Either from inspection of the graph or from our knowledge of probability theory we have P(x $\geq$ *0.58) = 1 - P(x ¡ 0.58) = 1 -* Φ*(0.58) = 1 - 0.7190 = 0.2810*

Example 51 *For N(0,1) find P(x* $\leq$ *-0.87)*

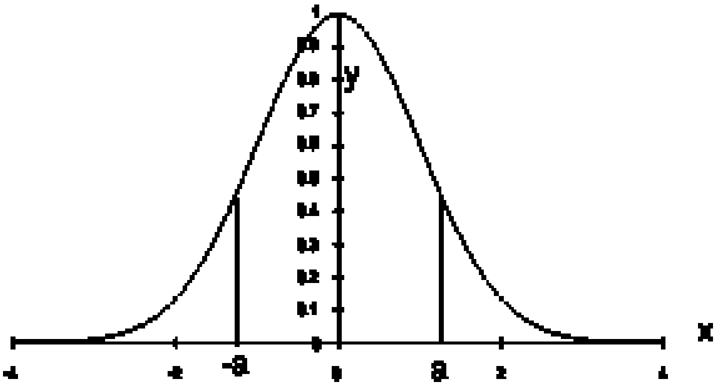
Solution.



You will notice that in the interest of saving trees the negative values of x are not printed in your tables. However this is no big problem since the standard normal distribution is symmetric about x = 0. We have P(x $\leq$ *-0.87) = P(x* $\geq$ *0.87) = 1 - P(x* $\leq$ *0.87) = 1- 0.8078) = 0.1922*

Example 52 *For N(0,1) find P(-1.6* $\leq$ *x* $\leq$ *2.4)*

Solution.



We have P(-1.6 $\leq$ x $\leq$ 2.4) = P(x $\leq$ 2.4) - P(x $\leq$ -1.6) = P(x $\leq$ 2.4) -[1 - P(x $\leq$ 1.6)] = 0.9918 -[1 - 0.9542)] = 0.937.

We now take note of the fact that it is extremely unlikely that an example arising from a real situation will exactly fit the standard normal distribution. We must then examine how we can modify the *standard normal* distribution so that we can use it to handle any *normal* distribution no matter what its mean and variance.

Suppose then that we have a distribution of a random variable x that we want to model as normal and which has an arbitrary mean μ and variance σ^2. We define a new random variable y in terms of the old one x by the equations

$$y = \sigma x + \mu \quad \text{and } x = \frac{y - \mu}{\sigma} \tag{10.12}$$

We now have $dy = \sigma dx$ and $1 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \frac{dy}{\sigma} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy.$

If we now define a new pdf. by $\phi\left(y\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$ the above line gives us the necessary normalization. Moreover the mean and variance of the new variable y will be μ and σ.

[e.g. $E(y) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} y e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$ Let $t = \frac{y-\mu}{\sigma} \quad y = \sigma t + \mu$

$$E(y) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \left(\sigma t + \mu\right) e^{-\frac{t^2}{2}} \sigma dt = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-\frac{t^2}{2}} dt + \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt \tag{10.13}$$

$$= 0 + \mu = \mu. \tag{10.14}$$

A similar reverse substitution and appeal to already calculated integrals will show that the variance of y is σ^2 as we hoped.]

We will not find tables calculated for the cumulative distribution function arising from this pdf. as we have for the standard normal distribution. Since an arbitrary normal distribution $N(\mu, \sigma)$ can have any value of μ and σ it would be quite impracticable to construct such tables. However we can proceed as follows.

First if we have a normal distribution $N(\mu, \sigma)$ and retaining the variable name x instead of the y above the cumulative distribution function is denoted by I and we have

$$I(a) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{a} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{10.15}$$

Let $t = \frac{x-\mu}{\sigma}$ $\quad dt = \frac{dx}{\sigma}$ and $I(a) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\frac{a-\mu}{\sigma}} e^{-\frac{t^2}{2}}\sigma dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{a-\mu}{\sigma}} e^{-\frac{t^2}{2}} dt.$

From this we deduce the final, and the most powerful result of our theory:

Theorem 11 *If F is the cumulative distribution function of the random variable x with normal distribution $N(\mu, \sigma)$ then*

$$F(a) = \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{10.16}$$

where Φ is the cumulative distribution function of the standard normal distribution.

Proof 10 *This is simply a summing up of the above analysis.*

Example 53 *For N(6, 2.5) find P(4 ≤ x ≤ 9)*

Solution.

Here μ = 6 and σ = 2.5.

$$\textit{We require } P\left(\frac{4-6}{2.5} \le t \le \frac{9-6}{2.5}\right) \textit{ for } N(0,1) \textit{ i.e. } P\left(-0.8 \le t \le 1.2\right). \tag{10.17}$$

$$\begin{array}{l} P(-0.8 \le t \le 1.2) = P(t \le 1.2) - P(t \ge 0.8) \\ \qquad = 0.8849 - (1 - 0.7881) = 0.6730 \end{array} \tag{10.18}$$

Example 54 *Given that P(x ≥ 1.7) = 0.6331 for a normal distribution of mean 2 find its variance.*

Solution.

$$\begin{array}{l} P(x \le 1.7) = 1 - 0.6331 = 0.3669 \\ \Phi\left(\frac{1.7-\mu}{\sigma}\right) = 0.3669 \end{array} \tag{10.19}$$

But 0.3669 is not in the table!

Since $\Phi\,(0) = 0.5 and$ the negative values of x are not in the table this will often happen. If x = -a (with a ¿ 0) is the value we need clearly

$$\begin{array}{l} 0.3669 = P(x \le -a) = P(x \ge a) = 1 - P(x \le a) \\ P(x \le a) = 0.6331 \end{array} \tag{10.20}$$

[We are back where we started! In practice the above digression is not needed then but it may help you to understand what is going on.]

From the table a = 0.34. So the value that we require is -0.34 and

$$\begin{array}{l} \frac{1.7-\mu}{\sigma} = -0.34 \Rightarrow \sigma = \frac{-0.3}{-0.34} = 0.882 \\ \textit{Thus the variance is } \sigma^2 = 0.779 \end{array} \tag{10.21}$$

Example 55 *Given P(x ≤ 5.4) = 0.8710 for a normal distribution with variance 4 find the mean.*

Solution.

$$\Phi\left(\frac{5.4-\mu}{2}\right) = 0.8710 \Rightarrow \frac{5.4-\mu}{2} = 1.13 \Rightarrow \mu = 3.14 \tag{10.22}$$

Let us now look at a more realistic problem involving a normal distribution.

Example 56 *The heights of a large number of young children are measured correct to the nearest centimeter and the mean and standard deviation of the resulting frequency distribution are calculated and found to have values 122cm and 5.2 cm respectively. Hence a statistical model is assumed of a distribution N(122, 5.2). Calculate the probability of a random measurement of a child of the group yielding a result in each of the class intervals:* $x \leq 105$, $105 < x \leq 110$, $125 < x \leq 130$.

Solution.

In view of the accuracy limitation (to the nearest centimeter) and the fact that we are assuming a continuous distribution the first probability interval will be better modelled by P(x ≤ 105.5) , the second by $P(105.5 < x \leq 110.5$*) and so on.*

[The analysis of the errors introduced by treating a discrete distribution as a continuous one and the effect of experimental error forms part of advanced statistics. However the fact remains that a simple common sense approach such as we are adopting here will give results that are as good as can be expected given the imperfections in the data.]

The transformed variable is $\frac{x-122}{5.2}$

$$P(x \leq 105.5) = \Phi\left(\frac{105.5-122}{5.2}\right) = \Phi\left(-3.173\right) = 0.0001 \tag{10.23}$$

$$P\left(105.5 < x \leq 110.5\right) = \Phi\left(\frac{110.5-122}{5.2}\right) - \Phi\left(\frac{105.5-122}{5.2}\right) \tag{10.24}$$

$$= \Phi\left(-2.212\right) - 0.0001 = 1 - 0.9864 - 0.001 = 0.014 \tag{10.25}$$

$$P\left(125.5 < x \leq 130.5\right) = \Phi\left(\frac{130.5-122}{5.2}\right) - \Phi\left(\frac{125.5-122}{5.2}\right) \tag{10.26}$$

$$= \Phi\left(1.635\right) - \Phi\left(0.673\right) = 0.949 - 0.749 = 0.20 \tag{10.27}$$

10.1.1 Exercises

1. For N(0,1) find a. P(x ≤ 0.79) b. P(x ≤ -0.48)

2. For N(0,1) find a. P(x ≥ 1.82) b. P(-1.1 ≤ x ≤ 0.41) c. P(x ≥ -2.8)

3. For N(5, 3) find a. P(x ≤ 3.2) b. P(x ≥ 4.82)

4. For N(1.6, 2) find a. $P\left(|x| \leq 1.4\right)$ b. P(x is negative) c. P(x ≥ 3.2)

5. For N(0, 1) find a value z for which a. P(Z ≥ z) = 0.5199 b. P(Z ≤ z) = 0.942 c. P(Z ≤ z) = 0.3665

6. A normal distribution $N\left(2, \sigma\right)$ is such that P(X ≤ 2.8) = 0.7123. Find σ.

7. A normal distribution $N\left(\mu, 3\right)$ is such that P(X ≥ 1.4) = 0.6179. Find μ.

8. A normal distribution N(4, 3) is such that $P\left(X \leq a\right) = 0.1489$. Find a.

9. A normal distribution N(0, σ) is such that $P\left(|X| \leq 7\right) = 0.5$. Find σ.

10. Find P(3Z ≤ 4.1) for N(0, 1).

11. Find $\int_{-1.4}^{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$

12. Find $\int_{5}^{\infty} \frac{1}{3\sqrt{2\pi}} e^{-\frac{(x-4)^2}{18}} dx$

13. The weights of army recruits form a normal distribution with mean 69.8 kg. standard deviation 6.2 kg. Out of 976 recruits how many can be expected to weigh more than 80 kg?

14. A machine making connecting pins produces them with lengths normally distributed about a mean of 1.8 cm. The variance is 0.0001 cm^2. If 3% are rejected for oversize and 4.5% for undersize, find the tolerance limits for acceptance. 15. A light bulb test shows that 10% last longer than 2140 hours and that the mean is 1630 hours. Assuming that the lifetime of the bulbs is normally distributed, find the standard deviation and the percentage which would fail to meet the guarantee of 1000 hours. 16. In a normal distribution, 12.5% of the distribution is less than 14 and 20% is less than 17. Find its mean and standard deviation. 17. Packets of sugar, nominally 1 kg. have weights normally distributed with mean 1.05 kg. and standard deviation 0017 kg. Find the percentage of packets which weigh less than 1 kg. On average, how many packets labelled '1kg' can be produced on this machine from 1 tonne of sugar? The setting of the machine is now altered. The standard deviation remains the same and the mean is reduced until 4.9% of the packets contain less than 1 kg. Find the new mean and the number of packets which can now be obtained from 1 tonne of sugar. 18. The IQ's of 500 students are assumed to be normally distributed with mean 105 and standard deviation 12. How many students may be expected:

    (a) to have an IQ greater than 140;

    (b) to have an IQ less than 90;

    (c) to have an IQ between 100 and 110.

15. Rods are manufactured with a mean length of 20.2 cm. and standard deviation 0.09 cm. the distribution being normal. If rods of length less than 20.1 cm are rejected what is the probability that a rod which is retained has a length in excess of 20.3 cm?

10.2 The Normal Approximation to the Binomial Distribution

(This section, like medicine, should be swallowed quickly and without question!)

We state without proof the following fact: As n becomes large, the binomial distribution $B(n,p)$ tends to the normal distribution . $N\left(np, \sqrt{npq}\right)$Even for relatively small values of n the normal distribution can be a useful approximation to the binomial. Since the normal distribution is continuous and the binomial distribution is discrete we have to make what is called a **continuity correction** also.

A continuity correction is made as follows: Each discrete value of the binomial distribution is replaced by a suitable range of values of the normal distribution. The convention is to replace a sequence of precise values say 12, 13, 14, 15 by the ranges 11.5-12.5, 12.5-13.5, 13.5-14.5, 14.5-15.5. As long as np ¿ 10 and p is not 'too small' this will give a reasonable approximation.

Example 57 *What is the probability of throwing 15 tails in 25 tosses of a fair coin?*

Solution.

The number of tails in 25 tosses will satisfy B(25,0.5). Using this binomial distribution we find the required probability to be

$$\binom{25}{15}\left(\frac{1}{2}\right)^{15}\left(\frac{1}{2}\right)^{10} = 0.0974 \tag{10.28}$$

The approximating normal distribution is N(12.5, 2.5). Using the continuity correction we require $P\left(14.5 \leq X \leq 15.5\right)$. *Standardizing* $z_1 = \frac{14.5-12.5}{2.5} = 0.8 \qquad z_2 = \frac{15.5-12.5}{2.5} = 1.2$

$$P\left(14.5 \leq X \leq 15.5\right) = P\left(0.8 \leq Z \leq 1.2\right) = 0.8849 - 0.7881 = 0.0968 \tag{10.29}$$

The approximation is quite good.

Example 58 *What is the probability of throwing 15 or more tails in 25 tosses of a fair coin?*

Solution.

The setup is the same as in the previous example but the exact answer now require us to calculate

$$\binom{25}{15}\left(\frac{1}{2}\right)^{25} + \binom{25}{16}\left(\frac{1}{2}\right)^{25} + \binom{25}{17}\left(\frac{1}{2}\right)^{25} + \ldots\ldots \tag{10.30}$$

which would be rather tedious even with a calculator.

Using the normal distribution approximation we require the sum of the probabilities of the ranges 14.5 - 15.5 ....24.5-∞ *(using the exact 25.5 renders the calculation more tiresome and serves no useful purpose since the probabilities are so small at the end of the range.) this is just* $P\left(X \geq 14.5\right) = P\left(Z \geq 0.8\right) = 1 - 0.7881 = 0.212$

10.2.1 Exercises

1. A die is thrown 1200 times. Use the normal approximation to find the probability that the number of sixes lies in the range 191 - 209.

2. A fair coin is thrown 100 times. What is the probability that there will be fewer than 40 heads?

3. Patients suffering from a certain disease have a 60% chance of recovering without treatment. A new drug is being tested but in fact has no effect. What is the probability that 15 or more patients in a group of 20 taking the new drug will recover?

4. To assess "public opinion" 80 inquisitors each ask 50 people to give their opinion either for or against the immolation of math teachers as heretics. If on average 60% of the population are for the measure what is the probability that an individual inquisitor will report a majority against.? What is the probability that more than 10% of the inquisitors will report a majority against.

5. Using the data of Example 39.14 plot on the same axes the probabilities obtained by using (a) the exact binomial distribution and (b) the probabilities obtained by using the normal approximation. Plot for say 8 to 18 tails as the probabilities become insignificant outside this range. Plot the binomial distribution as a step function letting the value calculated for 12 be constant on the interval $11.5 \leq x \leq 12.5$ etc. This will give you an empirical justification for the technique. It also shows that in a situation like that of Example 39.15 the errors cancel out to some extent.

10.3 Reading: The Application of Probability to Conduct by Lord John Maynard Keynes

10.3.1 Introduction to the Reading

10.3.2 The Paper

Chapter 11

Probability Distributions: The Chi-Squared Distribution

11.1 The Chi-Squared Distribution

Example 59 *A die was thrown 600 times with the results below:*

score	*1*	*2*	*3*	*4*	*5*	*6*
frequency	*83*	*105*	*95*	*104*	*111*	*102*

(11.1)

Is this significant evidence at the 5% level that the die is biased?

Solution.

We will go about this in the usual way until we run into difficulties. Null hypothesis: the die is not biased. Under the null hypothesis the expected value of each frequency is 100. [This does not mean that the actual value of each frequency has to be 100 of course.] We tabulate the differences observed frequency O - expected frequency E:

	1	*2*	*3*	*4*	*5*	*6*	*Totals*
O	*83*	*105*	*95*	*104*	*111*	*102*	*600*
E	*100*	*100*	*100*	*100*	*100*	*100*	*100*
O - E	*- 17*	*5*	*- 5*	*4*	*11*	*2*	*0*

(11.2)

Note that the total of O - E must be zero whatever the results.

Since with obvious notation $E_1 + E_2 + E_3 + E_4 + E_5 + E_6 = 0$ *these differences are not independent. Given 5 of them we could calculate the missing one. This fact will have significance when we begin to cast our spells later.*

Clearly $\sum |O - E|$ *or* $\sum (O - E)^2$ *are both measures of the goodness of fit of the observed to the expected values. But the test statistic widely used in the spell books is* χ^2 *defined by* $\chi^2 = \sum \frac{(O-E)^2}{E}$. *With our data we have*

$$\chi^2 = \frac{(-17)^2}{100} + \frac{5^2}{100} + \frac{(-5)^2}{100} + \frac{4^2}{100} + \frac{11^2}{100} + \frac{2^2}{100} = 4.80 \tag{11.3}$$

[The theory is rather ferocious. It is not unusual to see statisticians emerging bruised and bleeding from encounters with the χ^2 distribution. One unfortunate economic statistician was reported to have committed suicide after a nightmare in which he was attacked by three simultaneous χ^2 distributions with contradictory parameters.

For there is actually a complete bestiary of the things χ_ν for $\nu = 1, 2, \cdots$ and the definition is $\chi_\nu^2 = Z_1^2 + Z_2^2 + \cdots\cdots + Z_\nu^2$ where each Z_i is a copy of the standard normal distribution. For the connoisseur of Gaussian integrals the pdf can be found to be

$$f\left(\chi_\nu^2\right) = ke^{-\chi^2/2}\left(\chi^2\right)^{\frac{\nu}{2}-1} \qquad 0 \leq \chi^2 < \infty \tag{11.4}$$

For ν ¿ 2 its mean is ν and its variance is 2ν.

The cumulative distribution as might be expected can only be tabulated and in practice this is only done for the key percentage points as in the case of the t distribution. As indicated at the top of the tables the χ^2 distributions are not symmetrical. For example for χ_4 at 99% the value is 0.297 meaning that 99% of χ_4^2 values are $\geq$ 0.297. At 1% the value is 13.28 meaning that 1% of the χ_4^2 values are $\geq$ 13.18. Study the tables with this in mind.]

Returning to our problem it can be shown that for any set of observations $\sum\limits_{i=1}^{n} \frac{(O-E)^2}{E}$ is an approximate χ_ν^2 variate.The next problem is to find for which ν. ν is called the number of degrees of freedom.n is called the number of cells.

Then $\nu = n$ - the number of constraints. A constraint is defined as any parameter of the theoretical distribution which could be *calculated* from the data. In this example it is just the total number of observations (600) . (We have calculated that the ideal probabilities are 1/6 and the expectation values from the null hypothesis not from the data.) Thus ν = 6 - 1 = 5 and we have $\chi_5^2 = 4.80$

For χ_5^2 at 5% we have "only 5% of χ_5^2 values are greater than 11.07" i.e. at the 5% level we would only reject the null hypothesis if χ_5^2 ¿ 11.07.

Goodness of fit problems are essentially one tailed. All that is needed is care in interpreting the tea leaves (i.e. the tables). For this example we conclude that this experiment does not provide evidence that the die is biased.

[Fiddling with data in order to prove a point is not of course unknown. The smallness of our test value $\chi_5^2 = 4.8$ relative to the critical value 11.07 might cause suspicion that the data had been fiddled. A first way to check this is to look at the other end of the χ_5^2 distribution. At 99% 97.5% and 95% we find χ_5^2 = 0.554, 0.831and 1.15 respectively. These tell us e.g. that only 5% of the results should be less than 1.15. This enables us to conclude that our 4.80 has probably been obtained without too much fiddling.

Helpful Unproved Hints:

1. $\sum \frac{(O-E)^2}{E}$ gives a sound measure of the goodness of fit when used as above and heeding the advice below.

2. No cells should have values of E ¡ 1 and it is better if no cells have E ¡ 5. (There is one cell to each value considered i.e. N cells for an index running from i to N). If this criterion is not met you might try lumping some cells together.

3. The total number of observations should be ¿ 45.

4. If $\nu = 1$ forget it.

Example 60 *The number of girls in 368 families with 5 children is tabulated. We would like to fit a binomial distribution to this data. From the data in the observed line of the table below we easily find that the mean is 2.5163. If this really were a binomial distribution then the formula $\mu = np$ would give us p = 0.5033. Hence in the E line we have calculated the expected numbers of families with 0 to 5 girls from a collection of 368 families using B(5,0.5033) e.g.*

	0	1	2	3	4	5	*Total*
O(actual no. of girls)	*12*	*58*	*113*	*111*	*61*	*13*	*368*
E(theoretical no. 1dp.)	*11.1*	*56.4*	*114.2*	*115.7*	*58.6*	*11.8*	*367.8*
O - E	*0.9*	*1.6*	*- 1.2*	*- 4.7*	*2.4*	*1.2*	*0.2*

(11.5)

[A note on the rounding: Obviously the observed values are integers. The statisticians rule is to calculate theoretical expectation values to one more decimal place of precision than the observed values. So here 1 decimal place.]

Using the above values

$$\chi^2 = \frac{0.9^2}{11.1} + \frac{1.6^2}{56.4} + \frac{1.2^2}{114.2} + \frac{4.7^2}{115.7} + \frac{2.4^2}{58.6} + \frac{1.2^2}{11.8} = \underline{\underline{0.542}} \qquad (11.6)$$

There are 6 classes.

There are 2 constraints in this problem. Clearly $\sum E$ is calculated from the data to give one constraint as before. Also the mean of the theoretical distribution is calculated from the data. p of course need not be counted since it follows at once from the mean and E.

So $\nu = 4$ and our test statistic is $\chi^2_4 = 0.542$. From the tables we see that approximately 96% of the values are greater than this value of χ^2_4. This is a very good fit indeed and might lead one to conclude that the data had been fiddled. [In fact the example comes from a book where the authors did in fact fiddle the data to avoid problems in an earlier section on the binomial distribution. They admit it later.]

Example 61 *Random samples of 10 grains of wheat were weighed in grams and the results divided into class intervals as in the table below.*

wt. of 10 grains (gms)	*2.1 - 2.2*	*2.2 - 2.3*	*2.3 - 2.4*	*2.4 - 2.5*	*2.5 - 2.6*	*2.6 - 2.7*	*2.7 - 2.8*	*Total*
No. of samples	*5*	*46*	*183*	*342*	*289*	*115*	*20*	*1000*

(11.7)

Find the mean and standard deviation of these sample weights. Test at the 5% significance level the goodness of fit of the data with a normal distribution of the same mean and standard deviation.

Solution.

We use the mid points of the class intervals as random variable i.e. 2.15, 2.25, 2.35, 2.45, 2.55, 2.65, 2.75.

By using a guessed mean of 2.45 we have the table

x	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
f	5	46	183	342	289	115	20

(11.8)

$$\textit{Thus} \sum fx = 28.9 \textit{ and } \mu_x = 28.9/1000 = 0.0289 \qquad (11.9)$$

Whence the actual mean is 2.479.

$$\textit{Also} \sum fx^2 = 13.41 \textit{ and } \sigma^2 = E\left(X^2\right) - \mu_x^2 = \frac{13.41}{1000} - (0.0289)^2 = 0.0126 \quad (11.10)$$

The variance is not affected by a translation of the mean. So we have a mean of 2.479 and a standard deviation of 0.112 gm.

On the assumption that we have a normal distribution with the above mean and standard deviation we make the following calculations:

$P(w \leq 2.1) = P(z \leq \frac{2.1-2.479}{0.112}) = P(z \leq -3.3839) = 0.0004$
$P(w \leq 2.2) = P(z \leq -2.4911) = 0.0064$
$P(w \leq 2.3) = P(z \leq -1.5982) = 0.055$
$P(w \leq 2.4) = P(z \leq -0.7054) = 0.2256$
$P(w \leq 2.5) = P(z \leq 0.1875) = 0.5744$
$P(w \leq 2.6) = P(z \leq 1.0808) = 0.8601$
$P(w \leq 2.7) = P(z \leq 1.9732) = 0.9758$
$P(w \leq 2.8) = P(z \leq 2.8661) = 0.9980$

The expected frequencies then are calculated as (0.0064 - 0.0004) x 1000 etc.

We now have the following table:

wt. class	*2.15*	*2.25*	*2.35*	*2.45*	*2.55*	*2.65*	*2.75*	
O	*5*	*46*	*183*	*342*	*289*	*115*	*20*	*1000*
E	*6*	*48.6*	*170.6*	*348*	*285.7*	*115.7*	*22.2*	*996.8*

(11.11)

The test statistic is then

$$\chi^2 = \frac{1^2}{6} + \frac{2.6^2}{48.6} + \frac{12.4^2}{170.6} + \frac{6^2}{348} + \frac{3.3^2}{285.7} + \frac{0.7^2}{115.7} + \frac{2.2^2}{22.2} = \underline{\underline{1.57}} \quad (11.12)$$

We have 7 cells and three constraints (mean, variance and totals)

So we look for χ_4^2 at the 5% level finding 9.49.

Thus the critical region is χ_4^2 ¿ 9.49 and so we accept that we have a good fit.

(At 95% $\chi_4^2 = 0.711$ so we need not be too worried that this result is too good to be true.)

Another use of the χ^2 distribution is when working with contingency tables. Again we will illustrate how to cast the appropriate spells rather than try to build up the general theory.

Example 62 *In 1979 petrol tanker (gasoline truck) drivers voted on whether or not to accept a pay offer. The results were tabulated by company as follows.*

	Shell	*Esso*	*BP*	*Totals*
For	*1301*	*1058*	*1204*	*3563*
Against	*976*	*543*	*699*	*2218*
Total	*2277*	*1601*	*1903*	*5781*

(11.13)

This is called a contingency table. The essential feature of what makes it a contingency table appears to be that it makes sense to sum both vertically and horizontally. In our previous examples in this section vertical summation would have produced meaningless figures.

The question arising out of this particular contingency table is: "Is there evidence of any difference of response between drivers of different companies?"

The null hypothesis is that there is no difference. Now 3563 out of 5781 drivers voted 'for' and 2218 out of 5781 voted 'against'. Since there are 2277 Shell drivers then if the null hypothesis is true we would expect $\frac{3563}{5781} \times 2277 = 1403.4$ *to vote 'for'. In this way we complete the following table of expected values on the null hypothesis.*

	Shell	Esso	BP	Total
For	1403.4	986.7	1172.9	3563
Against	873.6	614.3	730.1	2218
Total	2277	1601	1903	5781

(11.14)

Now we tabulate the O - E values for these results.

	Shell	Esso	BP	Total
O - E for	- 102.4	71.3	31.1	0
O - E against	102.4	- 71.3	- 31.1	0
Total	0	0	0	0

(11.15)

[A little thought will convince you that the results must come out in this symmetric manner. Doing the complete calculation serves as a check on the arithmetic to this point.] χ^2 *computed from these differences is*

$$\chi^2 = \frac{(102.4)^2}{1403.4} + \frac{(102.4)^2}{873.6} + \frac{(71.3)^2}{986.7} + \frac{(71.3)^2}{614.3} + \frac{(31.1)^2}{1172.9} + \frac{(31.1)^2}{730.1} = \underline{\underline{35.05}} \quad (11.16)$$

The number of cells is obviously 6. We need to be careful in deciding the number of constraints.

1. The grand total 5781 is calculated from the data.

2. 1 row total. (When this has been computed the other row total can be obtained from it and the grand total with no further reference to the data.)

3. 2 column totals. (From these two the third can be found as above.)

Thus there are 4 independent constraints and $\nu = 2$.

For χ^2_2 *only 0.1% of the results should exceed 13.81 so our result is highly significant that the null hypothesis is not true and that there was a significant difference in the company voting patterns. I cannot resist pointing out that this is self evident if one merely computes the percentage from each company voting 'for'.*

11.1.1 Exercises

1. How many degrees of freedom are required for a χ^2 test, for testing data against the following distributions? Where parameters are known independently of the data they are given.

   (a) Binomial, 8 cells, p = 5/8

   (b) Binomial, 10 cells.

   (c) Normal, 12 cells

   (d) Normal, 7 cells, $\mu = 4.6$

(e) Normal, 15 cells, $\mu = 2.8$, $\sigma = 5.2$

(f) Poisson, 6 cells, $\lambda = 2.13$

(g) Poisson 8 cells.

2. Use tables to look up the following values. State whether the results are significant at the 5% or 1% levels, or whether they seem too good to be true.

   (a) $\chi_4^2 = 9.60$

   (b) $\chi_{11}^2 = 2.51$

   (c) $\chi_{20}^2 = 26.52$

   (d) $\chi_{12}^2 = 36.04$

3. Look up the following table values

   (a) χ_{12}^2 at 5%

   (b) χ_9^2 at 5%

   (c) χ_{100}^2 at 1%

   (d) χ_6^2 at 99%

4. How many degrees of freedom are required before $\chi^2 = 20.04$ is significant

   (a) at the 5% level

   (b) at the 1% level?

5. An experiment which was inconclusive was repeated three times and each time the value of χ^2 was computed. First experiment $\chi^2 = 9.32$ $\nu = 6$. Second experiment $\chi^2 = 16.51$ $\nu = 11$. Third experiment $\chi^2 = 13.82$ $\nu = 9$. What conclusion can be drawn?

6. A repeated experiment gave $\chi_1^2 = 3.29$ and then and then $\chi_3^2 = 6.64$. Is this result significant at the 5% level?

7. Five coins are tossed 320 times. The number of heads is as given:

Number of heads	0	1	2	3	4	5	
Frequency	11	55	107	102	38	7	Total 320

(11.17)

   Is this evidence of bias?

8. A die is tossed 120 times with the results given:

Score	1	2	3	4	5	6	
Frequency	21	17	24	22	15	21	Total 120

(11.18)

   Is this significant evidence at the 5% level that the die is not fair?

9. One hundred observations of a variable x were as listed:

x	0	1	2	3	4	5 or more	
Frequency	8	25	32	14	5	16	Total 100

(11.19)

Is it possible that these observations came from a Poisson distribution with mean 2?

10. Just before a by-election in a certain constituency, a public opinion poll gave the following percentages which are compared with the eventual by-election results. Is the difference significant if the poll was based on (a) 100 voters, (b) 1000 voters?

	Labour	Conservative	Alliance	SNP	
Opinion poll %	32	29	11	28	
By - election %	25	32	7	36	Total 100%

(11.20)

If the difference is to be significant at the 5% level, what is the smallest sample size for the opinion poll?

11. The women belonging to the Women's Institute in two villages were asked where they did most of their shopping.

	Shop locally	At nearby town	Total
Village A	51	37	88
Village B	85	71	156
Total	136	108	244

(11.21)

Is this significant evidence at the 5% level of different shopping habits in the two villages.

12. After a flu epidemic, 103 people were chosen at random and asked about the severity of their attack and also their previous history of flu.

	Never before	Previous attack more than5 years ago	Previous attack within 5 years	Total
Free	15	16	14	45
Slight	9	12	8	29
Severe	15	9	5	29
Total	39	37	27	103

(11.22)

Is this evidence significant at the 5% level of an association between this type of flu and a previous attack?

13. Three classes of different ages in a school were asked if they had been to the cinema in the previous month.

	More than once	Once only	Not at all	Total
Age 11	5	14	13	32
Age 13	6	11	14	31
Age 15	8	12	10	30
Total	19	37	37	93

(11.23)

Is this significant evidence at the 5% level of a difference of cinema going habits?

14. The different grades awarded by three examiners in English O level papers were as follows:

Examiner	R	S	T	Total
Grade A	21	14	38	73
Grade B	49	40	92	181
Grade C	55	43	103	201
Grade D or less	61	48	122	231
Total	186	145	355	686

(11.24)

Is there evidence of apparent variability of standards? Can you offer a possible explanation?

15. Show that for a 2 x 2 contingency table

			Total
	a	b	$a+b$
	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

$$\chi^2 = \frac{(ad-bc)^2\,(a+b+c+d)}{(a+c)\,(a+b)\,(b+d)\,(c+d)} \tag{11.25}$$

11.2 Reading: Sociology Learns the Language of Mathematics by Abraham Kaplan

11.2.1 Introduction to the Reading

11.2.2 The Paper

Chapter 12

Limit Theorems

12.1 The Law of Large Numbers

We considered the law of large numbers philosophically before and we wish to revisit it here in mathematical language. We give two versions without proof.

Theorem 12 (Weak Law of Large Numbers) *We consider a family of random variables $\{X_n\}$ that are independent and identically distributed with mean μ and finite variance σ^2, then $\overline{X} \to^P \mu$.*

Note that the overline denotes the mean over the random variables. The $\to^P$ means that that the left-hand side converges *in probability* to the right-hand side; that is the probability that the difference between the two sides is larger than some ϵ (for any ϵ as small as desired) tends to zero.

In words the law says, the mean of a random sample of a population converges to the mean of the distribution of the population. Even simpler, more trials give a better average. We can use the central limit theorem (see below) to show (which we won't) that the expected deviation of $\overline{X}$ from μ is of the order of magnitude of $1/\sqrt{N}$ where N is the number of trials made. This answers the question about an error terms from some lectures back.

If we wish to be pedantic, the previous theorem did not say that $\overline{X}$ will get close and stay close to μ; it merely said that this becomes more and more likely. The next theorem actually says this outright but please note the slightly changed assumptions!

Theorem 13 (Strong Law of Large Numbers) *We consider a family of random variables $\{X_n\}$ that are independent and identically distributed with mean μ and finite fourth moment. Let $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$, then $\overline{X} \to^{a.s.} \mu$.*

Note that $\to^{a.s.}$ means that the actual difference between the left and right-hand sides tends to zero and not only the probability of that. This is the crucial difference but for this we must have a finite fourth moment and not merely a finite variance (second moment).

12.2 Central Limit Theorem

Let's consider a number of different random variables $X_1, X_2, \cdots, X_N$ all of which are independent and each random variable has an arbitrary probability distribution α_i with and finite variance σ^2.

Theorem 14 (Central Limit Theorem) *The distribution of the random variable*

$$X_{norm} = \frac{X_1 + X_2 + \cdots + X_N}{\sqrt{i}} \tag{12.1}$$

converges to the normal distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{x^2}{\sigma^2}\right) \tag{12.2}$$

as $i \to \infty$.

We will not prove this theorem simply because of the length of the proof and the fact that is an exercise in Fourier integration which serves no purpose in this course. This central limit theorem holds under some conditions. There are two conditions that are both equivalent and which we now state (note the difference in wording that one uses the word "every" and the other "some"). We do this for completeness. There is no need to fully understand where these conditions come from. The discussion below will illustrate the theorem in more detail.

Theorem 15 (Lindeberg's Condition) *If, for every ϵ,*

$$\lim_{n\to\infty} \frac{1}{S_n^2} \sum_{i=1}^{n} \int\limits_{|x|\geq\epsilon S_n} x^2 d\alpha_i = 0 \tag{12.3}$$

then the central limit theorem holds.

Theorem 16 (Lyapunov's Condition) *If, for some ϵ,*

$$\lim_{n\to\infty} \frac{1}{S_n^{2+\epsilon}} \sum_{i=1}^{n} \int |x|^{2+\epsilon} d\alpha_i = 0 \tag{12.4}$$

then the central limit theorem holds.

The theorem is a bit of a mouthful and so we will reformulate it in different words. We do this by means of an example. Consider the random variable X_1 to be the value of a die (1, 2, 3, 4, 5 or 6), X_2 the side of a thrown coin (heads or tails), X_3 to be the time between two successive atomic decays of slab of uranium and X_4 to be the height of a randomly selected man. All four of these random variables are independent in that none influence the other by any means whatsoever. All of them have a probability distribution function. The first two have a uniform distribution, the third has a complicated one which we will not describe and the fourth has (by all accounts) a normal distribution. All distributions have a known mean and variance and the variances are all finite. So far, we have checked that the conditions of the theorem apply. This is very important because there are some circumstances when one or more of these conditions are false and the theorem does not apply.

Thus the random variable calculated from our four random variables and given by the equation 12.1 has a distribution function that, as the sample size gets large, is normally distributed. In this example, this knowledge does not get us very far but let us give another example to illustrate it further.

We have established that, as far as we can tell, the distribution of male heights is normal. We want to know why this is so. For this purpose we list a number of factors that influence a particular man's height:

1. The food regularly eaten.

2. The wealth of his family.

3. The place of residence.

4. His habits (smoking, sports, etc.)

Clearly some influence it more than others, some do not apply all the time, some can not easily be measured, others can not even be quantified in an established way. Basically the reason why a man is of a particular height is a mystery as far as predicting it goes. What we can say is that each of the factors can be viewed as a random variable with a certain distribution function. The central limit theorem now basically means that the heights of men are going to be normally distributed if only you measure enough of them.

As this can be done for virtually anything, the theorem essentially says: Anything complicated is almost always distributed normally if only the sample size is large enough and none of the contributing factors have infinite variance. This is not entirely and exactly true as the variate which is distributed normally is a very particular one (see equation 12.1). It is important to note however that it is in this sense that the theorem is almost always interpreted. The reason being that, in practise, the distribution functions and means etc. are rarely known for all the contributing factors. This this interpretation is not correct and that it leads to wrong conclusions in some circumstances is vital to be understood.

The conditions of the theorem must be checked, all the factors together with their means and variances known and then a particular variate is normally distributed if the sample is large enough. Note that the theorem makes no statement about another variate and it does also not tell you how large the sample has to be in order for the approximation to be good enough for your practical purposes. In short, this statement has to be taken with a lot of salt and is worth considerably less than is usually claimed and advertised.

12.3 Reading: Mathematics of War and Foreign Politics by Lewis Fry Richardson

12.3.1 Introduction to the Reading

12.3.2 The Paper

Part II

Statistics

Chapter 13

The Importance of Sampling

13.1 Introduction

Data consist of numbers, of course. But these numbers are fed into the computer, not produced by it. These are numbers to be treated with considerable respect, neither to be tampered with, nor subjected to a numerical process whose character you do not completely understand.

The analysis of data inevitably involves some trafficking with the field of *statistics*, that gray area which is not quite a branch of mathematics - and just as surely not quite a branch of science. In the following sections, you will repeatedly encounter the following paradigm:

1. apply some formula to the data to compute “a statistic”
2. compute where the value of that statistic falls in a probability distribution that is computed on the basis of some “null hypothesis”
3. if it falls in a very unlikely spot, way out on a tail of the distribution, conclude that the null hypothesis is *false* for your data set.

If a statistic falls in a *reasonable* part of the distribution, you must not make the mistake of concluding that the null hypothesis is “verified” or “proved.” That is the curse of statistics; that it can never prove things, only disprove them! At best, you can substantiate a hypothesis by ruling out, statistically, a whole long list of competing hypotheses, every one that has ever been proposed. After a while your adversaries and competitors will give up trying to think of alternative hypotheses, or else they will grow old and die, and *then your hypothesis will become accepted.* Sounds crazy, we know, but that’s how science works!

13.2 The Field of Statistics

All statistics aims at a single goal: *Validating some hypothesis about the real world by computing something from data.*

As such it suffers from a number of ailments: (1) A single counterexample can destroy a hypothesis but (2) this hypothesis can never be proven true unless it is about a finite number of things that can all be checked without error, (3) measurements in the real world usually come with errors and uncertainties and (4) the hypothesis must be phrased in a numerical way so that it can be checked against calculations. As statistics involves a lot of data and complicated computations only the final results of which are ever reported, there is a large scope for bias and cheating.

Usually statistics begins when the data is available for treatment and considers only the calculation of some figures of merit to be considered for the validation of the hypothesis. It is however quite important to consider how the data is obtained. If the hypothesis refers to a small number of objects, they can all be investigated and the hypothesis established as absolutely true or false. This is not a part of statistics. The question whether a hypothesis is true or false becomes a statistical one if we cannot measure all members of the population (the collection of objects of interest). We may establish that the hypothesis is false or that it is true with a to-be-determined degree of confidence.

Thus we must select from the population a few members that we will investigate and base our conclusions upon. This selection is called *sampling.* This is wrought with difficulties that we will discuss in a moment. It is clear that sampling is necessary.

13.3 Sampling Methods

13.3.1 Unbiased Samples

One of the most popular ways to cheat in statistics is to choose a biased sample. Suppose we are being paid by the American beef industry to establish the hypothesis that “eating meat makes you tall.” The implicit desire is to validate this and already we are in a fix because the scientific impartial attitude is lacking. Suppose we choose three samples of people: (1) only Americans, (2) only rural Africans and (3) a truly representative sample of the whole world. We shall find that the hypothesis is very likely to be true in the first case, almost totally wrong in the second and weakly true in the third. The reason is, of course, that it is not the meat that makes people tall but the growth hormones feed to the cows to make them grow fat and large quickly that are still contained in the meat. It is however only the industrial nations (first and foremost the USA) that can afford to do so. Rural Africa for example farms almost exclusively organically where cows eat grass and other naturally occurring foods. In conclusion, we must choose the first option to get what we want but we have to be aware that we are selling a lie because the hypothesis that we have *really* validated is: “eating American meat makes Americans tall.” Furthermore, the statistics say nothing about health and related factors *but* they can be made to be implied depending on how the results are written up for the newspapers. In this way, much nonsense can be promoted.

From this example we define

Definition 22 *An* unbiased *sample from a population is such that all properties calculated from the sample are identical within a previously agreed (and clearly reported) margin of error to these properties calculated for the whole population.*

As we cannot, in general, calculate the property of interest for the whole population (this being the point behind sampling in the first place) we must agree on methods that are going to produce an unbiased sample.

It is clear that *a biased sample may or may not give rise to a different conclusion with regard to a particular hypothesis than an unbiased sample.* As we are interested in the hypothesis as regards the population (and *not* the sample), *a biased sample is a wrong sample.*

Below we give some basic sampling methods. In real life, usually a combination of these is actually used for practical reasons and limitation of resources.

13.3.2 Uniformly Random Sampling

The simplest way to obtain an unbiased sample is to randomly choose members from the population, this is known as *random sampling.* If each member of the population is equally likely to be selected, it is *uniformly random sampling.*

This clearly suffers from the requirement that we must be in possession of a list of all members of the population! If we do not have such a list, random selection is impossible as there is nothing to be selected from.

To demonstrate the difficulties in organizing a truly uniformly random sampling strategy we give a few examples:

1. For a questionnaire at a supermarket, you position one person at each exit and entry point. This is not uniform as there is bound to be a more popular exit than others and so the probability of being selected there is lower than at the others. Furthermore the probability of being selected increases if, for some reason, a person has to enter and exit several times. Naturally this also depends on other factors such as time of day, weekday and so on.

2. For a survey, you phone people during the day. This is not uniform as some people do not have phones and many are not at home when you try to phone them.

3. For a medical study, you pay people a significant sum of money. This is not uniform as people with time but no money are more likely to join thus the study is biased towards the unemployed and poorer sections of the public which may or may not correlate with poorer health in general.

If the method of random sampling is not uniform, then the method is clearly biased towards that group that has a higher chance of being selected. The complexity is best illustrated by an unanswered exercise.

Example 63 *You are to do an election poll by asking 4000 people about their opinions regarding the candidates running for President in a country of more than 50 million people. The country is broken up into dozens of electoral districts each of which are counted separately. You want a representative sample that takes account of several factors: Geographic location (i.e. electoral district), age, financial resources and broad work category. How do you select these 4000 people?*

The answer will come in the next few sections.

Uniform random sampling is the best method theoretically speaking. Practical statistics claims that other methods are better but the list of reasons are all reasons of management. In short, unless you are severely strapped for cash or time and you have a list of the population, do uniform random sampling.

13.3.3 Stratified Random Sampling

This method divides the population into *strata* that are groups of individuals such as the electoral districts of a country. Then uniform random sampling is applied to each stratum. If the strata are large enough and the number of individuals selected proportional to the size of their stratum, then stratified random sampling approximates uniform random sampling. Often this is not the case as, for example, in USA presidential elections in which certain states have a higher proportion of votes than proportion of population. This gives rise to biased statistics. Whether or not this is wanted is up to the particular case at hand but it is generally considered bad practise to over-represent one group without explicitly mentioning this fact. So be careful to state what you have done if you have chosen a bias.

Clearly we have learned by now that the actual calculations involved in statistics are very simple. The complication comes in the interpretation of the results. Biased sampling may or may not lead to a legitimate and meaningful answer to a relevant question depending on this interpretation. One may choose this method over the uniform one if one is particularly interested in the distribution of means for the same question over the different strata as is the case, for example, for political parties that want to know which districts contains many supporters and which do not.

13.3.4 Cluster Random Sampling

Superficially this method looks the same as stratified random sampling but it is very different. Here the population is broken into *clusters* that are groups of individuals. The clusters to be measured are selected using uniform random sampling and then a census is taken in the selected clusters.

Definition 23 *A* census *is an exhaustive measurement that considers every individual in the group.*

This method suffers from the same caveats as the stratified sampling and may be chosen principally for financial reasons. It is simply easier to first limit oneself geographically to a few regions and then exhaustively question all individuals in that region.

13.3.5 Systematic Sampling

Suppose you have a list of all individuals and you pick every p^{th} individual for questioning. If there are L individuals on the list, you will achieve a sample size of $N = L/p$. This is known as systematic sampling. Generally the first individual in the sample (among the first p individuals on the list) is chosen randomly but not always. Often this method gives a biased sample because the list is usually arranged in some organized fashion.

13.4 The Question of Error

Suppose you are to be tested for a certain chemical in your blood. It stands to be assumed that you would prefer to give a sample rather than the population of your blood to be measured. In cases such as this, sampling methods are important. Practically, it is very rare when the whole population can be measured. All laboratory science includes a sampling process and so this topic is of vital importance.

Particularly, we want to know what a certain value is of the population and not the sample while only being able to measure the sample. That is, we want to know how far off the measured value is from the value of the population that we really want. The answer is very surprising.

Theorem 17 *If uniform random sampling is used on a population and N individuals are selected and measured to obtain a sample, the error is $1/\sqrt{N}$.*

If $N = 1600$, the error is thus $1/40 = 0.025 = 2.5\%$. If this sample had an average of 100 for some quantity, then the average of the population is going to lie in the interval $100 - 2.5\%$ to $100 + 2.5\%$, i.e. 97.5 to 102.5. This answer is surprising for the simple reason that it does not depend on the size of the population. Thus a sample of 100 individuals from a population of 1000 or a population of one million has the same error!

One could now speculate on the general sociological relevance of psychology experiments that are done with 15 psychology students that all need the money being paid for it...

13.5 Example of Bad Sampling

Each of the following examples are true historical cases. Read each carefully and try to determine the errors and how to correct them.

Example 64 *According to a poll taken among scientists and reported in the prestigious journal Science, scientists do not have much faith in either the public or the media. The article reported that, based on the results of a "recent survey of 1400 professionals" in science and in journalism, 82% of scientists "strongly or somewhat agree" with the statement "The U.S. public is gullible and believes in miracle cures and easy solutions," and 80% agreed that "the public doesn't understand the importance of federal funding for research." About the same percentage (82%) also trashed the media, agreeing with the statement that "The media do not understand statistics well enough to explain new findings."*

It is not until the end of the article that we learn who responded: "The study reported a 34% response rate among scientists, and the typical respondent was a white, male physical scientist over the age of 50 doing basic research." Remember that those who feel strongly about the issues in a survey are the most likely to respond. Clearly this is unrepresentative and the numbers reported are essentially meaningless as far as concluding anything about scientists in general.

Example 65 *On February 18, 1993, shortly after Bill Clinton became President of the United States, a television station in Sacramento, California, asked viewers to respond to the question: "Do you support the president's economic plan?" The next day, the results of a properly conducted study asking the same question were published in the newspaper. Here are the results:*

Answer	*TV Poll*	*Survey*
Yes, support	*42%*	*75%*
No, object	*58%*	*18%*
Not sure	*0%*	*7%*

The people dissatisfied with the plan were more likely do call the TV station and no one called in to say that they were not sure. In short, such polls are merely a count of received answers and not statistics at all. No conclusions as regards the general population can be made from them. It is irresponsible behavior to publish such numbers without the appropriate caveats.

Example 66 *Some years ago, the student newspaper at a California university announced as a front page headline: " Students ignorant, survey says." The article explained that a "random survey" indicated that American students were less aware of current events than international students were. However, the article quoted the undergraduate researchers, who were international students themselves, as saying that "the students were randomly sampled on the quad." The quad is an open-air, grassy area where students relax, eat lunch, and so on.*

Such samples are called haphazard samples *and can not be expected to representative of anything. The researchers are likely to approach friendly looking people who will support what the researchers want to establish. This is the same disaster as psychology studies being done on psychology students at university and leads to a waste of time.*

Example 67 *A much older example that initiated much of the study of sampling (in the USA) is as follows. Before the Presidential election of 1936 in the USA, a contest between Democratic incumbent Franklin Delano Roosevelt and Republican Alf Landon, the magazine Literary Digest had been extremely successful in predicting the results in U.S. presidential elections. But 1936 turned out to be the year of their downfall, when they predicted a 3-to-2 victory for Landon. To add insult to injury, young pollster George Gallup, who had just founded the American Institute of Public Opinion in 1935, not only correctly predicted Roosevelt as the winner of the election, he also predicted that the Literary Digest would get it wrong. He did this before they even conducted their poll. And Gallup surveyed only 50,000 people, whereas the Literary Digest sent questionnaires to 10 million people.*

The Literary Digest made two classic mistakes. First, the lists of people to whom they mailed the 10 million questionnaires where taken from magazine subscribers, car owners, telephone directories, and, in just a few cases, lists of registered voters. In 1936, those who owned telephones or cars, or subscribed to magazines, were more likely to be wealthy individuals who were not happy with the Democratic incumbent.

Despite what many accounts of this famous story conclude, the bias produced by the more affluent list was not likely to have been as severe as the second problem. The *main problem was volunteer response. The magazine received 2.3 million responses, a response rate of only 23%. Those who felt strongly about the outcome of the election were most likely to respond. And that included a majority of those who wanted a change, the Landon supporters. Those who were happy with the incumbent were less likely to bother to respond.*

Gallup, however, knew the value of random sampling. He was able not only to predict the election, but to predict the results of the Literary Digest poll within 1%. How did he do this? He just chose 3000 people at random from the same lists the Digest was going to use, and mailed them all a postcard asking them how they planned to vote.

This example illustrates the beauty of random sampling and idiocy of trying to base conclusions on non-random and biased samples. The Literary Digest went bankrupt the following year, and so never had a chance to revise its methods. The organization founded by George Gallup has flourished, although not without making a few sampling blunders of its own.

13.6 Reading: Sampling and Standard Error by L. C. Tippett

13.6.1 Introduction to the Reading

13.6.2 The Paper

Chapter 14

Hypothesis Testing: An Introduction

14.1 The Meaning of a Test

14.1.1 Hypotheses and Errors

Suppose you have collected a set of data and you make an observation about it. You would like to measure (quantitatively and objectively) how significant that observation is. A very significant observation is one that holds true in almost all cases.

Depending on the test, the answer will be numerical and will have to be interpreted. Clearly four different types of situation can occur:

1. Hypothesis true, test answer positive
2. Hypothesis true, test answer negative
3. Hypothesis false, test answer positive
4. Hypothesis false, test answer negative

Situations 1 and 4 are good in the sense that the test has told us the correct status of the hypothesis. Situations 2 and 3 are errors of *type I* and *type II* respectively; i.e. the test has given the wrong answer. For each test, there is a certain probability that it will give the wrong conclusion and we name these probabilities according to their types.

Definition 24 *The* sensitivity *of a test is the probability that the test returns a positive answer for a true hypothesis. The* 1-sensitivity *of a test is the probability that the test returns a negative answer for a false hypothesis. The* specificity *of a test is the probability that the test returns a negative answer for a true hypothesis. The* 1-specificity *of a test is the probability that the test returns a positive answer for a false hypothesis.*

We do not want errors to occur but generally decreasing the 1-sensitivity increases the 1-specificity of a test. The only way to decrease both is to increase the sample size and this is not always possible. We must resign ourselves to the simple fact that statistics is fraught with error.

14.1.2 Levels of Significance

Generally each different test is useful for testing a particular hypothesis and usually some conditions of use apply. For most tests, you will have to calculate a certain value and then interpret whether this is "good" or "bad." Statistics are collected in order verify a hypothesis or in order to choose between several hypotheses. The hypothesis to be verified or falsified is called the *null hypothesis.* Null hypotheses can take on several of the below forms and others:

1. The mean of a certain value is x. (Males in Germany are 180cm tall on average.)

2. Two populations have the same mean. (Males in the USA and Germany have the same average height.)

3. Two populations have the same variance. (Ninety percent of Males in the USA and Germany are between 170cm and 190cm tall.)

4. Two distributions are the same. (Male heights in both the USA and Germany are normally distributed.)

Each type of null hypothesis has one or more associated tests. One computes a number and on this basis one decides to accept or reject the hypothesis. One generally has to specify a significance level for a test.

Definition 25 *The* significance level α *of a statistical test is the largest probability of a type I error, i.e. the largest 1-sensitivity, that we are willing to accept.*

Generally α is either 0.05 or 0.01 but this choice *is purely conventional* and there is no theoretical reason to choose either one of them over anything else. If we choose the 0.05 confidence level, we can be 95% *confident* that the statistical test will return the correct answer.

14.1.3 Tails of Tests

A distribution like the normal distribution has two tails. If we are concerned with both tails in relation to our hypothesis, the test is called a *two-tailed test* and if we are concerned only with one, it is called a *one-tailed test.* This situation will occur for example if we are testing whether men in the USA are taller than those in Germany (one-tailed) or whether the men in the USA are taller or shorter than those in Germany (two-tailed). This changes the assessment of the significance level markedly in general and thus needs to be carefully taken account of.

14.1.4 An Example of a Test

Suppose, under some hypothesis, we have that a given sample statistic S is normally distributed with mean μ and standard deviation σ that have both been computed from the sample data. For this test, we need the concept of a z score.

Definition 26 *A z-score is a standardized variable that is normally distributed with mean 0 and variance 1.*

From the sample statistic, we obtain the z-score by

$$z = \frac{S - \mu}{\sigma} \tag{14.1}$$

The hypothesis should be accepted at the 0.05 significance level if the z-score lies somewhere in the symmetric region covering 95% of the area underneath the standard normal distribution (mean 0, variance 1). That is,

$$\int_{-p}^{p} \frac{1}{2\pi} \exp\left(\frac{-1}{2z^2}\right) dz = 0.95 \tag{14.2}$$

that we need to solve for p. This can be done by looking at standard tables or using a computer program like Mathematica. Note that this integral, as many like it in statistics, can not be done in closed form. The answer is $p = 1.96$ but needs to be recomputed for a different significance level, of course, as the right hand side will differ. If the test is at the 0.05 significance level but is one-tailed instead of two-tailed, we must solve

$$\int_{-p}^{p} \frac{1}{2\pi} \exp\left(\frac{-1}{2z^2}\right) dz = 0.9 \tag{14.3}$$

and we obtain $p = 1.645$.

We are lead to accept the hypothesis if the statistic S has a z-score inside the region from -1.96 to 1.96 and reject it otherwise. Note that at the 0.01 confidence level 1.96 needs to be replaced by 2.58.

In the next lecture, we discuss a number of tests that one can use in practise to test for a number of commonly occurring null hypotheses.

14.2 Reading: Mathematics of a Lady Tasting Tea by Sir Ronald A. Fisher

14.2.1 Introduction to the Reading

14.2.2 The Paper

Chapter 15

Hypothesis Testing: Basic Tests

15.1 Testing Whether Two Distributions have the Same Mean

Not uncommonly we want to know whether two distributions have the same mean. For example, a first set of measured values may have been gathered before some event, a second set after it. We want to know whether the event, a "treatment" or a "change in a control parameter," made a difference.

Our first thought is to ask "how many standard deviations" one sample mean is from the other. That number may in fact be a useful thing to know. It does relate to the strength or "importance" of a difference of means *if that difference is genuine.* However, by itself, it says nothing about whether the difference *is* genuine, that is, statistically significant. A difference of means can be very small compared to the standard deviation, and yet very significant, if the number of data points is large. Conversely, a difference may be moderately large but not significant, if the data are sparse. We will be meeting these distinct concepts of *strength* and *significance* several times in the next few sections.

A quantity that measures the significance of a difference of means is not the number of standard deviations that they are apart, but the number of so-called *standard errors* that they are apart. The standard error of a set of values measures the accuracy with which the sample mean estimates the population (or "true") mean. Typically the standard error is equal to the sample's standard deviation divided by the square root of the number of points in the sample.

15.1.1 Student's t-test for Significantly Different Means

Applying the concept of standard error, the conventional statistic for measuring the significance of a difference of means is termed *Student's t*. When the two distributions are thought to have the same variance, but possibly different means, then Student's t is computed as follows: First, estimate the standard error of the difference of the means, s_D, from the "pooled variance" by the formula

$$s_D = \sqrt{\frac{\sum\limits_{i\in A}^{N_A}(x_i-\overline{x_A})^2+\sum\limits_{i\in B}^{N_B}(x_i-\overline{x_B})^2}{N_A+N_B-2}\left(\frac{1}{N_A}+\frac{1}{N_B}\right)} \qquad (15.1)$$

where each sum is over the points in one sample, the first or second, each mean likewise refers to one sample or the other, and N_A and N_B are the numbers of points in the first and second samples, respectively. Second, compute t by

$$t = \frac{\overline{x_A} - \overline{x_B}}{s_D} \tag{15.2}$$

Third, evaluate the significance of this value of t for Student's distribution with $N_A + N_B - 2$ degrees of freedom, by equations (6.4.7) and (6.4.9), and by the routine betai (incomplete beta function) of 6.4.

The significance is a number between zero and one, and is the probability that $|t|$ could be this large or larger just by chance, for distributions with equal means. Therefore, a small numerical value of the significance (0.05 or 0.01) means that the observed difference is "very significant." The function $A(t|\nu)$ in equation (6.4.7) is one minus the significance.

The next case to consider is where the two distributions have significantly different variances, but we nevertheless want to know if their means are the same or different. (A treatment for baldness has caused some patients to *lose* all their hair and turned others into werewolves, but we want to know if it helps cure baldness *on the average*!) Be suspicious of the unequal-variance t-test: If two distributions have very different variances, then they may also be substantially different in shape; in that case, the difference of the means may not be a particularly useful thing to know. To find out whether the two data sets have variances that are significantly different, you use the *F-test*, described later on.

The relevant statistic for the unequal variance t-test is

$$t = \frac{\overline{x_A} - \overline{x_B}}{\sqrt{Var(x_A)/N_A + Var(x_B)/N_B}} \tag{15.3}$$

This statistic is distributed *approximately* as Student's t with a number of degrees of freedom equal to

$$\frac{\left(Var(x_A)/N_A + Var(x_B)/N_B\right)^2}{\frac{(Var(x_A)/N_A)^2}{N_A-1} + \frac{(Var(x_B)/N_B)^2}{N_B-1}} \tag{15.4}$$

Our final example of a Student's t test is the case of *paired samples*. Here we imagine that much of the variance in *both* samples is due to effects that are point-by-point identical in the two samples. For example, we might have two job candidates who have each been rated by the same ten members of a hiring committee. We want to know if the means of the ten scores differ significantly. We first try the t test above, and obtain a value of the probability that is not especially significant (e.g., > 0.05). But perhaps the significance is being washed out by the tendency of some committee members always to give high scores, others always to give low scores, which increases the apparent variance and thus decreases the significance of any difference in the means. We thus try the paired-sample formulas,

$$Cov\left(x_A, x_B\right) \equiv \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{Ai}-\overline{x_A}\right)\left(x_{Bi}-\overline{x_B}\right) \tag{15.5}$$

$$s_D = \left(\frac{Var(x_A)+Var(x_B)-2Cov\left(x_A, x_B\right)}{N}\right)^{1/2} \tag{15.6}$$

$$t = \frac{\overline{x_A}-\overline{x_B}}{s_D} \tag{15.7}$$

where N is the number in each sample (number of pairs). Notice that it is important that a particular value of i label the corresponding points in each sample, that is, the ones that are paired. The significance of the t statistic in 15.7 is evaluated for $N-1$ degrees of freedom.

15.2 Testing Whether Two Distributions have the Same Variance

15.2.1 F-Test for Significantly Different Variances

The *F-test* tests the hypothesis that two samples have different variances by trying to reject the null hypothesis that their variances are actually consistent. The statistic is the ratio of one variance to the other, so values either 1 or 1 will indicate very significant differences. The distribution of in the null case is given in equation (6.4.11), which is evaluated using the routine betai. In the most common case, we are willing to disprove the null hypothesis (of equal variances) by either very large or very small values of, so the correct significance is *two-tailed*, the sum of two incomplete beta functions. It turns out, by equation (6.4.3), that the two tails are always equal; we need compute only one, and double it. Occasionally, when the null hypothesis is strongly viable, the identity of the two tails can become confused, giving an indicated probability greater than one. Changing the probability to two minus itself correctly exchanges the tails.

15.3 Testing Whether Two Distributions are Different

Given two sets of data, we can generalize the questions asked in the previous section and ask the single question: Are the two sets drawn from the same distribution function, or from different distribution functions? Equivalently, in proper statistical language, "Can we disprove, to a certain required level of significance, the null hypothesis that two data sets are drawn from the same population distribution function?" Disproving the null hypothesis in effect proves that the data sets are from different distributions. Failing to disprove the null hypothesis, on the other hand, only shows that the data sets can be *consistent* with a single distribution function. One can never *prove* that two data sets come from a single distribution, since (e.g.) no practical amount of data can distinguish between two distributions which differ only by one part in 10^{10}.

Proving that two distributions are different, or showing that they are consistent, is a task that comes up all the time in many areas of research: Are the visible stars distributed uniformly in the sky? (That is, is the distribution of stars as a function of declination - position in the sky - the same as the distribution of sky area as a function of declination?) Are educational patterns the same in Brooklyn as in the Bronx? (That is, are the distributions of people as a function of last-grade-attended the same?) Do two brands of fluorescent lights have the same distribution of burnout times? Is the incidence of chicken pox the same for first-born, second-born, third-born children, etc.?

These four examples illustrate the four combinations arising from two different dichotomies: (1) The data are either continuous or binned. (2) Either we wish to compare one data set to a known distribution, or we wish to compare two equally unknown data sets. The data sets on fluorescent lights and on stars are continuous, since we can be given lists of individual burnout times or of stellar positions. The data sets on chicken pox and educational level are binned, since we are given tables of numbers of events in discrete categories: first-born, second-born, etc.; or 6th Grade, 7th Grade, etc. Stars and chicken pox, on the other hand, share the property that the null hypothesis is a known distribution (distribution of area in the sky, or incidence of chicken pox in the general population). Fluorescent lights and educational level involve the comparison of two equally unknown data sets (the two brands, or Brooklyn and the Bronx).

One can always turn continuous data into binned data, by grouping the events into specified ranges of the continuous variable(s): declinations between 0 and 10 degrees, 10 and 20, 20 and 30, etc. Binning involves a loss of information, however. Also, there is often considerable arbitrariness as to how the bins should be chosen. Along with many other investigators, we prefer to avoid unnecessary binning of data.

The accepted test for differences between binned distributions is the *chi-square test*. For continuous data as a function of a single variable, the most generally accepted test is the *Kolmogorov-Smirnov test*. We consider each in turn.

15.3.1 Chi-Square Test

Suppose that N_i is the number of events observed in the i^{th} bin, and that n_i is the number expected according to some known distribution. Note that the N_i's are integers, while the n_i's may not be. Then the chi-square statistic is

$$\chi^2 = \sum_i \frac{(N_i - n_i)^2}{n_i} \tag{15.8}$$

where the sum is over all bins. A large value of χ^2 indicates that the null hypothesis (that the N_i's are drawn from the population represented by the n_i's) is rather unlikely.

Any term j in 15.8 with $0 = n_j = N_j$ should be omitted from the sum. A term with $n_j = 0$, $N_j \neq 0$ gives an infinite χ^2, as it should, since in this case the N_i's cannot possibly be drawn from the n_i's!

The *chi-square probability function* $Q(\chi^2|\nu)$ is an incomplete gamma function, and was already discussed in 6.2 (see equation 6.2.18). Strictly speaking $Q(\chi^2|\nu)$ is the probability that the sum of the squares of ν random *normal* variables of unit variance (and zero mean) will be greater than χ^2. The terms in the sum 15.8 are not individually normal. However, if either the number of bins is large ($>> 1$), or the number of events in each bin is large ($>> 1$), then the chi-square probability function is a good approximation to the distribution of 15.8 in the case of the null hypothesis. Its use to estimate the significance of the chi-square test is standard.

The appropriate value of ν, the number of degrees of freedom, bears some additional discussion. If the data are collected with the model n_i's fixed - that is, not later renormalized to fit the total observed number of events $\sum N_i$ - then ν equals the number of bins N_B. (Note that this is *not* the total number of *events*!) Much more commonly, the n_i's are normalized after the fact so that their sum equals the sum of the N_i's. In this case the correct value for ν is $N_B - 1$, and the model is said to have one constraint. If the model that gives the n_i's has additional free parameters that were adjusted after the fact to agree with the data, then each of these additional "fitted" parameters decreases ν by one additional unit.

Next we consider the case of comparing *two* binned data sets. Let R_i be the number of events in bin i for the first data set, S_i the number of events in the same bin i for the second data set. Then the chi-square statistic is

$$\chi^2 = \sum_i \frac{(R_i + S_i)^2}{R_i + S_i} \tag{15.9}$$

Comparing 15.9 to 15.8, you should note that the denominator of 15.9 is *not* just the average of R_i and S_i (which would be an estimator of n_i in 15.8). Rather, it is twice the average, the sum. The reason is that each term in a chi-square sum is supposed to approximate the square of a normally distributed quantity with unit variance. The variance of the difference of two normal quantities is the sum of their individual variances, not the average.

If the data were collected in such a way that the sum of the R_i's is necessarily equal to the sum of S_i's, then the number of degrees of freedom is equal to one less than the number of bins, $N_B - 1$, the usual case. If this requirement were absent, then the number of degrees of freedom would be N_B. Example: A bird-watcher wants to know whether the distribution of sighted birds as a function of species is the same this year as last. Each bin corresponds to one species. If the bird-watcher takes his data to be the first 1000 birds that he saw in each year, then the number of degrees of freedom is $N_B - 1$. If he takes his data to be all the birds he saw on a random sample of days, the same days in each year, then the number of degrees of freedom is N_B. In this latter case, note that he is also testing whether the birds were more numerous overall in one year or the other: That is the extra degree of freedom. Of course, any additional constraints on the data set lower the number of degrees of freedom in accordance with their number.

Equation 15.9 applies to the case where the total number of data points is the same in the two binned sets. For unequal numbers of data points, the formula analogous to 15.9 is

$$\chi^2 = \sum_i \frac{\left(R_i\sqrt{S/R} - S_i\sqrt{R/S}\right)^2}{R_i + S_i} \tag{15.10}$$

where

$$R \equiv \sum_i R_i, \qquad S \equiv \sum_i S_i \tag{15.11}$$

are the respective numbers of data points.

15.3.2 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (or K-S) test is applicable to unbinned distributions that are functions of a single independent variable, that is, to data sets where each data point can be associated with a single number (lifetime of each lightbulb when it burns out, or declination of each star). In such cases, the list of data points can be easily converted to an unbiased estimator $S_N(x)$ of the *cumulative* distribution function of the probability distribution from which it was drawn: If the N events are located at values x_i, $i = 1, 2, \cdots, N$, then $S_N(x)$ is the function giving the fraction of data points to the left of a given value x. This function is obviously constant between consecutive (i.e., sorted into ascending order) x_i's, and jumps by the same constant $1/N$ at each x_i. (See Figure 14.3.1.)

Different distribution functions, or sets of data, give different cumulative distribution function estimates by the above procedure. However, all cumulative distribution functions agree at the smallest allowable value of x (where they are zero), and at the largest allowable value of x (where they are unity). (The smallest and largest values might of course be $\pm\infty$.) So it is the behavior between the largest and smallest values that distinguishes distributions.

One can think of any number of statistics to measure the overall difference between two cumulative distribution functions: the absolute value of the area between them, for example. Or their integrated mean square difference. The Kolmogorov-Smirnov D is a particularly simple measure: It is defined as the *maximum value* of the absolute difference between two cumulative distribution functions. Thus, for comparing one data set's $S_N(x)$ to a known cumulative distribution function $P(x)$, the K-S statistic is

$$D = \max_{-\infty<x<\infty} |S_N(x) - P(x)| \tag{15.12}$$

while for comparing two different cumulative distribution functions $S_{N_1}(x)$ and $S_{N_2}(x)$, the K-S statistic is

$$D = \max_{-\infty<x<\infty} |S_{N_1}(x) - S_{N_2}(x)| \tag{15.13}$$

What makes the K-S statistic useful is that *its* distribution in the case of the null hypothesis (data sets drawn from the same distribution) can be calculated, at least to useful approximation, thus giving the significance of any observed nonzero value of D. A central feature of the K-S test is that it is invariant under reparametrization of x; in other words, you can locally slide or stretch the x axis in Figure 15.1, and

Figure 15.1: Kolmogorov-Smirnov statistic D. A measured distribution of values in x (shown as N dots on the lower abscissa) is to be compared with a theoretical distribution whose cumulative probability distribution is plotted as $P(x)$. A step-function cumulative probability distribution $S_N(x)$ is constructed, one that rises an equal amount at each measured point. D is the greatest distance between the two cumulative distributions.

the maximum distance D remains unchanged. For example, you will get the same significance using x as using $\log x$.

The function that enters into the calculation of the significance can be written as the following sum:

$$Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2} \tag{15.14}$$

which is a monotonic function with the limiting values

$$Q_{KS}(0) = 1, \qquad Q_{KS}(\infty) = 0 \tag{15.15}$$

In terms of this function, the significance level of an observed value of D (as a disproof of the null hypothesis that the distributions are the same) is given approximately by the formula

$$\text{Probability}(D > \text{observed}) = Q_{KS}\left(\left[\sqrt{N_e} + 0.12 + 0.11/\sqrt{N_e}\right] D\right) \tag{15.16}$$

where N_e is the effective number of data points, $N_e = N$ for the case **??** of one distribution, and

$$N_e = \frac{N_1 N_2}{N_1 + N_2} \tag{15.17}$$

for the case 15.13 of two distributions, where N_1 is the number of data points in the first distribution, N_2 the number in the second. The nature of the approximation involved in 15.16 is that it becomes asymptotically accurate as the N_e becomes large, but is already quite good for $N_e \geq 4$, as small a number as one might ever actually use.

15.4 Reading: Foundations of Vital Statistics by John Graunt

15.4.1 Commentary on an Ingenious Army Captain and on a Generous and Many-sided Man

Statistics was founded by John Graunt of London, a "haberdasher of small-wares," in a tiny book called *Natural and Political Observations made upon the Bills of Mortality*[1]. It was the first attempt to interpret mass biological phenomena and

[1] The full title is, *Natural and Political Observations Mentioned in a following Index, and made upon the Bills of Mortality*

social behavior from numerical data—in this case, fairly crude figures of births and deaths in London from 1604 to 1661. Graunt's tract appeared in 1662. Thirty years later, the Royal Society published in its "Philosophical Transactions" a paper on mortality rates written by the eminent astronomer Edmund Halley. This famous article was entitled "An Estimate of the Degrees of the Mortality of Mankind, drawn from curious Tables of the Births and Funerals at the City of Breslaw; with an Attempt to ascertain the Prices of Annuities upon Lives." It was followed by "Some further Considerations on the Breslaw Bills of Mortality". Together, the papers are the foundation for all later work on life expectancy, indispensable of course to the solvency of life-insurance companies[2].

John Graunt was born in 1620 in Birchin Lane, London, "at the Sign of the Seven Stars," where his father kept a shop and home. He was early apprenticed to a merchant in small wares—buttons, needles and the like—and prospered in the trade. Success gave him the leisure to indulge interests somewhat broader than those of the notions counter. Aubrey describes him as "a very ingenious and studious person...[who] rose early—in the morning to his Study before shoptime."[3] He became a friend of Sir William Petty, later the author of a well—known book on the new study of political arithmetic, and probably discussed with him the ideas to be expressed in the *Observations.* The Bills of Mortality which attracted Graunt's attention were issued weekly by the company of parish clerks and listed the number of deaths in each parish, the causes, and also an "Accompt of all the Burials and Christnings, hapning that Week." They are described fully in the material selected from Graunt's book.

Charles II was so favorably impressed by the *Observations* that he specially proposed Graunt as an original member of the newly incorporated Royal Society. To forestall any possible objections on the ground that Graunt was a shopkeeper, "His Majesty gave this particular charge to his Society, that if they found any more such Tradesmen, they should be sure to admit them all, without any more ado."[4] He was elected F.R.S. in 1662.

The merit of the *Observations* was immediately recognized and encouraged the gathering and study of vital statistics on the Continent—particularly in France—as well as in England. The book went through several editions, the fifth of which, published after Graunt's death, was enlarged by Petty. Historians have long been vexed to decide how much Petty contributed to the original work. Aubrey, who delighted in retailing small malices, says only that Graunt had his "Hint" from Petty, but he implies much more. There seems no doubt that the book was a joint production. Graunt wrote by far the greater part, including the most valuable scientific portions; Petty, it may be supposed, added what Thomas Browne would have called "elegancy" and thereby increased the popularity of the book. Sir William

[2] "He not only gave a sound analysis of this problem (the calculation of annuity prices), but he put his results in such a convenient form that tills first table of mortality has remained the pattern for all subsequent tables, s to its fundamental form of expression."—Lowell J. Reed in the introduction to *Degrees of Mortality of Mankind* by Edmund Halley, a reprint of the papers noted, issued by the Johns Hopkins Press, Baltimore, 1942; p. iv. The selection by Halley is based on this reprint.

[3] *Aubrey's Brief Lives*, edited by Oliver Lawson Dick; London, 1950, p. 114.

[4] Tho. Sprat, *The History of the Royal Society of London, for the improving of Natural Knowledge*; 3rd Edition, London, 1722, p. 67.

was a bumptious and somewhat inflated man, unable to decide whether to patronize Graunt or to claim credit for his work. There is no evidence that he even understood the importance and originality of what his friend had done.[5] The last sentence—preface is unmistakably Graunt's: "For herein I have, like a silly Scholeboy, coming to say my Lesson to the World (that Peevish, and Tetchie Master) brought a bundle of Rods wherewith to be whipt, for every mistake I have committed."

Graunt served as a member of the city common council and in other offices, but on turning Catholic—he was raised a Puritan—"layd down trade and all other publique Employment." Aubrey tells us that he was a man generally beloved, "a faythfull friend," prudent and just. "He had an excellent working head, and was very facetious and fluent in his conversation." He was accused of having had "some hand" in the great fire of London, and the fact that he was a Catholic gave impetus to the charge. It was said that, as an officer of a water company, he had given orders stopping the water supply just before the fire started. A diligent eighteenth—century historian proved this false by showing that Graunt had had no connection with the company until a month after the fire. Graunt died of jaundice on Easter-eve 1674, and was buried "under the piewes" in St. Dunstan's church. "What pitty 'tis," wrote Aubrey, "so great an Ornament of the Citty should be buryed so obscurely!"

Unlike poor Graunt, whom my edition of the *Britannica* does not deign even to notice, Edmund Halley has been amply celebrated. I shall dispose of him as briefly as possible. He was born in London in 1658, the son of a wealthy "Soape-boyler," and he enjoyed every advantage, including an excellent education, that rich and indulgent parents could confer. His passion for mathematics and astronomy showed itself in his youth: when he arrived at Queen's College, Oxford, he brought with him a large assortment of astronomical Instruments, including a 24—foot telescope, whose use he had already mastered. His reputation as a theoretician and observer was established by the time be was twenty. He left the college before finishing his course, to make southern hemisphere observations at St. Helena. On his return, and by the King's command, he was awarded a Master of Arts degree; a few days later he was elected a Fellow of the Royal Society. He was then twenty-two. The next few years were spent on various astronomical labors which required him to travel widely on the Continent. Becoming deeply interested in the problem of gravity, he visited Newton at Cambridge in August 1684. It was a momentous meeting, for it resulted in the *Principia*, a work which might never have appeared except for Halley's extraordinary exertions. He suggested the project in the first place; he averted suppression of the third book; he bore all the expenses of printing and binding, corrected the proofs, and laid his own work entirely aside to see Newton's masterpiece through the press. The expense was assumed at a time when Halley could ill afford it. His father had suffered serious reverses before he died and had left an encumbered and almost worthless estate.

Halley's long life was crowded with literary and scientific activity. He was a classical scholar, hydrographer, mathematician, physicist, and astronomer. His writings include, besides a vast output in his specialty, such diverse items as "An Account

[5] For a meticulous sifting of the evidence as to Graunt vs. Petty see the introduction to a reprint of the *Observations* (Baltimore, The Johns Hopkins Press, 1939), by Walter F. Willcox. As to Petty, no inconsiderable person even if he was inflated and bumptious, see E. Strauss, *Sir William Petty, Portrait of a Genius*, Glencoe (III.), 1954.

of the Circulation of the Watery Vapours of the Sea, and of the Cause of Springs"; "Discourse tending to prove at what Time and Place Julius Caesar made his first Descent upon Britain"; "New and General Method of finding the Roots of Equations"; a translation from the Arabic - which language he learned for this purpose—of Apollonius' treatise *De sectione rationis* and a brilliant restoration of his two lost books *De sectione spatii*; an admirable edition of the same author's *Conics*; and more than eighty miscellaneous papers published by the Royal Society, which he served as secretary. In 1698 he commanded the war-sloop Paramour Pink in an expedition to the South Atlantic to study variations of the compass and to find new lands, if possible. On this journey he "fell in with great islands of ice, of so incredible a height and magnitude that I scarce dare write my thoughts of it." He was made Savilian professor of geometry at Oxford in 1703 and astronomer royal in 1721. One of his greatest achievements was a study of the orbits of comets, of which he described no less than twenty-four. Three of these were so much alike that he was convinced that the comets of 1531, 1607, and 1682 were one body. Assuming its period to be seventy-six years, he predicted its return in 1758. On Christmas Day of that year his conjecture was verified, and Halley's comet has since appeared in 1835 and 1910.

Halley died at the age of eighty-six. He was a generous, easygoing person, "free from rancor or jealousy," who spoke and acted with an "uncommon degree of sprightliness and vivacity." He enjoyed his work, had excellent health and owned a large circle of friends, among them Peter the Great of Russia to whose table he always had access. Bishop Berkeley thought Halley an "infidel," and it is true that in 1691 he was refused the Savilian professorship of astronomy at Oxford because of his alleged "materialistic views." The evidence is that he was a sensible man spoke his mind freely—dangerous practice in any age.

Halley's concern with the "curious tables" of Breslaw was one of his diversions. This Silesian city had, for more than a century before y into the problem, kept regular records of its births and deaths. Dr. Caspar Neumann, a scientist and clergyman of Breslaw had analyzed data, "disproving certain current superstitions with regard of the phases of the moon and the so-called 'climacteric' health." [6] His results were submitted to Leibniz who sent them al Society. It was at about this time that the Society resumed publication of the "Transactions" after a lapse of several years. Halley to furnish five sheets in twenty of the forthcoming issues. He was never hard up for ideas, nor for the energy and ingenuity to express them. His Breslaw papers may therefore be regarded as a kind of filler for the "Transactions", to keep his word until something better came along. Nevertheless, the analysis reflects the exceptional power of his mind.

15.4.2 The Paper

[6] Lowell J. Reed

Chapter 16

Hypothesis Testing: Tests and Pitfalls

16.1 Non-parametric Tests

So far, our tests required us to *assume* that our data was distributed in a particular way, usually normally. This is warranted in many cases by the central limit theorem but what happens if we cannot be sufficiently sure that our data has the required distribution? After all, statistics is already so error-prone, we do not want to make matters worse than they already are.

Definition 27 Non-parametric tests *allow statistical testing independent of the population distribution and related parameters.*

These tests are particularly useful if the population is highly skewed or if the data is non-numerical, for example if people are asked to rate something in order of preference.

16.1.1 The Sign Test for Equal Distributions

Suppose that there are two machines that make the same product. They are identical for all we know and they produce the same number of a certain product per day. They do, however, produce a few defect products per day and this number varies. In order to test if one machine is more reliable than the other, we produce a table of the number of defectives for each machine over a range of days. We then subtract the first machine's number from the second machine's total for each day and keep only the sign of this difference.

The test of the null hypothesis that the two machines are indeed the same, would have us expect to receive as many minus signs as plus signs. In fact, this is the same as testing whether a coin after a certain number of throws of heads and tails is fair, i.e. we need the binomial distribution.

The probability of getting x heads given N throws is

$$P_N(x) = \binom{N}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{N-x} = \binom{N}{x} \left(\frac{1}{2}\right)^N \tag{16.1}$$

As we want to test if the machines are the same, we have to use a two-tailed test (we would use a one-tailed one if we predicted the first machine to be better than the second). Remembering that N is given, we now add the probabilities until the total exceeds one-half of the chosen significance level.

Suppose we choose the 0.05 significance level and $N = 12$, then we get

$$P_{12}(0,1,2) = 0.01928 \tag{16.2}$$
$$P_{12}(0,1,2,3) = 0.07299 \tag{16.3}$$

and so at the 0.05 significance level we have to accept the hypothesis that the two machines are equally good if we have at least 3 minuses and 3 pluses in 12 days of testing (as the probability for 0,1,2 is less than 0.025 and the probability for 0,1,2,3 is larger).

We remark that if the difference between the two defective totals of the machines is zero one day, this day must be ignored as we can not ascribe a sign to this difference. Furthermore, you may use a normal approximation to the binomial distribution if you wish.

16.1.2 The Mann-Whitney U Test for Two Samples Belonging to the Same Population

We have taken two samples and we wish to test if they came from the same population, i.e. if there is a difference between the two samples. First of all, we list all test values of both samples in a single list sorted from least to greatest value and we assign a rank to each value. If there are duplicates we assign to each value the average rank of the ranks of each individual value. That is if ranks 12 and 13 both have a 2 in it, we assign the rank 12.5 to both of these 2's. Next we compute the sum of the ranks for each sample and denote these two sums by R_1 and R_2 with N_1 and N_2 being the sample sizes. For convenience, let's choose $N_1 \leq N_2$.

If there is a significant difference between rank sums, then there is a significant difference between the samples and so we want to test this. We use the statistic U,

$$U_1 = N_1 N_2 + \frac{N_1 (N_1 + 1)}{2} - R_1; \qquad U_2 = N_1 N_2 + \frac{N_2 (N_2 + 1)}{2} - R_2 \tag{16.4}$$

corresponding to samples 1 and 2 respectively. Note that

$$U_1 + U_2 = N_1 N_2; \qquad R_1 + R_2 = \frac{N(N+1)}{2} \tag{16.5}$$

where $N = N_1 + N_2$.

The distribution of U is symmetric and has mean and variance given by

$$\mu_U = \frac{N_1 N_2}{2}; \qquad \sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12}. \tag{16.6}$$

If both samples have at least 8 individuals, then the distribution is almost normal. Thus the z-score is normally distributed with mean 0 and variance 1,

$$z = \frac{U - \mu_U}{\sigma_U} \tag{16.7}$$

and so the null hypothesis that the two sample are from the same population is acceptable at the 0.05 significance level if $-1.96 < z < 1.96$.

16.1.3 The Kruskal-Wallis H Test for k Samples Belonging to the Same Population

This is the generalization of the Mann-Whitney test for k instead of 2 samples. The k samples have sizes N_1, N_2, $\cdots$, N_k with $N = N_1 + N_2 + \cdots + N_k$ and we assume that all values have been ranked together and the rank sums computed for each of the samples and denoted by R_1, R_2, $\cdots$, R_k. Then we define

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{N_j} - 3(N+1). \tag{16.8}$$

It is possible to show that H is nearly a chi-square distribution with $k-1$ degrees of freedom provided that each sample contains at least five individuals.

If there are any ties, they must be ranked as described above and the statistic H must be corrected by a factor. Let T represent the number of ties for an observation. Then the value of H must be divided by

$$1 - \frac{\sum (T^3 - T)}{N^3 - N} \tag{16.9}$$

If there are no ties $T = 0$ and the factor is equal to 1. In practise, this correction usually does not change the decision but there are circumstances in which it does, so this possibility should not be ignored.

16.1.4 The Runs Test for Randomness

Quite frequently in statistics, an opponent can allege that a found correlation or a test-verified hypothesis is spurious (correctly calculated but based on chance and not a causality of events). One possible source of spurious conclusions is if the data collected is actually random, that is it does not show an observable regularity. In this case, all correlations or conclusions (apart from the one that the data is random) are spurious.

Of course, it is often desirable for data to be random and so this does not need to be a negative aspect. The runs test is a simple test to see if a given set of data is random or not.

Definition 28 *In a linearly arranged sequence of symbols drawn from two possible symbols, a* run *is the longest local subsequence of identical letters.*

If the data is

$$aaaabbbbaaaabbbb \tag{16.10}$$

then we have four runs with the first and third run being “aaaa” and the second and fourth run being “bbbb.” In this case, we easily observe the pattern and thus the test for randomness should fail. We will essentially count runs and see if there are enough but not too many. For example the series

$$ababababababab \tag{16.11}$$

has runs of length one but is also not random. Thus a random sequence can be said to have a moderate number of runs.

Suppose that there are N_1 a's, N_2 b's and V runs in the sequence. It can be shown that this sampling has a distribution of mean and variance

$$\mu_V = \frac{2N_1N_2}{N_1+N_2} + 1; \qquad \sigma_V^2 = \frac{2N_1N_2\left(2N_1N_2 - N_1 - N_2\right)}{\left(N_1+N_2\right)^3\left(N_1+N_2-1\right)} \tag{16.12}$$

If both N_1 and N_2 are at least 8, then the sampling distribution of V is almost a normal distribution and thus we can use the tried and true method of z-scores again,

$$z = \frac{V - \mu_V}{\sigma_V} \tag{16.13}$$

If the data is numerical, arrange it in the order it was taken and replace each number with a if it is lower than the median, b if it is above the median and ignore it if it is equal to the median of the sequence. Then apply the above test to see if the numerical data is random or not.

16.2 Pitfalls of Testing

Be aware of the following important point when you test any hypothesis using these or other tests and try to address these points if and when you have to test something.

1. There is always a chance that the test gives rise to the wrong decision.
2. Never put complete trust in confirmatory statistics, i.e. if your long-held pet theory turned into a null hypothesis is confirmed by a test, do not believe that it is true now. It has merely been confirmed by the data collected. If, on the other hand, the test is negative, you may assume that it is wrong or at least does not hold with the assumed generality.
3. Be certain that the measured sample is representative of the population.
4. Think of all biases and counteract them as much as possible.
5. Be certain that the data source is reliable. A representative sample of liars is also useless.
6. Think of all parts of the experiment where you might be leading the test person or introducing some incentive to go one way, to keep back information, to embellish information or even to lie and counteract them as much as possible.
7. Choose the test that best matches your null hypothesis.
8. Investigate carefully the truth of the assumptions of the test (if any). You may have to use other tests to ascertain this.
9. Every test gives an answer at some significance level or with some probability of error.

10. In any write-up, carefully identify the manner in which the data was obtained and the testing done. If the data was not collected by you, be sure to investigate how it was collected so that you can be sure of reliability and representativity of the data.

16.3 Reading: First Life Insurance Tables by Edmund Halley

16.3.1 Introduction to the Reading

16.3.2 The Paper

Chapter 17

A Matrix Review

17.1 Matrices and Relatives

This is not a course on matrices. The basics of matrices are expected to be an old hat. Hence a brief review follows. Some topics will be new but matrices are not exactly rocket science and so this should be a sweet ride. Nonetheless, buckle up!

17.1.1 Basic Concepts

A *matrix* A is a number of elements A_{ij} arranged in n rows and m columns and is labelled by the row and column in which it appears. A_{ij} can be any mathematical object but is usually a number. Square matrices ($n = m$) have an associated determinant $|A|$. Special case of matrices are row vectors, column vectors and diagonal matrices, these look like

$$\text{row vector} = \begin{pmatrix} v_1 & v_2 & \cdots & v_m \end{pmatrix} \tag{17.1}$$

$$\text{column vector} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \tag{17.2}$$

$$\text{diagonal matrix} = \begin{pmatrix} v_{11} & 0 & 0 & \cdots & 0 \\ 0 & v_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & v_{nm} \end{pmatrix} \tag{17.3}$$

There is a special matrix called the identity matrix

$$I = \delta_{ij} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{17.4}$$

and the elements are denoted by δ_{ij} which is the Kroenecker delta symbol. So we may write any diagonal matrix as

$$D = D_{ii}\delta_{ij} \tag{17.5}$$

17.1.2 Matrix Arithmetic

We define the usual concepts as always.

Definition 29 *Two matrices A and B are* equal *if and only if $A_{ij} = B_{ij}$ for all i and j.*

Definition 30 *The* zero *and* one *matrices are $Z = Z_{ij} = 0$ and $I = \delta_{ij}$ for all i and j.*

Definition 31 *Two matrices A and B are added to produce their* sum $A + B$ *and their* difference $A - B$ *by*

$$\begin{aligned} C &= A + B = A_{ij} + B_{ij} \quad (17.6) \\ D &= A - B = A_{ij} + B_{ij} \quad (17.7) \end{aligned}$$

We note that it is necessary for two matrices to have the same size for the sum and difference to be defined.

Definition 32 *A matrix A is multiplied by a number a by $aA = aA_{ij}$ for all i and j.*

Definition 33 *Two matrices A and B are multiplied to produce their* product $C = A \cdot B$ *only if the number of columns of A equals the number of rows of B. If A is $n \times k$ and B is $k \times m$, then C is $n \times m$ and*

$$C_{ij} = \sum_{l=1}^{k} A_{il} B_{lj} \quad (17.8)$$

We note that in general $A \cdot B \neq B \cdot A$ and usually one of the two is even undefined. We also note that multiplication with δ_{ij} leaves the arguments unchanged. This is also commutative.

Definition 34 *If $A = A_{ij}$ we define the transpose A^T by $A^T_{ij} = A_{ji}$, that is we switch the identity of rows and columns.*

Note that $\left(A^T\right)^T = A$ and we define some special cases.

Definition 35 *A* symmetric *matrix has $A^T = A$, an* anti-symmetric *matrix has $A^T = -A$ and an* orthogonal *matrix has $A^T \cdot A = A \cdot A^T = I$.*

Furthermore if $C = A \cdot B$ then $C^T = B^T A^T$.

Definition 36 *The* trace *of a matrix is the sum of the diagonal elements $T(A) = \sum\limits_{i=1}^{k} A_{ii}$.*

17.1.3 Determinants

Every square $n \times n$ matrix A has an associated determinant $\det A$,

$$\begin{aligned} D(A) &= \begin{vmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nm} \end{vmatrix} && (17.9) \\ &= \sum \alpha = 1^n \sum \beta = 1^n \cdots \sum \nu = 1^n A_{1\alpha} A_{2\beta} \cdots A_{n\nu} \epsilon_{\alpha\beta\cdots\nu} && (17.10) \end{aligned}$$

where

$$\epsilon_{\alpha\beta\cdots\nu} = \begin{cases} 1 & \text{even permutation} \\ -1 & \text{odd permutation} \\ 0 & \text{any two indices equal} \end{cases} \tag{17.11}$$

and we note that 123, 231 and 312 are examples of even permutation whereas 132, 213 and 321 are examples of odd permutations.

Definition 37 *The i, j* minor determinant *of order $n-1$ is formed from the $n \times n$ determinant D by removing row i and column j from D; it is denoted by $M_{ij}^{(n-1)}$.*

Definition 38 *The i, j* co-factor *is $\alpha_{ij}^{(n)} = (-1)^{i+j} M_{ij}^{(n)}$*

Thus we have

$$D = \sum_{k=1}^{n} \alpha_{ij}^{(n-1)} A_{ik} \tag{17.12}$$

for any i.

The properties of the determinant include

1. Multiplying a row or column by a number a, multiplies the determinant by a also.

2. $D = 0$ when either (1) all elements in a single row or column are zero or (2) any row or column is equal to another row or column or (3) any row or column can be expressed as a linear combination or other rows or columns.

3. Interchanging two rows or columns changes the sign of the determinant.

4. The determinant remains unchanged when it is transposed or when any row or column is added to or subtracted from any other row or column.

17.1.4 Division

Definition 39 *The* adjoint *matrix A^{adj} is formed from a square $n \times n$ matrix A by*

$$A_{ij}^{adj} = \alpha_{ji}^{(n-1)} \tag{17.13}$$

Dividing is more complicated and not always possible.

Definition 40 *The inverse A^{-1} of a matrix A is defined by $AA^{-1} = A^{-1}A = I$ and is equal to*

$$A^{-1} = \frac{A^{adj}}{\det A} \tag{17.14}$$

where the inverse is understood not to exist if $\det A = 0$*. In this case A is called a* singular matrix*. Thus division is not always possible.*

We can now work out that if $C = A \cdot B$, then $C^{-1} = B^{-1} \cdot A^{-1}$.

17.1.5 Null Space

The *null space* of an m by n matrix A is the set of all those vectors in R^n that A maps to the zero vector in R^m, i.e.

$$Null(A) = \{X \in R^n : A \cdot X = 0\} \tag{17.15}$$

A basis for a vector space is a set of linearly independent vectors $\{x_1, x_2, \cdots, x_n\}$ such that any vector in the space can be written as a linear combination of the basis vectors, i.e.

$$X = a_1x_1 + a_2x_2 + \cdots a_nx_n \tag{17.16}$$

It is the finding of a basis for the null space that we shall need. It turns out to be the columns of the matrix known as the reduced echelon form of the matrix A.

17.1.6 Reduced Echelon Form for a Matrix

Definition 41 *An $m \times n$ matrix is in* echelon form *if*

1. *All rows that consist entirely of zeros are grouped together at the bottom of the matrix;*
2. *In every nonzero row, the first nonzero entry (moving left to right) is* 1*;*
3. *If the $(i+1)^{st}$ row contains nonzero entries, then the first nonzero entry is in a column to the right of the first nonzero entry in the i^{th} row.*

The following matrix is in echelon form:

$$\begin{pmatrix} 0 & 1 & -9 & 7 & 0 \\ 0 & 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{17.17}$$

This matrix isn't in echelon form (why?):

$$\begin{pmatrix} 0 & 1 & -9 & 7 & 0 \\ 0 & 0 & 1 & 1 & 3 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{17.18}$$

Definition 42 *A matrix in echelon form is in* ***reduced echelon form*** *if the first nonzero entry in any row (a* 1*) is the only nonzero entry in its column.*

The following matrix is in reduced echelon form:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 9 \\ 0 & 0 & 1 & 0 & 5 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{17.19}$$

17.1.7 Simultaneous Linear Equations

Suppose we have to solve the equations

$$a_1 x_1 + a_2 x_2 = a_3 \tag{17.20}$$

$$b_1 x_1 + b_2 x_2 = b_3 \tag{17.21}$$

for the variables x_1 and x_2 given all the others. We reformulate in terms of the above language

$$\begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_3 \\ b_3 \end{pmatrix} \tag{17.22}$$

and represent the coefficients by matrices such that

$$A = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}; \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}; \qquad B = \begin{pmatrix} a_3 \\ b_3 \end{pmatrix} \tag{17.23}$$

and so we have

$$A \cdot x = B \tag{17.24}$$

$$A^{-1} \cdot A \cdot x = A^{-1} \cdot B \tag{17.25}$$

$$x = A^{-1} \cdot B \tag{17.26}$$

if and only if $\det A \neq 0$. If $\det A = 0$, then these equations have no solution. This method of course applies no matter what is in the matrices A and B and in fact linear algebra is largely concerned with this type of equation.

We call the equations homogenous when $B = 0$. In this case, we always have $x = 0$ as a solution. This solution is the only one (a unique solution) if and only if $\det A \neq 0$ for the above reasons. If $\det A = 0$, then $x = 0$ is a solution but not the only one. In the inhomogenous case, we have a solution if and only if $\det A \neq 0$.

17.1.8 Eigenvalues and Eigenvectors

Definition 43 *If A is an $n \times n$ matrix and x is a column vector of length n and if $A \cdot x = \lambda x$ where λ is a constant number, then x is called an* eigenvector *of A corresponding to λ which is called the* eigenvalue *of A. In general, we have n different eigenvalues and eigenvectors for a matrix A but they may be duplicated in some cases.*

We have

$$A \cdot x = \lambda I \cdot x \tag{17.27}$$

and so

$$(A - \lambda I) \cdot x = 0 \tag{17.28}$$

which is homogenous and to be solved for both λ and x. To have non-trivial (i.e. $x \neq 0$) solutions we must have $\det(A - \lambda I) = 0$. We determine λ from this,

$$\det(A - \lambda I) = \begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{17.29}$$

The previous equation will yield a polynomial in the variable λ which will have n roots $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ that may or may not be all different. These are the eigenvalues. Having obtained a particular λ_i, we may obtain its associated eigenvector x_i by

$$(A - \lambda_i I) \cdot x_i = 0 \tag{17.30}$$

and non-trivial solutions must exist as we have ensured that the matrix on the left has vanishing determinant.

17.2 Human Behavior and the Principle of Least Effort by George Kingsley Zipf

Nearly twenty-five years ago it occurred to me that we might gain considerable insight into the mainsprings of human behavior if we viewed it purely as a natural phenomenon like everything else in the universe, and if we studied it with the same dispassionate objectivity with which one is wont to study, say, the social behavior of bees, or the nestbuilding habits of birds, The present book reports the results of the extended inquiry that ensued in the course of those years, and which led to the disclosure of some fundamental principles that seem to govern important aspects of our behavior, both as individuals and as members of social groups.

17.2.1 The question of Practical Application

It is inescapable that the attitude of the natural scientist towards human behavior will differ from that of the man or woman of affairs who is faced with the practical need of solving urgent human problems, even though the two attitudes are by no means irreconcilable, Thus the scientist hopes that from an objective study of the actual ways that we human beings do in fact behave, he may disclose the nature of the underlying principles that govern our conduct, But though the scientist's interests are said to stop at that point, he must nevertheless admit that a knowledge of these underlying principles will inevitably help others to live *more efficiently*, whether as individuals, or as members of teams that cooperate and compete - and whether in the roles of those who primarily do the leading, or in the roles of those who primarily do the following.

After all, the basic natural principles that govern man's responses to the incentives of prospective rewards, or that govern the proportion of executive leaders to followers (in corporation, labor union, army, political Party, or social club), or that govern the rise and decline of fashions, or that govern the distribution of relative power and control in any dominance system, from the family unit to the nation, or that govern the circulation of persons, goods, services, ideas, and information over the earth's surface in the exchange of the products of our labor - just to enumerate a few of the principles that we shall treat - are not likely to remain idle curiosities, so long as each man must daily cooperate and compete with others in order to live. Nor are these principles particularly out of place at the present time, when we seem to be faced with an impending planned economy in Which a few persons will tell many others how they *should behave - often* Perhaps without regard to how people *do* behave.

On the other hand, just because we treat objectively and dispassionately of the mainsprings of human behavior, without any particular reference to how people "should" behave, that does not necessarily mean that I for one feel personally obliged to deprecate the almost universal belief that all mankind "should" co-operate and get along together like one happy team that is bent upon "social progress." Nevertheless I do note that, in spite of this universal belief, there is virtually no agreement as to the particular ways and means whereby the worthwhile objective of universal human cooperation is to be achieved.

It is obvious that some persons and groups have personal and group *preconceptions* as to precisely how the world "should" co-operate. These preconceptions are sometimes so deeply rooted that the individuals in question can barely talk with others whose similarly profound preconceptions happen not to agree with their own, In so doing they seem to block *communication*, and thereby to impede the better world understanding and co-operation they so fervently desire, It is further obvious that many of these persons and groups are so rigid and inflexible in their preconceptions that they are not to be budged from them either by incentive rewards of any amount or by threats of direst harm.

Neither the natural scientist nor the practical social engineer can afford to ignore the power of these preconceptions, to which even the best intended incentives are often subordinate and from which, only too often, the gravest individual and group misery can result.

Nevertheless, to the natural scientist, man's preconceptions do not belong to some other world, but instead are merely further natural phenomena, As such they are a part of the total natural phenomenon of human behavior and merit an investigation into their mainsprings quite as much as the rest of human behavior, Indeed, in many situations, the preconceptions involved are largely determinative of the rest of the behavior.

Our emphasis upon the effect of man's preconceptions is by no means new to present-day thinking, even though, in actual practice, a given person's attitude towards a particular *vocabulary* of individual or group preconceptions that confront him may depend upon what his particular problem is.

Thus, for example, the personnel man in a sizable factory, store, labor union, or armed group, who is obliged to deal daily with a large number of persons of diverse cultural and socio-economic backgrounds, has the continual task of understanding and of reconciling his group's conflicting preconceptions, so that the group can effectively *communicate* in reference to its common tasks, incentives, and risks, The personnel man does not need to be told that it is efficient for his group to have a common *language*, in the broad sense of having like responses and evaluations to like stimuli within the functioning of the group, His task is essentially that of understanding the existing diverse preconceptions, so that he can restructure them into a more harmonious whole.

The case is somewhat different, however, with the politician who wants votes, or with the marketer who invents styles or has something to sell, For here the game may be simply the most effective manipulation of the existing preconceptions, without any thought of altering them, A politician, though meritorious, who casually ignores his constituents' preconceptions, or else tries to superimpose his own logic upon them, is only too likely to fall before another and even far less capable politician who knows the preconceptions of his constituents, and who says, in substance, "My friends, I understand your feelings perfectly, and am heartily in accord with them."

Yet just because one man's preconceptions often flaunt another man's logic in what seems to him to be a highly capricious manner, we may by no means suppose that man's preconceptions are random and haphazard, and without a certain logic of their own, On the contrary, in our study of the dynamics of language and the structure of the personality, we shall find that a vocabulary of preconceptions is quite orderly and is governed by quite definite underlying principles. Nor are we in any way disposed to argue that the deliberate use of these underlying principles by the personnel man, politician, and marketer will not help him to alter or to manipulate more effectively the particular vocabulary of preconceptions with which he happens to be confronted.

It is perhaps well at this point to elucidate our more general terms *language* and a *vocabulary of preconceptions*, lest they be confused with the more familiar terms, *words* and *speech*, with which, incidentally, they are intimately related.

To this end, since we have just spoken of the marketer's problem, let us by way of illustration begin with a "brand name" of goods (for example, G.E., Frigidaire, Chesterfield), A given brand name may be so favorably known that many persons will prefer, and even pay more for, the brand goods than for unnamed goods, although even a connoisseur could not tell the difference. In short, a *specific brand name tends to evoke a specific response in reference to human wants*, and in so doing may be said to represent a sort of preconception.

Let us note, however, that a brand name is also a *word*, and nothing but a *word*, Whatever the principles may be that govern the behavior of words in their favorable and unfavorable connotations, and in their fashionableness and obsolescence, will also govern to a considerable extent the ups and downs and competition of brand names. (Hence our later study of *words* and *speech* is germane to a study of fashions and preconceptions.)

But let us go further, Instead of a brand name, let us consider a trademark which identifies a particular product or service quite as effectively as a brand name, but which contains not a single word, This trademark is a sign or a *signal* which, like a brand name, evokes a more or less stereotyped response. Although a trademark is not a word, and therefore not an element of speech, we shall later see that it is in fact an element of what we may call the group's language, (And we shall attempt to demonstrate that things like trademarks will behave in a manner quite similar to that of words.)

But we do not need to stop with the trademark, There are many stereotyped things, such as kinds and qualities of clothes, ways of doing one's hair, manners of gesticulating and of talking, places and times where one is seen or not seen, which convey information about the person in question, Although these things are neither words, brand names, nor trademarks, they tend to evoke more or less stereotyped responses and, in so doing, they belong to the *language* of the group in question - quite as much as the group's words, and phrases, and sentences.

To illustrate the import of this broader concept of language, let us paint a picture. It is evening; a costly automobile with livened chauffeur drives up before the opera house; and out steps milady elegantly gowned and jeweled. She fumbles in her purse and conspicuously gives the beggar a coin, and then skips up the stairs. That is the picture.

All the parts of this picture relate to the problem of the production and distribution of goods and services and of rewards, Yet, as we shall note in detail, all parts of the picture - the car, chauffeur, opera, woman, gown, jewels, coin, and beggar - are *also* problems in *language* (and in *preconceptions*).

And so if at times in the opening chapters of our demonstration, we seem to be almost pedantically concerned with the phonetic and semantic minutiae of human speech which are apparently of little importance in the robustness of everyday life, may the reader of imagination reflect that we may thereby be gaining insight into the total *language* of the group, whose minutiae may at times be extremely important in everyday life, in politics, in marketing, or in just plain getting along together.

In thus placing a study of the principles of language before that of the economy of geography, or of the distribution of economic power and social status, or of the waxing and waning of prestige symbols and cultural vogues, we perhaps confess to a certain personal preconception as to what is likely to be most important in the difficulties of actual everyday practical human problems, from which confusion, heartache, and misery arise.

17.2.2 The Question of Natural Science

Although we have not evaded the question of the possible practical value of scientific principles in the solution of actual problems in human behavior, we nevertheless must point out that the present study is offered purely as a work of science.

More specifically, it is the expressed purpose of this book to establish The Principle of Least Effort as the primary principle that governs our entire individual and collective behavior of all sorts, including the behavior of our language and preconceptions.

An investigator who undertakes to propound any such primary scientific principle of human behavior must discharge three major obligations towards his reader. First, his argument must be supported by a large number and variety of verifiable observations of relevant phenomena. Second, his theory must be logically self-consistent throughout, with all terms and operations relating to his observations explicitly defined, Third, his entire demonstration should be presented in such a way that it will be readily understandable to the reader who, in the present case, is assumed to have a genuine interest in the underlying principles of human behavior, without necessarily having any specialized or technical training in the fields in question.

As to the first point - the number and variety of observations - we may claim in all modesty to have increased the number of our observations to such a point that they may be viewed as empiric natural laws, regardless of the correctness of any of our theoretical interpretations. In other words, by means of the accepted methods of the exact sciences, we have established an orderliness, or natural law, that governs human behavior, Moreover, the variety of our observations, which extend from the minutiae of phonetic and semantic behavior to the gross distributions of human populations, goods, services, and wealth, is sufficient, I believe, to give pause to the superficial opinion that the observed orderliness on these fundamental matters has nothing to do with the practical affairs of everyday life, We stress this fact in the hope that any person who may sincerely wish to apply these findings to the solution of his own problems may do so with a feeling of confidence, even though some of the findings may not be entirely in line with current preconceptions about how people "should" behave.

As to the second point - the theoretical aspect of the study - that is, the theoretical demonstration of the Principle of Least Effort - we submit that our theory, *like all other theories in natural science*, does *not* claim either that no other theory can be found that will also rationalize our data, or that no other data will ever be found that do not controvert our theory. On the contrary, the reader is invited not only to weigh our own theory, but to find a more cogent theory and more instructive data, To this end, we have tendered suggestions for further elaborational research, and have tried to open further theoretical vistas for the possible use of others, whether these others be professional students who are interested in original research of their own, or nonprofessional laymen who simply like to adventure with new ideas.

As to the third point - the manner of presentation of the material - we have prepared the manuscript in such a way that it will be entirely understandable to anyone interested in the topic, regardless of his previous training. In short, every step in observation, analysis, and statistical description has been explained simply and in detail for the reader's complete comprehension if he has only a modicum of patience, Nor is this simplified presentation entirely amiss for the specialized reader, not every one of Whom may be supposed to be familiar with all the fields upon which the present study touches (e.g., economics, sociology, cultural anthropology, Psychology-both general and Freudian-linguistics, and semantics), In this connection it might be remarked that we have restricted our bibliographical references to those publications germane to the discussion at hand and which will serve to orient the reader further in the bibliography of any of the special fields.[1]

[1] Because the preparation of our manuscript was essentially complete at the time, we WCCC unable to include a discussion of the recently appeared "Kinsey Report" in connection With our own discussion of an individual's homosexual-heterosexual balance, in which we have arrived at conclusions-on the basis of entirely different kinds of data-that undeniably support the 'Kinsey" findings. Hence reference is here made: A. C. Kinsey. W. B. Pomeroy, and C. E. Martin. *Sexual Behavior in the Human Male*, Philadelphia: Saunders, l948.

17.2.3 Introduction and Orientation

Everyone in the course of his daily living must to some extent move about in his environment. And in so moving he may be said to take paths. Yet these paths that he takes in his environment do not constitute his entire activity. For even when a person is comparatively at rest, there is still a continual movement of matter-energy into his system, through his system, and out of his system if only in the accomplishment of his metabolistic processes. This movement of matter-energy also proceeds over paths. Indeed, a person's entire body may be viewed as an aggregate of matter that is in transit at differing speeds over different paths within his system. His system in turn moves about as a unit whole over paths in his external environment.

We stress this concept of movement over paths because we shall attempt to demonstrate in the course of our following chapters that every individual's movement, of whatever sort, will always be over paths and will always tend to be governed by one single primary principle which, for the want of a better term, we shall call the *Principle of Least EfJort.* Moreover, we shall attempt to demonstrate that the structure and organization of an individual's entire being will tend always to be such that his entire behavior will be governed by this Principle.

And yet what is this Principle? In simple terms, the Principle of Least Effort means, for example, that a person in solving his immediate problems will view these against the background of his probable future problems, *as estimated by himself.* Moreover he will strive to solve his problems in such a way as to minimize the *total work* that he must expend in solving *both* his immediate problems *and* his probable future problems. That in turn means that the person will strive to minimize the *probable average rate of his work-expenditure* (over time), And in so doing he will be minimizing his *effort*, by our definition of effort. Least effort, therefore, is a variant of least work.

In the interest of defining and of elucidating the Principle of Least Effort, and of orienting ourselves in the problem of its demonstration, we can profitably devote this opening chapter to a preliminary disclosure of the Principle, if only on the basis of commonplace cases of human behavior that are admittedly oversimplified for the sake of a more convenient initial exposition.

17.2.4 The Selection of a Path

Sometimes it is not difficult to select a path to one's objective. Thus if there are two cities, A and B, that are connected by a straight level highway with a surface of little friction, then this highway represents simultaneously the *shortest*, the *quickest*, and the *easiest* path between the two cities - or, as we might say, the highway is at once a path of *least distance* and of *least time* and of *least work.* A traveler from one city to the other would take the same path regardless of whether he was *minimizing* distance, time, or work.

On the other hand, if the two cities happen to be separated by an intervening mountain range, then the respective paths of least distance, and of least time, and of least work will by no means necessarily be the same. Thus if a person wanted to go by foot from one city to another by least distance, he would be obliged to tunnel through the base of the mountain chain at a very great expense of work, His quickest course might be over the tops of the mountains at a great cost of labor and at great risk, His easiest path, however, might be a tortuous winding back and forth through the mountain range over a very considerable distance and during a quite long interval of time.

These three paths are obviously not the same. The foot-traveler between the two cities cannot, therefore, simultaneously minimize distance, time, and work in a single path between the two cities as the problem now stands. Which path, therefore, will he take? Or, since the above case is fairly typical of life's daily problems, in which impediments of various sorts obstruct our way, which path do we actually take? Clearly our selection of a path will be determined by the particular *dynamic minimum* in operation.

17.2.5 The Singleness of the Superlative

The preceding discussion of the selection of paths not only illustrates the meaning of a *minimum* in a problem in dynamics but also prepares the ground for a consideration of the concept of the *"singleness of the superlative"* which, incidentally, will provide an intellectual tool of considerable value for our entire inquiry.

The concept of the "singleness of the superlative" is simple: no problem in dynamics can be properly formulated in terms of more than one superlative, whether the superlative in question is stated as a *minimum* or as a *maximum* (e.g., a *minimum* expenditure of work can also be stated as a *maximum* economy of work). If the problem has more than one superlative, the problem itself becomes completely meaningless and indeterminate.

We do not mean that a particular situation will never arise in which the minimizing of one factor will not *incidentally* entail the minimizing of another or other factors. Indeed, in our preceding section we noted a situation in which the easiest path between two cities might be a straight level highway that also represented the shortest and quickest path. Instead we mean that a general statement in dynamics cannot contain more than one superlative if it is to be sensible and determinate, since a situation may arise in which the plural superlatives are in conflict.

Perhaps the simplest way to emphasize the singleness of the superlative is to present as an example a statement with a single superlative that is meaningful and determinate. Then we shall note how meaningless and indeterminate the statement immediately becomes when a second superlative is added.

As a suitable example we might take the imaginary case of a prize offered to the submarine commander who sinks the *greatest number* of ships in a given interval of time; in this case, *maximum number* is the single superlative in the problem. Or we might alter the terms of the problem and offer a prize to the submarine commander who sinks a given number of ships in the *shortest possible* time; in this second case, *time* is the *minimum*; and, since it is the only superlative in the statement, the problem is quite meaningful and determinate. In either of the above examples the submarine commander can understand what the precise terms of the prize are.

Yet when we offer a prize to the submarine commander who sinks the *greatest number* of ships in the *shortest possible time*, we have a double superlative-a *maximum* number and a *minimum* time-which renders the problem completely meaningless and indeterminate, as becomes apparent upon reflection.

Double superlatives of this sort, which are by no means uncommon in present-day statements, can lead to a mental confusion with disastrous results.[2]

In the present study we are contending that the entire behavior of an individual is at all times motivated by the urge to minimize effort.

The sheer idea that there may be only one dynamic minimum in the entire behavior of all living individuals need not by itself dismay us. The physicists are certainly not dismayed at the thought that all physical process throughout the entire time-space continuum is governed by the one single superlative, *least action.*[3] Indeed, the presence of only one single superlative for all physical process throughout the entire time-space continuum can even be derived logically from the basic postulate of science that there is a unity of nature and a continuity of natural law (in the sense that the same laws of nature govern all events in time-space). For, according to this postulate, the entirety of time-space, with all its happenings, may be viewed as constituting a single problem in dynamics which in turn can have only one single superlative-a superlative which in the opinion of physicists is that of *least action.*

By the same token, the sheer idea of there being one single superlative for all living process is not in and for itself an *a priori* incredibility.

On the other hand, there is also admittedly no *a priori* necessity for our believing that all living process does in fact behave at all times according to one single invariable superlative, such as that of least effort. That, after all, must first be established empirically, as was done with the principle of least action. We can even now note how bizarre the effect would be if a person behaved at one moment according to one dynamic minimum, and at the next moment according to an entirely different dynamic minimum. Nor would the effect be any less bizarre if one person's life were governed throughout by one superlative while his neighbor's life followed a totally different superlative.

In order to emphasize the ludicrousness of a variety of different superlatives, let us assume that each person consists of two parts, and that each part has a different dynamic superlative of its own, For example, let us assume that one part of the person is governed by least work while the other is governed by least time. In that case the person will represent two distinct problems in dynamics, with the result that he will be, effectively, two distinctly different individuals with two distinct sets of dynamical principles. One part of him, in its eagerness to save work, might conceivably even "get lost" from the other part of him, in its eagerness to save time.

Nor would the situation be different if we assume that a person now minimizes one factor and now another without any single governing principle behind the total phenomenon. For if the person's entire metabolistic and procreational system is organized, say, for the purpose of minimizing work in all its action, then there would have to be a simply staggering alteration of structure and of operation if the

[2] As pointed out years ago, the frequent statement, "in a democracy we believe in the *greatest* good for the *greatest* number" contains a double superlative and therefore is meaningless and indeterminate, (In Part Two we shall see that the distribution of goods and services are in fact governed by a single superlative.) Intimately connected with the "singleness of the superlative" is what might be called the *singleness of the objective* whose implications are often overlooked (i.e., the pursuit of one objective may preclude or frustrate the pursuit of the second objective). These two concepts apply to all studies in ecology.

[3] The principle of least action was first propounded by Maupertuis in the eighteenth century, and has been subsequently conceptually sharpened by others.

person in question were suddenly to minimize time. Since sudden alterations of such proportions are unknown, we are perhaps not overbold in suspecting *a fortiori* that an individual's entire activity from birth to death is governed throughout by the same one single superlative which, in our opinion, is least effort.

But that is not all. If we remember the extent to which offspring inherit the forms and functions of their parents, we may suspect that this inheritance is possible only if the offspring also inherit the parental dynamic drive that governs the parental forms and functions that are inherited.

Furthermore, if we view the present-day variety of living process as the result of slow evolutionary changes from an initial similarity of form and function, then we can understand *a fortiori* how the one initial single common dynamic superlative might well remain unchanged from generation to generation, regardless of how enormous the changes in forms and functions might become; and that, in turn, will mean that all individuals, regardless of their differences in form and function, will still be governed by the same single superlative.

But though we may argue at length as to the plausibility of one single superlative for all living process, yet, even if this were the case, we should still need to disclose what, in fact, the particular superlative in question is.

An actual disclosure of the single hypothetical superlative in question may be difficult for quite obvious reasons, If we take our previous example of the two cities with an intervening mountain chain, in which the paths of least distance, least time, and least work are three different paths, we are obliged in all candor to admit that sometimes *one* of these paths is taken and sometimes another. For that matter, a tunnel may be dug through the base of the mountain to save distance, while airplanes are flown over the same mountain to save time, while pack horses continue to take the easier and more leisurely winding route. Or, to take another case, sometimes the reader will dart through traffic at considerable risk in order to save time in crossing a Street; and sometimes he will take the longer and safer path to the corner, where he will wait for the traffic light. Even if we assume that we are all governed by the same one single dynamic superlative, which superlative is it?

But although the superlatives in the foregoing examples seem to be different, are they nevertheless irreconcilable? Before answering this question, let us remember the physicists' claim that according to their law of falling bodies, all free-standing bodies will fall (by least action) to the earth. Yet, despite this claim, we have all observed how leaves sometimes rise in the air, or how birds take off from the ground and fly out of sight, much as if they were exceptions to the law of falling bodies. Of course we know from a more careful inspection of the problem that these leaves and birds are by no means exceptions to the law of falling bodies; on the contrary, if all the factors in the problem are taken into consideration, they are behaving quite in accordance to the physical law in question.

May not the same be true of the three different paths to the other side of the mountain? Even though each of these paths may be taken simultaneously by someone, and even though a given person may now take one path and now another, there remains the possibility that the adoption of one or another by an individual under varying sets of circumstances is governed by the operation of some further single dynamic minimum that forever remains invariant, In any event, we shall argue that such is the case.

More specifically, we shall argue that if we view the above types of situations against the broad background of the individual's present and future problems, we shall find that an extraordinary expenditure of work at one moment, or an extraordinary haste in one situation, may simply be temporary devices for reducing the probable rate of the individual's work expenditure over subsequent periods of his life.

In short, we shall argue that the invariable minimum that governs all varying conduct of an individual is least effort,

17.2.6 The Principle of Least Effort

Perhaps the easiest way to comprehend the meaning and implications of the Principle of Least Effort is to show the inadequacies of sheer *Least work*, to which *least effort* is closely related, This is all the more worth doing because some persons (see below) apparently believe that least work is the basic minimum of living process, as often seems to be the case in particular situations that are considered out of context.

If we remember, however, that an individual's life continues over a longer or shorter length of time, then we can readily understand how the least work solution to one of his problems may lead to results that will inevitably increase the amount of work that he must expend in solving his subsequent problems. In other words, the minimizing of work in solving today's problems may lead to results that will increase tomorrow's work beyond what would have been necessary if today's work had not been completely minimized. Conversely, by expending more work than necessary today, one may thereby save still more work tomorrow.

And, as we have argued about the functional relatedness of today and tomorrow, we may argue about the functional relatedness of the entire succession of events throughout the individual's whole life, in which the rate of his expenditure of work at one moment may affect the minimizing of his work at subsequent moment(s).

In view of the implications of the above quite obvious considerations, we feel justified in taking the stand that it is the person's *average rate of work-expenditure over time* that is minimized in his behavior, and not just his work-expenditure at any moment or in any one isolated problem, without any reference to his future problems.

Yet a sheer *average rate of work-expenditure over time* is not an entirely meaningful concept, since no mortal can know for certain what his future problems are going to be. The most that any individual can do is to estimate what his future problems are *likely to be*, and then to govern his conduct accordingly. In other words, before an individual can minimize his average rate of work-expenditure over time, he must first estimate the probable eventualities of his future, and then select a path of least average rate of work through these.

Yet in so doing the individual is no longer minimizing an average rate of work, but *a probable average rate of work*; or he is governed by the principle of the *least average rate of probable work*.[4]

For convenience, we shall use the term *least effort* to describe the preceding least average rate of probable work. We shall argue that an individual's entire behavior is

[4] To avoid a possible verbal confusion, let us note that we are not discussing *least probable* average rate of work, but a *probably least* average rate of work.

subject to the minimizing of effort. Or, differently stated, every individual's entire behavior is governed by the Principle of Least Effort.

Now that we have described what the Principle of Least Effort is, let us briefly illustrate it.

At the risk of being tedious, let the first example be our previous case of the two towns, A and B, that are separated by an intervening mountain range. Here we can see the enormous amount of work that could be saved in travel and trade if the two towns were connected by a tunnel of least distance through the base of the mountain; we can also see the enormous amount of work that it would take to construct such a tunnel. We are simply arguing that when the probable cost in work of digging the tunnel is estimated to be less than the probable work of not having the tunnel, then, if the necessary work for construction is available, the tunnel will be dug. The problem relates, therefore, to the probable amounts of work involved, as estimated by one or more persons. Naturally, these persons can have been mistaken in their estimates, with the result that the tunnel can either succeed beyond their wildest hopes, or dismally fail. For we do not deny that "a person's hindsight is generally better than his foresight." We merely claim that a person acts on the basis of his "foresight" - with all that that will later be found to imply - according to the Principle of Least Effort.

The above type of argument will also apply to a path of least time over the mountain. Thus the enormous cost of flying munitions over the mountain to save time in supplying an army in combat on the other side may be more than justified by the future probable work that is thereby saved.

These cases of the different paths to the other side of the mountain represent instances of collective action and of collective economies, since, for example, a tunnel through a mountain is obviously not constructed by a single person but by the collective effort of a great many persons.

And yet we are not restricted to examples of collective effort in illustrating our Principle of Least Effort, which we contend also applies to an individual's own behavior. We might take the case of a student whose particular path of least effort out of his classroom would seem offhand to be the path that leads from his seat to the nearest aisle, and thence out of the door, through the hall, to the nearest stairway. On the other hand, in the event of a fire, the student might conceivably prefer to run with least time to the nearest window and adopt a path that is simultaneously a path of least work and of least time and of least distance to the ground. This path will also be a path of least effort, as estimated by himself, even at the risk of months in the hospital with a broken back. Other students may prefer to take paths through the smoke-filled corridors. These paths are also paths of least effort, as estimated by the students in question. Afterwards, when, as, and if all the students foregather, they can decide which of them, in the light of subsequent events, actually were the shrewdest gamblers in the sense of having both most correctly comprehended the nature and estimated the probabilities of the problem in their lives that was caused by the unexpected fire.

From this second example we note that the operation of the Principle of Least Effort is contingent upon the *mentation* of the individual, which in turn includes the operations of *"comprehending"* the "relevant" elements of a problem, of *"assessing their probabilities;"* and of "solving the problem in terms of least effort," We mention this vital consideration of *mentation* right here and now, so that we may prepare ourselves for the task of defining mentation, and of showing that the structure and operation of mentation are also governed throughout by the Principle of Least Effort, since an individual's mentation is clearly a part of his total behavior, and hence subject to our Principle of Least Effort.

The foregoing examples suffice to illustrate what the Principle of Least Effort is, and what its implications may be for everyday problems. By and large, our explanations of the above commonplace examples are pretty much in line with the way the reader himself would have explained them. We mention this consideration in order to suggest that our chief task may not be that of persuading the reader to adopt a totally new way of thinking, but rather of formally describing and of scientifically establishing the basic principle of our habitual way of thinking.

17.2.7 The Scope of the Principle: Tools and Jobs

Our previous examples have illustrated the theoretical operation of the Principle of Least Effort in particular situations that involved either individual or collective behavior. They did not illustrate, however, our contention that the Principle governs the *totality* of a person's behavior at all times. Since we shall find it necessary to devote considerable space to a demonstration of the economy of mentation, which is only a part of a person's total behavior, we obviously cannot hope in the course of a few paragraphs to illustrate the economy of the *totality* of a person's behavior by means of one single telling example.

Nevertheless it may be useful for the sake of preliminary orientation to suggest how a great deal of a person's total behavior can be expressed in the form of a simple problem of tools-and-jobs whose elements are quite familiar in everyday life, The problem of *tools-and jobs* is the same as the problem of *means* and *ends*, or of *instruments* (or agents) and *objectives.* We shall adopt the homelier term, *tools-and-jobs*, to emphasize the commonplace nature of the problem under discussion.

Regardless of the terms employed, it is evident upon reflection that according to the Principle of Least Effort there are two aspects to the economy of the *tools-and jobs* in question. In the first place, there is the economy of *tools.* In the second place, there is the economy of *jobs.*

To clarify the significance of these two economies, let us briefly illustrate them in terms of carpentry *tools* and carpentry *jobs.*

We all know from experience that when a person has a carpentry *job* to be performed, he directly or indirectly seeks a set of carpentry *tools* to perform the job. And, in general, we may say that *jobs seek tools.*

But what is often overlooked is the equally obvious fact that when a person owns a set of carpentry *tools*, then, roughly speaking, he directly or indirectly seeks a carpentry *job* for his tools to perform. Thus we may say that *tools seek jobs.*

This reciprocal matching of tools to jobs and of jobs to tools may conveniently be described by the phrase, *tools-seek-jobs-and-jobs-seek-tools.*

Upon further reflection, we note that the concept of this italicized phrase is ultimately unsatisfactory because it defines tools in reference to jobs and jobs in reference to tools. Hence, unless either the tools or the jobs are fixed, the companion term remains undefined. In subsequent chapters, we shall find a third frame of reference that will serve to define both *tools* and *jobs* in conditions where neither is fixed.

For the purpose of a preliminary orientation, however, we can illustrate superficially some of the more obvious implications of the reciprocal economy of matching tools and jobs under the assumption that either the tools or the jobs are fixed, We shall begin with the example of an automobile manufacturer, and then turn to the case of an imaginary person called John.

If the owner of a manufacturing plant has the *job* of making pleasure automobiles, then, theoretically, he will seek to use those tools that will manufacture the automobiles with a maximum economy, as is observably the case with all automobile manufacturers. The same urge of economy presumably activates the manufacturers of other kinds of goods. In short, the kinds of jobs (or objectives) that a person has to perform will determine the kinds of tools (or means) that he employs for their performance.

Nevertheless, the above situation might well be immediately changed with the sudden outbreak of war that introduces a whole new set of national jobs-or objectives-while suppressing a great many of the erstwhile peacetime ones. During the war the automobile manufacturer may no longer be permitted to manufacture his peacetime pleasure cars; and the same will doubtless be true of many other kinds of manufactures, That will not mean, however, that the manufacturing plants in question will remain idle for the duration of the war, On the contrary the plants will be "converted to war work." That is, they will perform the new kinds of war jobs.

What, more precisely, takes place under this "conversion to war work"? Theoretically, each plant will seek to perform that particular kind of war job most nearly adapted to its present peacetime toolage; that is, it will seek to perform the particular job for which it can re-tool with least work (effort) - Thus the automobile factory may make tanks, or jeeps, or gun carriages of some sort. Generalizing upon this case, we may say that tools-seek-jobs-and-jobs-seek-tools throughout the entire nation at war.[5]

After the war is over, the manufacturers again face the need of "conversion." This does not mean that the manufacturers will necessarily revert to production of their prewar lines of goods, although they may. They may find it easier to convert their wartime toolage to the production of some entirely different kind of peacetime goods. In this conversion to peacetime activity, we may again say that *tools-seek jobs-and-jobs-seek-tools.*

The foregoing example of the manufacturing plant is instructive for two reasons, *First*, it shows what might be called the complete *relativism* of the problem, with little that is permanently stable over any extended period of time; for by just such successive steps of "adaptive evolution" a plant that begins with the manufacture of microscopes may find itself manufacturing perfumes a century later, without a single one of the original kinds of tools and processes still employed. *Secondly*, the example of conversion of tools to war and then to peace indicates that our problem

[5] The case of the automobile manufacturer is oversimplified since we have deliberately 'Shored problems of labor, management, and raw materials, which will be treated in detail in Chapters 9 and 11. Theoretically, the total supply of war jobs will be distributed to the total supply of manufacturers in such a way that the total work of re-tooling and of manufacture of the desired items will be least (cf. Chap. 5).

of economy is twofold, since it involves not only the more familiar important problem of the Selection of economical means (tools) but also the somewhat less familiar but no less important problem of the selection of economical objectives (jobs), in the *reciprocal economy* of matching tools to jobs and jobs to tools.

The above example is tendered only to illustrate the general relativism of the fundamental problem of tools-and-jobs as well as the twofold economy of their reciprocal matching. Since these two considerations are obviously important in our theory, it is not unreasonable for the reader to ask even now just how we may hope to study them *quantitatively.*

Curiously enough, we shall find in the forms and functions of the entities of human speech an almost perfect example of the general relativism of tools and jobs and of the twofold economy of selection. For, as we shall later see more specifically, the forms and meanings of words represent merely a special case of tools that perform jobs. We shall find that the forms and functions of words are quite capable of being studied quantitatively by means of the objective methods of empiric Science, with results that will be applicable to the general economy of all tools and jobs.

Indeed, it will be the precise information that we gain from a study of the case of speech that will suggest how every individual may be viewed in his *entirety* as a single set of integrated tools-and-jobs; and that the *total* behavior of the individual can be viewed as a case in which tools-seek-jobs-and-jobs-seek-tools with a maximum economy of effort, This view of an individual as a set of tools-jobs does no violence to our commonsense feelings about the matter, as we shall now see as we turn from the case of the automobile manufacturer to that of an imaginary person called John who, we shall suppose, is in love with Mary.

John, after work, tidies himself before going to see Mary, who is similarly tidying herself to see John. In these two persons we have, theoretically, a case where jobs-seek-tools-and-tools-seek-jobs. Each may be viewed as both a set of tools and as a set of jobs for the other person. Together they face a period of reciprocal adjustment, during which each side alters its tools and jobs to effect a more economical "match." In this respect John (or Mary) is quite similar to our previously discussed automobile manufacturer, who also had to alter his tools and jobs to match them more economically to the jobs and tools of others.

Yet in saying that John (or Mary) is a set of tools-and-jobs that is seeking a complementary set of jobs-and-tools, we are obviously dealing with two different economic problems in each person, Thus, in the case of John, there is the *first* problem of organizing John's own individual set of tools in such a way that they will operate with maximum economy in performing their jobs of self-support, self-defense, and procreation. Then there is the *second* problem of economically moving John as a unit system of tools over the earth's surface in the quest of jobs for his tools and of tools for his jobs (e.g., John seeks Mary). Clearly, these two economies to which John is continually subject are not the same, Yet they have one salient feature in common: in either case there is always the problem of moving matter-energy over paths of least effort, whether the matter-energy thus moved represents John's individual tools in operation, or whether it represents John as a total whole.

In other words, we may say two things about John in reference to our theoretical paths of least effort, First, we may say that John *is* a *set of paths* over which matter-energy proceeds into his system of toolage, through his system of toolage, and out of his system of toolage. Secondly, John, as a unit, *takes paths.* According to our Principle, all these paths, of whatever sort, will be paths of least effort, even though John's *intrasystematic* set of paths may seem to be rigidly stereotyped and determinate, whereas John's *extra-systematic* unit action may seem to be comparatively optional and indeterminate.

Paths of least effort are only probable paths, regardless of the comparative degrees of probabilities that the paths in question will be taken, If we now inspect more closely the manner in which John selects his *extra-systematic* path to his rendezvous with Mary, we shall gain further insight into the degree of precision to which any path of least effort is calculated, and also into the general economy of a perseveration, or a repetition, of activity that accounts for the apparently stereotyped rigidity of John's *intrasystematic* paths.

To begin, let us ask whether John, in selecting a path of least effort to Mary, will perchance take a slide rule and surveyor's transit, and calculate his path to Mary with the precision of a civil engineer who is surveying a roadbed for a railway that is to wind through a mountain range where every inch counts. Obviously not, and for a very good economic reason: *the work of calculating a path of least effort must be included in the total work of taking the path of least effort.* Nothing is gained by calculating a particular path of least effort to a greater degree of precision, when the added work of so calculating it is not more than offset by the work that is saved by using the more precisely calculated path. John, therefore, in selecting an easiest probable path to his objective, will minimize the total effort of calculating and of taking the path in question. The same will theoretically be the case of every other path of least effort.

This consideration leads to a second one. If John expects to take the same path repeatedly, then he can afford to calculate it more precisely. Since the additional work of calculation can be distributed over the repeated trips. From this we see that there is an inherent economy of effort in repeatedly taking the same paths, since one saves thereby the work of repeated calculations, In short, there is an economy in the repetitiveness of ones acts of behavior, Thence the growth of "habits."

We mention this consideration in order to suggest that the *intrasystemic* paths over which John's individual tools behave *within* his own system will still be paths of least effort, even though they may seem to be Stereotyped to the point of complete rigidity because of the relatively high frequency of recurrence of the action in question. The sheer fact that our Physiological behavior is highly predictable does not preclude the possibility that it takes place over paths of least effort; on the contrary, as we shall later argue in detail, intrasystematic paths are stereotyped because they are frequent, *and the reverse.*

This view of John as simultaneously both *taking* paths and *being* paths leads ultimately to a profound question in dynamics that will occupy our attention in the course of our demonstration.

If we restrict our attention to John as a set of paths in reference to which matter-energy moves into John's system, through John's system, and out of John's system, we note that there is nothing in this transient matter-energy that can be called permanently "John." And since the matter-energy is not John-or, if one will, since the actual physical tools of his system are not John-what then is John?

All that is left, according to our theory, is the *system* of paths itself, over which the matter-energy moves while acting in the capacity of being John's tools in operation. Yet does even this system of paths represent what is permanently John? Obviously not, since these paths are only probable paths and by no means fixed. Indeed, we know that the particular system of paths of an aged man are far from being the same as his system of paths when he was an embryo. Even as matter-energy is continually moving over paths, so too paths are continually changing.

And yet if all the matter-energy in John's system of tools in operation is transient and ephemeral to the system, and if the paths, or processes, are also variable, what, then, is left over in the total phenomenon to represent that apparently enduring entity called John?

This is clearly a primary question that is inherent in our concept of tools-and-jobs and one which will confront every other investigator into the field of biosocial dynamics. The question is simple: What is John?

At present we merely point to the existence of this question, which we shall candidly face in the course of our demonstration, when we shall attempt to answer it. If we mention it now, it is only in order to suggest what is attached to the view that an individual is a set of tools-and-jobs which in turn seeks jobs-and-tools.

We shall later define more precisely what we mean by *tools* and by *jobs* and by the reciprocal economy of *matching* the two. Our present discussion of the topic is intended merely to indicate the scope of the Principle of Least Effort, and to suggest a possibly fruitful manner of studying the problem of the economy of a person's *total* behavior.

This *total* behavior, as we have seen, consists of two economies. John is not only subject to the economy of organizing his own *individual* self; he is also subject to the economy of a *collective* organization with what lies outside his system. For the convenience of exposition, we shall divide our demonstration into two parts, the first devoted primarily to a discussion of the economy of the organization of the individual, and the second to the economy of the organization of the *collective* group of individuals in reference to each other and to the rest of nature,

We can see from the very nature of the case that the one economy of the individual will influence the other economy of the collective group, and the reverse.

17.2.8 Previous Studies

Now that we have discussed the meaning of the Principle of Least Effort and have indicated its general scope, let us review earlier studies of the topic before presenting a prospectus of the chief steps of our own ensuing demonstration. We shall begin *(A)* with the earlier studies of *collective* human economy, as represented by the economists; then *(B)* we shall consider studies of *individual* economy as conducted by psychologists. This summary review will also serve to let us take a position towards these earlier studies.

Collective Economy

It is questionable whether our analysis of the particular case of the economy of digging a tunnel through the mountain added one single consideration that would not have been completely covered by the older schools of "classical" or "orthodox" economics. These schools have never been in any doubt that the factor that is minimized in such collective enterprises is "labor," or work. Nor do they view their topics so narrowly that they fail to take into account the concept of risk which, in turn, is intimately connected with the concept of probable work. A great deal of economic theory, therefore, belongs to the history of the topic we are considering.

Closely related to the field of economics are the fields of sociology and of cultural anthropology, to whose thinking we are also indebted. Earlier quantitative observations in these fields, as well as in the field of general biological ecology, will often be referred to in the course of our demonstration.

Individual Economy

Less widely known than the above studies, yet surely no less important, are the studies of the factor of work in the motivation of human behavior as conducted with painstaking care by experimental psychologists, The results of these studies have led to the theoretical formulations of Drs. N. F. Miller and J, Dollard, and to those of Dr. C. L. Hull (see below), who is himself an experimentalist. Although the actual experimentation in the field of motivation has been too extensive to be discussed in detail here, we shall at least give a short account of some of the main steps of the theoretical development of the concept of economy.

Among the very earliest experimentalists to be concerned with the topic of work were three persons: (1) Dr. J. A. Gengerelli in his "The Principle of Maxima and Minima in Learning," '(2) Dr. L. S. Tsai in his "The Laws of Minimum Effort and Maximum Satisfaction in Animal Behavior," and (3) Dr. R, H. Waters in his "The Principle of Least Effort in Learning." By the term *effort* in these titles, the authors without exception mean *work*. In addition, there is also the "principle of least action" that was borrowed wholesale from the physicists by Dr. R. H. Wheeler, who advanced it as a Primary psychological principle without any supporting proof;[6] we mention Dr. Wheeler's "principle of least action in psychology" lest we otherwise seem to be unaware of it; and we discard it for the reason pointed out by Dr. Waters: Dr. Wheeler has in fact done nothing scientifically except, at best, to enunciate a postulate which he saw fit to make, yet for the validity of which there is not the slightest bit of evidence.

As to the other three persons, Dr. Gengerelli states his *Principle of Maxima and Minima* as follows:

"The behavior of an organism elicited by a given stimulating situation which affords relief to an internal need of that organism tends, with repetition, to approach, in terms of time, space, and effort involved, the minimal limit compatible with the relief of that need; the nature of the limit being defined by the structure of the organism and of the external situation."

Dr. Tsai in turn states his *Law of Minimum Effort* as follows:

"Among several alternatives of behavior leading to equivalent satisfaction of some potent organic need, the animal, within the limits of its discriminative ability, tends

[6] In the original simple terms of Maupertuis. the principle of least action states that when a mass, M, moves from a given point at a given moment of time to another point at another moment of time, it will proceed along a course in which the sum of all products of all masses when multiplied by their respective distances moved and by their respective velocities will be a minimum.

finally to select that which involves the least expenditure of energy."

Of these two principles, the one by Tsai is preferable because, as Dr. Waters pointed out (see below), Tsai frankly recognizes the limiting effect of the discriminative capacity of an animal; in short, an animal's "choices" are limited to those of whose existence the animal is aware. We should like to point out that both of the above propositions contain terms that need to be defined objectively.

Dr. Waters, after criticizing the above, showed the need of devising experimental situations that will separate the respective paths of least work, of least distance, and of least time, so that the preferred minimum can be observed. Dr. Waters own experimental setup to this end did not disclose the unambiguous minimum he had expected. The rats in Dr. Waters' experimental labyrinth showed a marked preference for paths along a wall - a preference that has itself been studied quantitatively with great insight by the physiologist, Dr. W. J. Crozier and G. Pincus.

Dr. Waters presented quite cautiously a rationalization of his observations in terms of least work, to which he added many qualifications that merit careful study. In closing his argument, Dr. Waters made the following statement (p. 17):

"Thus Theseus, after slaying the minotaur, found his way out of the labyrinth and to his loved one by following the string which he had carried with him into the labyrinth. Perhaps this was not the most direct route in terms of distance, time, or effort, but it was the only sure way he had of escaping. Likewise our rats found that by sticking to the outside pathways they more readily achieved the goal."[7]

This simple statement represents one of those anomalies in the history of science that always merits mention, though in mentioning it we most certainly do not mean to discredit Dr. Waters' excellent experimentation, much less Dr. Waters' admirable cautiousness about the principle of least work. We merely call attention to the fact that the case of Theseus suffices to disprove the Gengerelhi-Tsai principle of *least work* that Dr. Waters was trying to establish, and inadvertently provides instead an excellent example of *least average probable rate of work* (or our Least Effort) which we believe is the correct minimum. But to continue, in 1943 Dr. C. L. "in his *Principles of Behavior* set forth three postulates that relate to reactive inhibitions, on the basis of which he developed a corollary law of *less work* (p. 294) as follows:

"If two or more behavior sequences, each involving a different amount of work (W), have been equally well reinforced an equal number of times, the organism will gradually learn to choose the less laborious behavior sequence leading to the attainment of the reinforcing state of affairs."

Hull's principle of *less work* was based upon the research of those whom we have already discussed above, and upon that of Dr. R. S. Crutchfield's study of "Psychological Distance as a Function of Psychological Need." In 1944, a year after Hull's publication, Dr. M. E, Thompson in "An Experimental Investigation of the Gradient of Reinforcement in Maze Learning"" provided a further excellent set of observations in support of Hull's corollary principle of less work."

Hull's corollary principle, as stated by him, seems, like that of Tsai's, to be also a corollary of our Principle of Least Effort, as will be apparent from our demonstration, Thus, for example, when there are two or more possible courses of activity from one given point-moment to another given point-moment, for which the prerequisites are

[7] Perhaps rats running alonr a wall are less likely to he hit.

the same and of which the consequences are the same, then theoretically the course of least work will be adopted, since this course is also the course of Least Effort. We mention this consideration because many of the phenomena of speech, if viewed in isolation over a short period of time, will be found to conform to Hull's corollary of less work. The same would seem to apply to many other restricted experimental situations.

In addition to the above approach by experimental psychologists, of which the foregoing account is only a barest outline, there was the general theoretical study of "vectorial analysis" by Kurt Lewin in his *Principles of Topological Psychology.* This study, though stimulating to read, suffered both from a paucity of supporting data and from the lack of a satisfactory theory to account "vectorially" for the progression of a person from One situation to the other, On the other hand, the gifted experimentalists, Drs. J. F. Brown and A. C. Voth, have observed instances of "topological vectors" in the field of vision." Nor should we fail to mention, if only in P8ssing, the extremely interesting observations of the social behavior of ants as made by the topological psychologist and biologist, Dr. T. C. Schneirla.[8]

The writer's publications in the field of the dynamics of behavior began in 1929, in which cases of least work were found in the speech process.

The foregoing outline is by no means complete. In the following chapters we shall discuss in detail the observations of other investigators, including those of sociologists, which for that reason are not included above. The above outline serves to suggest both the increasing interest on the part of objective scientists in the topic of biosocial dynamics, and a growing feeling of the urgent need for a single unifying biosocial principle.

Only by venturing to set up general unifying principles for others to criticize and to prune and to restructure, even as those persons we have just mentioned have ventured and are venturing to do-and as we shall also attempt to do-may we hope, I believe, ultimately to disclose the desired principle.

[8] See, for example, O. H. Mowrer and H. M. Jones, "Extinction and Behavior Variability as Function of Effortfulness of Task"; J. E. Dc Camp, "Relative distance as a factor in the white rat's selection of a path.' etc.

Chapter 18

The Maximum Likelihood Method

18.1 Data Modelling

Given a set of observations, one often wants to condense and summarize the data by fitting it to a "model" that depends on adjustable parameters. Sometimes the model is simply a convenient class of functions, such as polynomials or Gaussians, and the fit supplies the appropriate coefficients. Other times, the model's parameters come from some underlying theory that the data are supposed to satisfy; examples are coefficients of rate equations in a complex network of chemical reactions, or orbital elements of a binary star. Modeling can also be used as a kind of constrained interpolation, where you want to extend a few data points into a continuous function, but with some underlying idea of what that function should look like.

The basic approach in all cases is usually the same: You choose or design a *figure-of-merit function* ("merit function," for short) that measures the agreement between the data and the model with a particular choice of parameters. The merit function is conventionally arranged so that small values represent close agreement. The parameters of the model are then adjusted to achieve a minimum in the merit function, yielding *best-fit parameters*. The adjustment process is thus a problem in minimization in many dimensions.

There are important issues that go beyond the mere finding of best-fit parameters. Data are generally not exact. They are subject to *measurement errors* (called *noise* in the context of signal-processing). Thus, typical data never exactly fit the model that is being used, even when that model is correct. We need the means to assess whether or not the model is appropriate, that is, we need to test the *goodness-of-fit* against some useful statistical standard.

We usually also need to know the accuracy with which parameters are determined by the data set. In other words, we need to know the likely errors of the best-fit parameters.

Finally, it is not uncommon in fitting data to discover that the merit function is not unimodal, with a single minimum. In some cases, we may be interested in global rather than local questions. Not, "how good is this fit?" but rather, "how sure am I that there is not a *very much better* fit in some corner of parameter space?" This kind of problem is generally quite difficult to solve.

The important message we want to deliver is that fitting of parameters is not the end-all of parameter estimation. To be genuinely useful, a fitting procedure should provide (i) parameters, (ii) error estimates on the parameters, and (iii) a statistical measure of goodness-of-fit. When the third item suggests that the model is an unlikely match to the data, then items (i) and (ii) are probably worthless. Unfortunately, many practitioners of parameter estimation never proceed beyond item (i). They deem a fit acceptable if a graph of data and model "looks good." This approach is known as *chi-by-eye*. Luckily, its practitioners get what they deserve.

18.2 Maximum Likelihood Estimators

Suppose that we are fitting N data points (x_i, y_i) $i = 1, 2, \cdots, N$, to a model that has M adjustable parameters a_j, $j = 1, 2, \cdots, M$. The model predicts a functional relationship between the measured independent and dependent variables,

$$y(x) = y\left(x; a_1, a_2, \cdots, a_M\right) \tag{18.1}$$

where the dependence on the parameters is indicated explicitly on the right-hand side. What, exactly, do we want to minimize to get fitted values for the a_j's? The first thing that comes to mind is the familiar least-squares fit,

$$\text{minimize over } a_1, a_2, \cdots, a_M : \quad \sum_{i=1}^{N} \left[y_i - y\left(x_i; a_1, a_2, \cdots, a_M\right)\right]^2 \tag{18.2}$$

But where does this come from? What general principles is it based on? The answer to these questions takes us into the subject of *maximum likelihood estimators.*

Given a particular data set of x_i's and y_i's, we have the intuitive feeling that some parameter sets $a_1, a_2, \cdots, a_M$ are very unlikely - those for which the model function $y(x)$ looks *nothing like* the data -while others may be very likely -those that closely resemble the data. How can we quantify this intuitive feeling? How can we select fitted parameters that are "most likely" to be correct? It is not meaningful to ask the question, "What is the probability that a particular set of fitted parameters $a_1, a_2, \cdots, a_M$ is correct?" The reason is that there is no statistical universe of models from which the parameters are drawn. There is just one model, the correct one, and a statistical universe of data sets that are drawn from it!

That being the case, we can, however, turn the question around, and ask, "*Given a particular set of parameters*, what is the probability that this data set could have occurred?" If the y_i's take on continuous values, the probability will always be zero unless we add the phrase, "...plus or minus some fixed Δy on each data point." So let's always take this phrase as understood. If the probability of obtaining the data set is infinitesimally small, then we can conclude that the parameters under consideration are "unlikely" to be right. Conversely, our intuition tells us that the data set should not be too improbable for the correct choice of parameters.

In other words, we identify the probability of the data given the parameters (which is a mathematically computable number), as the *likelihood* of the parameters given the data. This identification is entirely based on intuition. It has no formal mathematical basis in and of itself; as we already remarked, statistics is *not* a branch of mathematics!

Once we make this intuitive identification, however, it is only a small further step to decide to fit for the parameters $a_1, a_2, \cdots, a_M$ precisely by finding those values that *maximize* the likelihood defined in the above way. This form of parameter estimation is *maximum likelihood estimation.*

We are now ready to make the connection to 18.2. Suppose that each data point y_i has a measurement error that is independently random and distributed as a normal (Gaussian) distribution around the "true" model $y(x)$. And suppose that the standard deviations σ of these normal distributions are the same for all points. Then the probability of the data set is the product of the probabilities of each point,

$$P \text{ proportional to } \prod_{i=1}^{N} \left[\exp\left(-\frac{1}{2} \left(\frac{y_i - y\,(x_i)}{\sigma} \right)^2 \right) \Delta y \right] \tag{18.3}$$

Notice that there is a factor Δy in each term in the product. Maximizing 18.3 is equivalent to maximizing its logarithm, or minimizing the negative of its logarithm, namely,

$$\left[\sum_{i=1}^{N} \frac{(y_i - y\,(x_i))^2}{2\sigma^2} \right] - N \log \Delta y \tag{18.4}$$

Since N, σ and Δy are all constants, minimizing this equation is equivalent to minimizing 18.2.

What we see is that least-squares fitting *is* a maximum likelihood estimation of the fitted parameters *if* the measurement errors are independent and normally distributed with constant standard deviation. Notice that we made no assumption about the linearity or nonlinearity of the model $y(x; a_1, a_2, \cdots, a_M)$ in its parameters $a_1, a_2, \cdots, a_M$. Just below, we will relax our assumption of constant standard deviations and obtain the very similar formulas for what is called "chi-square fitting" or "weighted least-squares fitting." First, however, let us discuss further our very stringent assumption of a normal distribution.

For a hundred years or so, mathematical statisticians have been in love with the fact that the probability distribution of the sum of a very large number of very small random deviations almost always converges to a normal distribution. (the *central limit theorem*) This infatuation tended to focus interest away from the fact that, for real data, the normal distribution is often rather poorly realized, if it is realized at all. We are often taught, rather casually, that, on average, measurements will fall within $\pm\sigma$ of the true value 68 percent of the time, within $\pm 2\sigma$ 95 percent of the time, and within $\pm 3\sigma$ 99.7 percent of the time. Extending this, one would expect a measurement to be off by $\pm 20\sigma$ only one time out of 2×10^{88}. We all know that "glitches" are much more likely than *that*!

In some instances, the deviations from a normal distribution are easy to understand and quantify. For example, in measurements obtained by counting events, the measurement errors are usually distributed as a Poisson distribution, whose cumulative probability function was already discussed in 6.2. When the number of counts going into one data point is large, the Poisson distribution converges towards a Gaussian. However, the convergence is not uniform when measured in fractional accuracy. The more standard deviations out on the tail of the distribution, the larger the number of counts must be before a value close to the Gaussian is realized. The sign of the effect is always the same: The Gaussian predicts that "tail" events are much less likely than they actually (by Poisson) are. This causes such events, when they occur, to skew a least-squares fit much more than they ought.

Other times, the deviations from a normal distribution are not so easy to understand in detail. Experimental points are occasionally just *way off.* Perhaps the power flickered during a point's measurement, or someone kicked the apparatus, or someone wrote down a wrong number. Points like this are called *outliers.* They can easily turn a least-squares fit on otherwise adequate data into nonsense. Their probability of occurrence in the assumed Gaussian model is so small that the maximum likelihood estimator is willing to distort the whole curve to try to bring them, mistakenly, into line.

The subject of *robust statistics* deals with cases where the normal or Gaussian model is a bad approximation, or cases where outliers are important. We will discuss robust methods briefly later. All the sections between this one and that one assume, one way or the other, a Gaussian model for the measurement errors in the data. It it quite important that you keep the limitations of that model in mind, even as you use the very useful methods that follow from assuming it.

Finally, note that our discussion of measurement errors has been limited to *statistical errors*, the kind that will average away if we only take enough data. Measurements are also susceptible to *systematic errors* that will not go away with any amount of averaging. For example, the calibration of a metal meter stick might depend on its temperature. If we take all our measurements at the same wrong temperature, then no amount of averaging or numerical processing will correct for this unrecognized systematic error.

18.2.1 Chi-Square Fitting

If each data point (x_i, y_i) has its own, known standard deviation σ_i, then equation 18.3 is modified only by putting a subscript i on the symbol σ. That subscript also propagates docilely into 18.4, so that the maximum likelihood estimate of the model parameters is obtained by minimizing the quantity

$$\chi^2 \equiv \sum_{i=1}^{N} \left(\frac{y_i - y\left(x_i; a_1, a_2, \cdots, a_M\right)}{\sigma_i} \right)^2 \tag{18.5}$$

called the "chi-square."

To whatever extent the measurement errors actually *are* normally distributed, the quantity χ^2 is correspondingly a sum of N squares of normally distributed quantities, each normalized to unit variance. Once we have adjusted the $a_1, a_2, \cdots, a_M$ to minimize the value of χ^2, the terms in the sum are not all statistically independent. For models that are linear in the a's, however, it turns out that the probability distribution for different values of χ^2 at its minimum can nevertheless be derived analytically, and is the *chi-square distribution for $N - M$ degrees of freedom.* We learned how to compute this probability function using the incomplete gamma function. In particular, equation (6.2.18) gives the probability that the chi-square should exceed a particular value χ^2 by chance, where $\nu = N - M$ is the *number of degrees of freedom.* The quantity, or its complement $P \equiv 1 - Q$, is frequently tabulated in appendices to statistics books. It is quite common, and usually not too wrong, to assume that the chi-square distribution holds even for models that are not strictly linear in the a's.

This computed probability gives a quantitative measure for the goodness-of-fit of the model. If is a very small probability for some particular data set, then the apparent discrepancies are unlikely to be chance fluctuations. Much more probably either (i) the model is wrong - can be statistically rejected, or (ii) someone has lied to you about the size of the measurement errors σ_i - they are really larger than stated.

It is an important point that the chi-square probability Q does not directly measure the credibility of the assumption that the measurement errors are normally distributed. It assumes they are. In most, but not all, cases, however, the effect of non-normal errors is to create an abundance of outlier points. These decrease the probability Q, so that we can add another possible, though less definitive, conclusion to the above list: (iii) the measurement errors may not be normally distributed.

Possibility (iii) is fairly common, and also fairly benign. It is for this reason that reasonable experimenters are often rather tolerant of low probabilities Q. It is not uncommon to deem acceptable on equal terms any models with, say, $Q > 0.001$. This is not as sloppy as it sounds: Truly *wrong* models will often be rejected with vastly smaller values of Q, 10^{-18}, say. However, if day-in and day-out you find yourself accepting models with $Q \approx 10^{-3}$, you really should track down the cause.

If you happen to know the actual distribution law of your measurement errors, then you might wish to *Monte Carlo simulate* some data sets drawn from a particular model. You can then subject these synthetic data sets to your actual fitting procedure, so as to determine both the probability distribution of the χ^2 statistic, and also the accuracy with which your model parameters are reproduced by the fit. The technique is very general, but it can also be very expensive.

At the opposite extreme, it sometimes happens that the probability Q is too large, too near to 1, literally too good to be true! Non-normal measurement errors cannot in general produce this disease, since the normal distribution is about as "compact" as a distribution can be. Almost always, the cause of too good a chi-square fit is that the experimenter, in a "fit" of conservativism, has *overestimated* his or her measurement errors. Very rarely, too good a chi-square signals actual fraud, data that has been "fudged" to fit the model.

A rule of thumb is that a "typical" value of χ^2 for a "moderately" good fit is $\chi^2 \approx \nu$. More precise is the statement that the χ^2 statistic has a mean ν and a standard deviation $\sqrt{2\nu}$, and, asymptotically for large ν, becomes normally distributed. In some cases the uncertainties associated with a set of measurements are not known in advance, and considerations related to χ^2 fitting are used to derive a value for σ. If we assume that all measurements have the same standard deviation, $\sigma_i = \sigma$, and that the model does fit well, then we can proceed by first assigning an arbitrary constant σ to all points, next fitting for the model parameters by minimizing χ^2, and finally recomputing

$$\sigma^2 = \sum_{i=1}^{N} \frac{(y_i - y(x_i))^2}{N - M} \tag{18.6}$$

Obviously, this approach prohibits an independent assessment of goodness-of-fit, a fact occasionally missed by its adherents. When, however, the measurement error is not known, this approach at least allows *some* kind of error bar to be assigned to the points.

If we take the derivative of equation 18.5 with respect to the parameters a_k, we obtain equations that must hold at the chi-square minimum,

$$0 = \sum_{i=1}^{N} \left(\frac{y_i - y\left(x_i\right)}{\sigma_i^2} \right) \left(\frac{\partial y\left(x_i; a_1, a_2, \cdots, a_M\right)}{\partial a_k} \right) \qquad k = 1, 2, \cdots, M \qquad (18.7)$$

which is, in general, a set of M nonlinear equations for the M unknown a_k. Several of the procedures described subsequently derive from this equation and its specializations.

18.3 Reading: Theory of Games by S. Vajda

18.3.1 Introduction to the Reading

18.3.2 The Paper

Chapter 19

The Linear Least Squares Method 1

19.1 Fitting Data to a Straight Line

A concrete example will make the considerations of the previous section more meaningful. We consider the problem of fitting a set of N data points (x_i, y_i) to a straight-line model

$$y(x) = y(x; a, b) = a + bx \tag{19.1}$$

This problem is often called *linear regression*, a terminology that originated, long ago, in the social sciences. We assume that the uncertainty σ_i associated with each measurement y_i is known, and that the x_i's (values of the dependent variable) are known exactly.

To measure how well the model agrees with the data, we use the chi-square merit function 18.5, which in this case is

$$\chi^2(a, b) = \sum_{i=1}^{N} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2 \tag{19.2}$$

If the measurement errors are normally distributed, then this merit function will give maximum likelihood parameter estimations of a and b; if the errors are not normally distributed, then the estimations are not maximum likelihood, but may still be useful in a practical sense.

Equation 19.2 is minimized to determine a and b. At its minimum, derivatives of $\chi^2(a, b)$ with respect to a, b vanish.

$$0 = \frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2} 0 = \frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^{N} \frac{x_i \left(y_i - a - bx_i \right)}{\sigma_i^2} \tag{19.3}$$

These conditions can be rewritten in a convenient form if we define the following sums:

$$S \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2}; S_x \equiv \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}; S_y \equiv \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}; S_{xx} \equiv \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}; S_{xy} \equiv \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} \tag{19.4}$$

With these definitions 19.3 becomes

$$aS + bS_x = S_y; \qquad aS_x + bS_{xx} = S_{xy} \tag{19.5}$$

The solution of these two equations in two unknowns is calculated as

$$\Delta \equiv SS_{xx} - (S_x)^2\,; a = \frac{S_{xx}S_y - S_xS_{xy}}{\Delta}; b = \frac{S_{xy}S - S_xS_y}{\Delta} \tag{19.6}$$

Equation 19.6 gives the solution for the best-fit model parameters a and b.

We are not done, however. We must estimate the probable uncertainties in the estimates of a and b, since obviously the measurement errors in the data must introduce some uncertainty in the determination of those parameters. If the data are independent, then each contributes its own bit of uncertainty to the parameters. Consideration of propagation of errors shows that the variance χ_f^2 in the value of any function will be

$$\sigma_f^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial f}{\partial y_i}\right)^2 \tag{19.7}$$

For the straight line, the derivatives of a and b with respect to y i can be directly evaluated from the solution:

$$\frac{\partial a}{\partial y_i} = \frac{S_{xx} - S_x x_i}{\sigma_i^2 \Delta}; \frac{\partial b}{\partial y_i} = \frac{S x_i - S_x}{\sigma_i^2 \Delta} \tag{19.8}$$

Summing over the points as in 19.7, we get

$$\sigma_a^2 = \frac{S_{xx}}{\Delta}; \sigma_b^2 = \frac{S}{\Delta} \tag{19.9}$$

which are the variances in the estimates of a and b, respectively. We will see later that an additional number is also needed to characterize properly the probable uncertainty of the parameter estimation. That number is the *covariance* of a and b, and (as we will see below) is given by

$$Cov(a, b) = \frac{-S_x}{\Delta} \tag{19.10}$$

The coefficient of correlation between the uncertainty in a and the uncertainty in b, which is a number between -1 and 1, follows from 19.10,

$$r_{ab} = \frac{-S_x}{\sqrt{SS_{xx}}} \tag{19.11}$$

A positive value of r_{ab} indicates that the errors in a and b are likely to have the same sign, while a negative value indicates the errors are *anti-correlated*, likely to have opposite signs.

We are *still* not done. We must estimate the goodness-of-fit of the data to the model. Absent this estimate, we have not the slightest indication that the parameters a and b in the model have any meaning at all! The probability Q that a value of chi-square as *poor* as the value 19.2 should occur by chance is

$$Q = Gamma\left(\frac{N-2}{2}, \frac{\chi^2}{2}\right) \tag{19.12}$$

If Q is larger than, say, 0.1, then the goodness-of-fit is believable. If it is larger than, say, 0.001, then the fit *may* be acceptable if the errors are non-normal or have been moderately underestimated. If Q is less than 0.001 then the model and/or estimation procedure can rightly be called into question.

If you do not know the individual measurement errors of the points σ_i, and are proceeding (dangerously) to use equation 18.6 for estimating these errors, then here is the procedure for estimating the probable uncertainties of the parameters a and b: Set $\sigma_i \equiv 1$ in all equations through 19.6, and multiply σ_a and σ_b, as obtained from equation 19.9, by the additional factor $\sqrt{\chi^2/(N-2)}$, where χ^2 is computed by 19.2 using the fitted parameters a and b. As discussed above, this procedure is equivalent to *assuming* a good fit, so you get no independent goodness-of-fit probability Q.

We promised a relation between the linear correlation coefficient r and a goodness-of-fit measure, χ^2 (equation 19.2). For unweighted data (all $\sigma_i = 1$), that relation is

$$\chi^2 = \left(1 - r^2\right) Nvar\left(y_1, y_2, \cdots, y_N\right) \tag{19.13}$$

where

$$Nvar\left(y_1, y_2, \cdots, y_N\right) \equiv \sum_{i=1}^{N} \left(y_i - \overline{y}\right)^2 \tag{19.14}$$

For data with varying weights σ_i, the above equations remain valid if the sums in equation (14.5.1) are weighted by $1/\sigma_i^2$.

When the weights σ are known in advance, the calculations exactly correspond to the formulas above. However, when weights σ are unavailable, the routine *assumes* equal values of σ for each point and *assumes* a good fit.

19.2 Straight-Line Data with Errors in Both Coordinates

If experimental data are subject to measurement error not only in the y_i's, but also in the x_i's, then the task of fitting a straight-line model

$$y(x) = a + bx \tag{19.15}$$

is considerably harder. It is straightforward to write down the χ^2 merit function for this case,

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_{y_i}^2 + b^2\sigma_{x_i}^2} \tag{19.16}$$

where σ_{x_i} and σ_{y_i} are, respectively, the x and y standard deviations for the i^{th} point. The weighted sum of variances in the denominator of equation 19.16 can be understood both as the variance in the direction of the smallest χ^2 between each data point and the line with slope b, and also as the variance of the linear combination $y_i - a - bx_i$ of two random variables x_i and y_i,

$$Var\left(y_i - a - bx_i\right) = Var\left(y_i\right) + b^2 Var\left(x_i\right) = \sigma_{y_i}^2 + b^2\sigma_{x_i}^2 \equiv \frac{1}{w_i} \tag{19.17}$$

The sum of the square of N random variables, each normalized by its variance, is thus χ^2-distributed.

We want to minimize equation 19.16 with respect to a and b. Unfortunately, the occurrence of b in the denominator of equation 19.16 makes the resulting equation for the slope $\partial\chi^2/\partial b = 0$ nonlinear. However, the corresponding condition for the intercept, $\partial\chi^2/\partial a = 0$, is still linear and

$$a = \frac{\left[\sum_i w_i \left(y_i - bx_i\right)\right]}{\sum_i w_i} \tag{19.18}$$

where the w_i's are defined by equation 19.17. A reasonable strategy, now, is to use an optimization method for minimizing a general one-dimensional function to minimize with respect to b, while using equation 19.18 at each stage to ensure that the minimum with respect to b is also minimized with respect to a.

Because of the finite error bars on the x_i's, the minimum χ^2 as a function of b will be finite, though usually large, when b equals infinity (line of infinite slope). The angle $\theta \equiv \tan^{-1} b$ is thus more suitable as a parametrization of slope than b itself. The value of χ^2 will then be periodic in with period π (not 2π!). If any data points have very small σ_y's but moderate or large σ_x's, then it is also possible to have a maximum in χ^2 near zero slope, $\theta \approx 0$. In that case, there can sometimes be two χ^2 minima, one at positive slope and the other at negative. Only one of these is the correct global minimum. It is therefore important to have a good starting guess for b (or θ). Our strategy, implemented below, is to scale the y_i's so as to have variance equal to the x_i's, then to do a conventional linear fit with weights derived from the (scaled) sum $\sigma_{y_i}^2 + \sigma_{x_i}^2$. This yields a good starting guess for b if the data are even *plausibly* related to a straight-line model.

Finding the standard errors σ_a and σ_b on the parameters a and b is more complicated. We will see that, in appropriate circumstances, the standard errors in a and b are the respective projections onto the a and b axes of the "confidence region boundary" where χ^2 takes on a value one greater than its minimum, $\chi^2 = 1$. In the linear case, these projections follow from the Taylor series expansion

$$\Delta\chi^2 \approx \frac{1}{2}\left[\frac{\partial^2\chi^2}{\partial a^2}\left(\Delta a\right)^2 + \frac{\partial^2\chi^2}{\partial b^2}\left(\Delta b\right)^2\right] + \frac{\partial^2\chi^2}{\partial a \partial b}\Delta a \Delta b \tag{19.19}$$

Because of the present nonlinearity in b, however, analytic formulas for the second derivatives are quite unwieldy; more important, the lowest-order term frequently gives a poor approximation to χ^2. Our strategy is therefore to find the roots of $\chi^2 = 1$ numerically, by adjusting the value of the slope b away from the minimum. It may occur that there are no roots at all -for example, if all error bars are so large that all the data points are compatible with each other. It is important, therefore, to make some effort at bracketing a putative root before refining it.

Because a is minimized at each stage of varying b, successful numerical root-finding leads to a value of a that minimizes χ^2 for the value of b that gives $\chi^2 = 1$. This (see Figure 19.1) directly gives the tangent projection of the confidence region onto the b axis, and thus σ_b. It does not, however, give the tangent projection of the confidence region onto the a axis. In the figure, we have found the point labeled

Figure 19.1: Standard errors for the parameters a and b. The point B can be found by varying the slope b while simultaneously minimizing the intercept a. This gives the standard error σ_b, and also the value s. The standard error σ_a can then be found by the geometric relation $\sigma_a^2 = s^2 + r^2$.

B; to find σ_a we need to find the point A. Geometry to the rescue: To the extent that the confidence region is approximated by an ellipse, then you can prove (see figure) that $\sigma_a^2 = s^2 + r^2$. The value of s is known from having found the point B. The value of r follows from equations 19.16 and 19.17 applied at the χ^2 minimum (point O in the figure), giving

$$r^2 = \frac{1}{\sum\limits_i w_i} \tag{19.20}$$

Actually, since b can go through infinity, this whole procedure makes more sense in (a, θ) space than in (a, b) space. That is in fact how the following program works. Since it is conventional, however, to return standard errors for a and b, not a and, we finally use the relation

$$\sigma_b = \frac{\sigma_\theta}{\cos^2\theta} \tag{19.21}$$

We caution that if b and its standard error are both large, so that the confidence region actually includes infinite slope, then the standard error σ_b is not very meaningful.

A final caution is that if the goodness-of-fit is not acceptable (returned probability is too small), the standard errors σ_a and σ_b are surely not believable. In dire circumstances, you might try scaling all your x and y error bars by a constant factor until the probability is acceptable (0.5, say), to get more plausible values for σ_a and σ_b.

19.3 General Linear Least Squares

An immediate generalization of the previous material is to fit a set of data points (x_i, y_i) to a model that is not just a linear combination of 1 and x (namely $a + bx$), but rather a linear combination of *any* M specified functions of x. For example, the functions could be 1, x, x^2, $\cdots$, x^{M-1}, in which case their general linear combination,

$$y(x) = a_1 + a_2x + a_3x^2 + \cdots + a_Mx^{M-1} \tag{19.22}$$

is a polynomial of degree $M - 1$. Or, the functions could be sines and cosines, in which case their general linear combination is a harmonic series.

The general form of this kind of model is

$$y(x) = \sum_{k=1}^{M} a_k X_k(x) \tag{19.23}$$

where $X_1(x)$, $X_2(x)$, $\cdots$, $X_M(x)$ are arbitrary fixed functions of x, called the *basis functions*.

Note that the functions $X_k(x)$ can be wildly nonlinear functions of x. In this discussion "linear" refers only to the model's dependence on its *parameters* a_k.

For these linear models we generalize the discussion of the previous section by defining a merit function

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - \sum\limits_{k=1}^{M} a_k X_k(x_i)}{\sigma_i} \right]^2 \tag{19.24}$$

As before, σ_i is the measurement error (standard deviation) of the i^{th} data point, presumed to be known. If the measurement errors are not known, they may all be set to the constant value $\sigma = 1$.

Once again, we will pick as best parameters those that minimize χ^2. There are several different techniques available for finding this minimum. Two are particularly useful, and we will discuss both in this section. To introduce them and elucidate their relationship, we need some notation.

Let A be a matrix whose $N \times M$ components are constructed from the M basis functions evaluated at the N abscissas x_i, and from the N measurement errors σ_i, by the prescription

$$A_{ij} = \frac{X_j(x_i)}{\sigma_i} \tag{19.25}$$

The matrix A is called the *design matrix* of the fitting problem. Notice that in general A has more rows than columns, $N \geq M$, since there must be more data points than model parameters to be solved for. (You can fit a straight line to two points, but not a very meaningful quintic!) The design matrix is shown schematically in Figure 19.2.

Also define a vector b of length N by

$$b_i = \frac{y_i}{\sigma_i} \tag{19.26}$$

and denote the M vector whose components are the parameters to be fitted, a_1, a_2, $\cdots$, a_M, by a.

19.3.1 Solution by Use of the Normal Equations

The minimum of 19.24 occurs where the derivative of χ^2 with respect to all M parameters a_k vanishes. Specializing equation 18.7 to the case of the model 19.23, this condition yields the M equations

$$0 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[y_i - \sum j = 1^M a_j X_j(x_i) \right] X_k(x_i) \qquad k = 1, 2, \cdots, M \tag{19.27}$$

Interchanging the order of summations, we can write 19.27 as the matrix equation

$$\sum j = 1^M \alpha_{kj} a_j = \beta_k \tag{19.28}$$

$$\begin{array}{cc|cccc} & & \multicolumn{4}{c}{\longleftarrow\ \text{basis functions}\ \longrightarrow} \\ & & X_1(\ ) & X_2(\ ) & \cdots & X_M(\ ) \\ \uparrow & x_1 & \frac{X_1(x_1)}{\sigma_1} & \frac{X_2(x_1)}{\sigma_1} & \cdots & \frac{X_M(x_1)}{\sigma_1} \\ & x_2 & \frac{X_1(x_2)}{\sigma_2} & \frac{X_2(x_2)}{\sigma_2} & \cdots & \frac{X_M(x_2)}{\sigma_2} \\ \text{data points} & \vdots & \vdots & & & \vdots \\ \downarrow & x_N & \frac{X_1(x_N)}{\sigma_N} & \frac{X_2(x_N)}{\sigma_N} & \cdots & \frac{X_M(x_N)}{\sigma_N} \end{array}$$

Figure 19.2: Design matrix for the least-squares fit of a linear combination of M basis functions to N data points. The matrix elements involve the basis functions evaluated at the values of the independent variable at which measurements are made, and the standard deviations of the measured dependent variable. The measured values of the dependent variable do not enter the design matrix.

where

$$\alpha_{kj} = \sum_{i=1}^{N} \frac{X_j(x_i) X_k(x_i)}{\sigma_i^2} \text{ or } [\alpha] = A^T \cdot A \tag{19.29}$$

an $M \times M$ matrix, and

$$\beta_k = \sum_{i=1}^{N} \frac{y_i X_k(x_i)}{\sigma_i^2} \text{ or } [\beta] = A^T \cdot b \tag{19.30}$$

a vector of length M.

The equations 19.27 or 19.28 are called the *normal equations* of the least-squares problem. They can be solved for the vector of parameters a by standard methods, notably LU decomposition and backsubstitution, Choleksy decomposition or Gauss-Jordan elimination. In matrix form, the normal equations can be written as either

$$[\alpha] \cdot a = [\beta] \text{ or } \left(A^T \cdot A\right) \cdot a = A^T \cdot b \tag{19.31}$$

The inverse matrix $C_{jk} \equiv [\alpha]^{-1}_{jk}$ is closely related to the probable (or, more precisely, *standard*) uncertainties of the estimated parameters a. To estimate these uncertainties, consider that

$$a_j = \sum_{k=1}^{M} [\alpha]^{-1}_{jk} \beta_k = \sum_{k=1}^{M} C_{jk} \left[\sum_{i=1}^{N} \frac{y_i X_k(x_i)}{\sigma_i^2} \right] \tag{19.32}$$

and that the variance associated with the estimate a_j can be found as in 19.7 from

$$\sigma^2(a_j) = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a_j}{\partial y_i} \right)^2 \tag{19.33}$$

Note that a_{jk} is independent of y_i , so that

$$\frac{\partial a_j}{\partial y_i} = \sum_{k=1}^{M} \frac{C_{jk} X_k(x_i)}{\sigma_i^2} \tag{19.34}$$

Consequently, we find that

$$\sigma^2(a_j) = \sum_{k=1}^{M} \sum_{l=1}^{M} C_{jk} C_{jl} \left[\sum_{i=1}^{N} \frac{X_k(x_i) X_l(x_i)}{\sigma_i^2} \right] \tag{19.35}$$

The final term in brackets is just the matrix $[\alpha]$. Since this is the matrix inverse of $[C]$, 19.35 reduces immediately to

$$\sigma^2(a_j) = C_{jj} \tag{19.36}$$

In other words, the diagonal elements of $[C]$ are the variances (squared uncertainties) of the fitted parameters a. It should not surprise you to learn that the off-diagonal elements C_{jk} are the covariances between a_j and a_k; but we shall defer discussion of these to the next lecture.

Since we wish to compute not only the solution vector a but also the covariance matrix $[C]$, it is most convenient to use Gauss-Jordan elimination to perform the linear algebra. In theory, since $A^T \cdot A$ is positive definite, Cholesky decomposition is the most efficient way to solve the normal equations. However, in practice most of the computing time is spent in looping over the data to form the equations, and Gauss-Jordan is quite adequate.

We need to warn you that the solution of a least-squares problem directly from the normal equations is rather susceptible to roundoff error. An alternative, and preferred, technique involves QR decomposition of the design matrix A. This is essentially what we did for fitting data to a straight line, but without invoking all the machinery of QR to derive the necessary formulas.

Frequently it is a matter of “art” to decide which parameters a_k in a model should be fit from the data set, and which should be held constant at fixed values, for example values predicted by a theory or measured in a previous experiment. One wants, therefore, to have a convenient means for “freezing” and “unfreezing” the parameters a_k.

19.4 Reading: On the Average and Scatter by M. J. Moroney

19.4.1 Introduction to the Reading

19.4.2 The Paper

Chapter 20

The Linear Least Squares Method 2

20.1 Confidence Limits on Estimated Model Parameters

Several times already in this chapter we have made statements about the standard errors, or uncertainties, in a set of M estimated parameters a. We have given some formulas for computing standard deviations or variances of individual parameters, as well as some formulas for covariances between pairs of parameters.

In this section, we want to be more explicit regarding the precise meaning of these quantitative uncertainties, and to give further information about how quantitative confidence limits on fitted parameters can be estimated. The subject can get somewhat technical, and even somewhat confusing, so we will try to make precise statements, even when they must be offered without proof.

Figure 20.1 shows the conceptual scheme of an experiment that "measures" a set of parameters. There is some underlying true set of parameters a_{true} that are known to Mother Nature but hidden from the experimenter. These true parameters are statistically realized, along with random measurement errors, as a measured data set, which we will symbolize as $\mathcal{D}_{(0)}$. The data set $\mathcal{D}_{(0)}$ *is* known to the experimenter. He or she fits the data to a model by χ^2 minimization or some other technique, and obtains measured, i.e., fitted, values for the parameters, which we here denote $a_{(0)}$.

Because measurement errors have a random component, $\mathcal{D}_{(0)}$ is not a unique realization of the true parameters a_{true}. Rather, there are infinitely many other realizations of the true parameters as "hypothetical data sets" each of which *could* have been the one measured, but happened not to be. Let us symbolize these by $\mathcal{D}_{(1)}$, $\mathcal{D}_{(2)}$, $\cdots$. Each one, had it been realized, would have given a slightly different set of fitted parameters, $a_{(1)}$, $a_{(2)}$, $\cdots$, respectively. These parameter sets $a_{(i)}$ therefore occur with some probability distribution in the M-dimensional space of all possible parameter sets a. The actual measured set $a_{(0)}$ is one member drawn from this distribution.

Even more interesting than the probability distribution of $a_{(i)}$ would be the distribution of the difference $a_{(i)} - a_{true}$. This distribution differs from the former one by a translation that puts Mother Nature's true value at the origin. If we knew *this* distribution, we would know everything that there is to know about the quantitative uncertainties in our experimental measurement $a_{(0)}$.

So the name of the game is to find some way of estimating or approximating the probability distribution of $a_{(i)} - a_{true}$ without knowing a_{true} and without having available to us an infinite universe of hypothetical data sets.

20.1.1 Monte Carlo Simulation of Synthetic Data Sets

Although the measured parameter set $a_{(0)}$ is not the true one, let us consider a fictitious world in which it *was* the true one. Since we hope that our measured parameters are not *too* wrong, we hope that that fictitious world is not too different from the actual world with parameters a_{true}. In particular, let us hope — no, let us *assume* — that the shape of the probability distribution $a_{(i)} - a_{(0)}$ in the fictitious world is the same, or very nearly the same, as the shape of the probability distribution $a_{(i)} - a_{true}$ in the real world. Notice that we are not assuming that $a_{(0)}$ and a_{true} are equal; they are certainly not. We are only assuming that the way in which random errors enter the experiment and data analysis does not vary rapidly as a function of a_{true}, so that $a_{(0)}$ can serve as a reasonable surrogate.

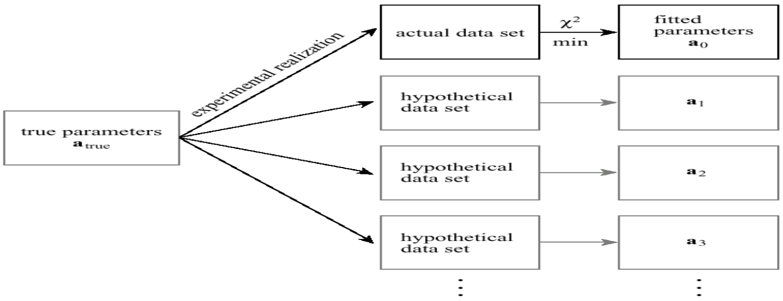


Figure 20.1: A statistical universe of data sets from an underlying model. True parameters a_{true} are realized in a data set, from which fitted (observed) parameters $a_{(0)}$ are obtained. If the experiment were repeated many times, new data sets and new values of the fitted parameters would be obtained.

Now, often, the distribution of $a_{(i)} - a_{(0)}$ in the fictitious world *is* within our power to calculate (see Figure 20.2). If we know something about the process that generated our data, given an assumed set of parameters $a_{(0)}$, then we can usually figure out how to *simulate* our own sets of "synthetic" realizations of these parameters as "synthetic data sets." The procedure is to draw random numbers from appropriate distributions so as to mimic our best understanding of the underlying process and measurement errors in our apparatus. With such random draws, we construct data sets with exactly the same numbers of measured points, and precisely the same values of all control (independent) variables, as our actual data set $\mathcal{D}_{(0)}$. Let us call these simulated data sets $\mathcal{D}^S_{(1)}$, $\mathcal{D}^S_{(2)}$, $\cdots$. By construction these are supposed to have exactly the same statistical relationship to $a_{(0)}$ as the $\mathcal{D}_{(i)}$'s have to a_{true}. (For the case where you don't know enough about what you are measuring to do a credible job of simulating it, see below.)

Next, for each $\mathcal{D}^S_{(j)}$, perform exactly the same procedure for estimation of parameters, e.g., χ^2 minimization, as was performed on the actual data to get the parameters $a_{(0)}$, giving simulated measured parameters $a^S_{(1)}$, $a^S_{(2)}$, $\cdots$. Each simulated measured parameter set yields a point $a^S_{(i)} - a_{(0)}$. Simulate enough data sets and enough derived simulated measured parameters, and you map out the desired probability distribution in M dimensions.

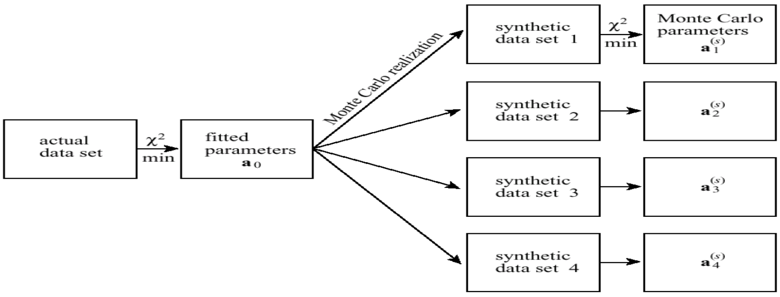


Figure 20.2: Monte Carlo simulation of an experiment. The fitted parameters from an actual experiment are used as surrogates for the true parameters. Computer-generated random numbers are used to simulate many synthetic data sets. Each of these is analyzed to obtain its fitted parameters. The distribution of these fitted parameters around the (known) surrogate true parameters is thus studied.

In fact, the ability to do *Monte Carlo simulations* in this fashion has revolutionized many fields of modern experimental science. Not only is one able to characterize the errors of parameter estimation in a very precise way; one can also try out on the computer different methods of parameter estimation, or different data reduction techniques, and seek to minimize the uncertainty of the result according to any desired criteria. Offered the choice between mastery of a five-foot shelf of analytical statistics books and middling ability at performing statistical Monte Carlo simulations, we would surely choose to have the latter skill.

20.1.2 Quick-and-Dirty Monte Carlo: The Bootstrap Method

Here is a powerful technique that can often be used when you don't know enough about the underlying process, or the nature of your measurement errors, to do a credible Monte Carlo simulation. Suppose that your data set consists of N *independent and identically distributed* (or iid) "data points." Each data point probably consists of several numbers, e.g., one or more control variables (uniformly distributed, say, in the range that you have decided to measure) and one or more associated measured values (each distributed however Mother Nature chooses). "Iid" means that the sequential order of the data points is not of consequence to the process that you are using to get the fitted parameters a. For example, a χ^2 sum does not care in what order the points are added. Even simpler examples are the mean value of a measured quantity, or the mean of some function of the measured quantities.

The *bootstrap method* uses the actual data set $\mathcal{D}^S_{(0)}$, with its N data points, to generate any number of synthetic data sets $\mathcal{D}^S_{(1)}$, $\mathcal{D}^S_{(2)}$, $\cdots$, also with N data points. The procedure is simply to draw N data points at a time *with replacement* from the set $\mathcal{D}^S_{(0)}$. Because of the replacement, you do not simply get back your original data set each time. You get sets in which a random fraction of the original points, typically $\sim 1/e \approx 37\%$, are replaced by *duplicated* original points. Now, exactly as in the previous discussion, you subject these data sets to the same estimation procedure as was performed on the actual data, giving a set of simulated measured parameters $a^S_{(1)}, a^S_{(2)}, \cdots$. These will be distributed around $a_{(0)}$ in close to the same way that $a_{(0)}$ is distributed around a_{true}.

Sounds like getting something for nothing, doesn't it? In fact, it has taken more than a decade for the bootstrap method to become accepted by statisticians. By now, however, enough theorems have been proved to render the bootstrap reputable. The basic idea behind the bootstrap is that the actual data set, viewed as a probability distribution consisting of delta functions at the measured values, is in most cases the best — or only — available estimator of the underlying probability distribution. It takes courage, but one can often simply use *that* distribution as the basis for Monte Carlo simulations.

Watch out for cases where the bootstrap's "iid" assumption is violated. For example, if you have made measurements at evenly spaced intervals of some control variable, then you can *usually* get away with pretending that these are "iid," uniformly distributed over the measured range. However, some estimators of a (e.g., ones involving Fourier methods) might be particularly sensitive to all the points on a grid being present. In that case, the bootstrap is going to give a wrong distribution. Also watch out for estimators that look at anything like small-scale clumpiness within the N data points, or estimators that sort the data and look at sequential differences. Obviously the bootstrap will fail on these, too. (The theorems justifying the method are still true, but some of their technical assumptions are violated by these examples.)

For a large class of problems, however, the bootstrap does yield easy, *very quick*, Monte Carlo estimates of the errors in an estimated parameter set.

20.1.3 Confidence Limits

Rather than present all details of the probability distribution of errors in parameter estimation, it is common practice to summarize the distribution in the form of *confidence limits*. The full probability distribution is a function defined on the M-dimensional space of parameters a. A *confidence region* (or *confidence interval*) is just a region of that M-dimensional space (hopefully a small region) that contains a certain (hopefully large) percentage of the total probability distribution. You point to a confidence region and say, e.g., "there is a 99 percent chance that the true parameter values fall within this region around the measured value."

It is worth emphasizing that you, the experimenter, get to pick both the *confidence level* (99 percent in the above example), and the shape of the confidence region. The only requirement is that your region does include the stated percentage of probability. Certain percentages are, however, customary in scientific usage: 68.3 percent (the lowest confidence worthy of quoting), 90 percent, 95.4 percent, 99 percent, and 99.73 percent. Higher confidence levels are conventionally "ninety-nine point nine ... nine." As for shape, obviously you want a region that is compact and reasonably centered on your measurement $a_{(0)}$, since the whole purpose of a confidence limit is to inspire confidence in that measured value. In one dimension, the convention is to use a line segment centered on the measured value; in higher dimensions, ellipses or ellipsoids are most frequently used.

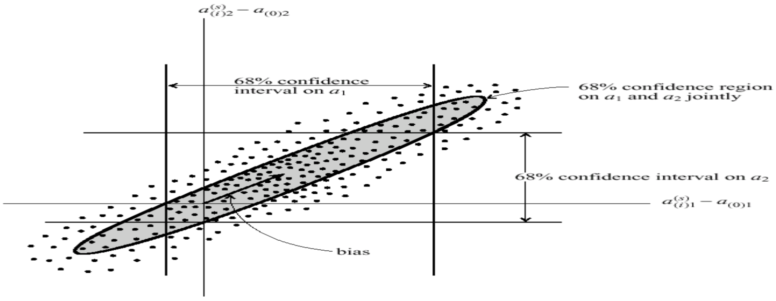

$a^{(s)}_{(i)2} - a_{(0)2}$
68% confidence interval on a_1
68% confidence region on a_1 and a_2 jointly
68% confidence interval on a_2
$a^{(s)}_{(i)1} - a_{(0)1}$
bias

Figure 20.3: Confidence intervals in 1 and 2 dimensions. The same fraction of measured points (here 68%) lies (i) between the two vertical lines, (ii) between the two horizontal lines, (iii) within the ellipse.

You might suspect, correctly, that the numbers 68.3 percent, 95.4 percent, and 99.73 percent, and the use of ellipsoids, have some connection with a normal distribution. That is true historically, but not always relevant nowadays. In general, the probability distribution of the parameters will not be normal, and the above numbers, used as levels of confidence, are purely matters of convention.

Figure 20.3 sketches a possible probability distribution for the case $M = 2$. Shown are three different confidence regions that might usefully be given, all at the same confidence level. The two vertical lines enclose a band (horizontal interval) which represents the 68 percent confidence interval for the variable a_1 without regard to the value of a_2. Similarly the horizontal lines enclose a 68 percent confidence interval for a_2. The ellipse shows a 68 percent confidence interval for a_1 and a_2 jointly. Notice that to enclose the same probability as the two bands, the ellipse must necessarily extend outside of both of them (a point we will return to below).

20.1.4 Constant Chi-Square Boundaries as Confidence Limits

When the method used to estimate the parameters $a_{(0)}$ is chi-square minimization, as in the previous sections of this chapter, then there is a natural choice for the shape of confidence intervals, whose use is almost universal. For the observed data set $\mathcal{D}_{(0)}$, the value of χ^2 is a minimum at $a_{(0)}$. Call this minimum value χ^2_{min}. If the vector a of parameter values is perturbed away from $a_{(0)}$, then χ^2 increases. The region within which χ^2 increases by no more than a set amount $\Delta\chi^2$ defines some M-dimensional confidence region around $a_{(0)}$. If $\Delta\chi^2$ is set to be a large number, this will be a big region; if it is small, it will be small. Somewhere in between there will be choices of $\Delta\chi^2$ that cause the region to contain, variously, 68 percent, 90 percent, etc. of probability distribution for a's, as defined above. These regions are taken as the confidence regions for the parameters $a_{(0)}$.

Very frequently one is interested not in the full M-dimensional confidence region, but in individual confidence regions for some smaller number ν of parameters. For example, one might be interested in the confidence interval of each parameter taken separately (the bands in Figure 20.3), in which case $\nu = 1$. In that case, the natural

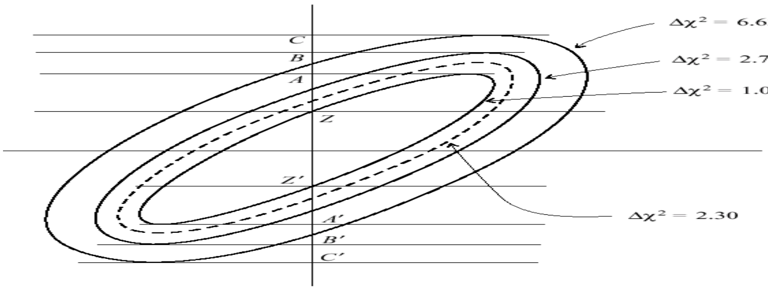


Figure 20.4: Confidence region ellipses corresponding to values of chi-square larger than the fitted minimum. The solid curves, with $\Delta\chi^2 = 1.00, 2.71, 6.63$ project onto one-dimensional intervals AA', BB', CC'. These intervals — not the ellipses themselves — contain 68.3%, 90%, and 99% of normally distributed data. The ellipse that contains 68.3% of normally distributed data is shown dashed, and has $\Delta\chi^2 = 2.30$. For additional numerical values, see accompanying table.

confidence regions in the ν-dimensional subspace of the M-dimensional parameter space are the *projections* of the M-dimensional regions defined by fixed $\Delta\chi^2$ into the ν-dimensional spaces of interest. In Figure 20.4, for the case $M = 2$, we show regions corresponding to several values of $\Delta\chi^2$. The one-dimensional confidence interval in a_2 corresponding to the region bounded by $\Delta\chi^2 = 1$ lies between the lines A and A'.

Notice that the projection of the higher-dimensional region on the lower-dimension space is used, not the intersection. The intersection would be the band between Z and Z'. It is *never* used. It is shown in the figure only for the purpose of making this cautionary point, that it should not be confused with the projection.

20.1.5 Probability Distribution of Parameters in the Normal Case

You may be wondering why we have, in this section up to now, made no connection at all with the error estimates that come out of the χ^2 fitting procedure, most notably the covariance matrix C_{ij}. The reason is this: χ^2 minimization is a useful means for estimating parameters even if the measurement errors are not normally distributed. While normally distributed errors are required if the χ^2 parameter estimate is to be a maximum likelihood estimator, one is often willing to give up that property in return for the relative convenience of the χ^2 procedure. Only in extreme cases, measurement error distributions with very large "tails," is χ^2 minimization abandoned in favor of more robust techniques.

However, the formal covariance matrix that comes out of a χ^2 minimization has a clear quantitative interpretation only if (or to the extent that) the measurement errors actually are normally distributed. In the case of *non*-normal errors, you are "allowed"

1. to fit for parameters by minimizing χ^2

2. to use a contour of constant $\Delta\chi^2$ as the boundary of your confidence region

3. to use Monte Carlo simulation or detailed analytic calculation in determining *which* contour $\Delta\chi^2$ is the correct one for your desired confidence level

4. to give the covariance matrix C_{ij} as the "formal covariance matrix of the fit."

You are not allowed to use formulas that we now give for the case of normal errors, which establish quantitative relationships among $\Delta\chi^2$, C_{ij}, and the confidence level.

Here are the key theorems that hold when (i) the measurement errors are normally distributed, and either (ii) the model is linear in its parameters or (iii) the sample size is large enough that the uncertainties in the fitted parameters a do not extend outside a region in which the model could be replaced by a suitable linearized model.

Theorem A. χ^2_{min} min is distributed as a chi-square distribution with $N - M$ degrees of freedom, where N is the number of data points and M is the number of fitted parameters. This is the basic theorem that lets you evaluate the goodness-of-fit of the model. We list it first to remind you that unless the goodness-of-fit is credible, the whole estimation of parameters is suspect.

Theorem B. If $a^S_{(j)}$ is drawn from the universe of simulated data sets with actual parameters $a_{(0)}$, then the probability distribution of $\delta a = a^S_{(j)} - a_{(0)}$ is the multivariate normal distribution

$$P(\delta a)da_1 \cdots da_M = \text{ constant } \times \exp\left(-\frac{1}{2}\delta a \cdot [\alpha] \cdot \delta a\right) da_1 \cdots da_M \qquad (20.1)$$

where $[\alpha]$ is the curvature matrix.

Theorem C. If $a^S_{(j)}$ is drawn from the universe of simulated data sets with actual parameters $a_{(0)}$, then the quantity $\Delta\chi^2 \equiv \chi^2\left(a_{(j)}\right) - \chi^2\left(a_{(0)}\right)$ is distributed as a chi-square distribution with M degrees of freedom. Here the χ^2's are all evaluated using the fixed (actual) data set $\mathcal{D}_{(0)}$. This theorem makes the connection between particular values of $\Delta\chi^2$ and the fraction of the probability distribution that they enclose as an M-dimensional region, i.e., the confidence level of the M-dimensional confidence region.

Theorem D. Suppose that $a^S_{(j)}$ is drawn from the universe of simulated data sets (as above), that its first ν components a_1, a_2, $\cdots$, a_ν are held fixed, and that its remaining $M - \nu$ components are varied so as to minimize χ^2. Call this minimum value χ^2_ν. Then $\Delta\chi^2_\nu \equiv \chi^2_\nu - \chi^2_{min}$ is distributed as a chi-square distribution with ν degrees of freedom. If you consult figure 20.4, you will see that this theorem connects the *projected* $\Delta\chi^2$ region with a confidence level. In the figure, a point that is held fixed in a_2 and allowed to vary in a_1 minimizing χ^2 will seek out the ellipse whose top or bottom edge is tangent to the line of constant a_2, and is therefore the line that projects it onto the smaller-dimensional space.

As a first example, let us consider the case $\nu = 1$, where we want to find the confidence interval of a single parameter, say a_1. Notice that the chi-square distribution with $\nu = 1$ degree of freedom is the same distribution as that of the square of a single normally distributed quantity. Thus $\Delta\chi^2_\nu < 1$ occurs 68.3 percent of the time ($1 - \sigma$ for the normal distribution), $\Delta\chi^2_\nu < 4$ occurs 95.4 percent of the time ($2 - \sigma$ for the normal distribution), $\Delta\chi^2_\nu < 9$ occurs 99.73 percent of the time ($3 - \sigma$ for the normal distribution), etc. In this manner you find the $\Delta\chi^2_\nu$ that corresponds

p	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$	$\nu = 5$	$\nu = 6$
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

Table 20.1: $\Delta\chi^2$ as a function of confidence level and degrees of freedom.

to your desired confidence level. (Additional values are given in the accompanying table.)

Let δa be a change in the parameters whose first component is arbitrary, δa_1, but the rest of whose components are chosen to minimize the $\Delta\chi^2$. Then Theorem D applies. The value of $\Delta\chi^2$ is given in general by

$$\Delta\chi^2 = \delta a \cdot [\alpha] \cdot \delta a \tag{20.2}$$

Since δa by hypothesis minimizes χ^2 in all but its first component, the second through M^{th} components of the normal equations continue to hold. Therefore, the solution is

$$\delta a = [\alpha]^{-1} \cdot \begin{pmatrix} c \\ 0 \\ \vdots \\ 0 \end{pmatrix} = [C]^{-1} \cdot \begin{pmatrix} c \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{20.3}$$

where c is one arbitrary constant that we get to adjust to make 20.2 give the desired left-hand value. Plugging 20.3 into 20.2 and using the fact that $[C]$ and $[\alpha]$ are inverse matrices of one another, we get

$$c = \frac{\delta a_1}{C_{11}} \text{ and } \Delta\chi^2_\nu = \frac{(\delta a_1)^2}{C_{11}} \tag{20.4}$$

or

$$\delta a_1 = \pm\sqrt{\Delta\chi^2_\nu}\sqrt{C_{11}} \tag{20.5}$$

At last! A relation between the confidence interval $\pm\delta a_1$ and the formal standard error $\sigma_1 \equiv \sqrt{C_{11}}$. Not unreasonably, we find that the 68 percent confidence interval is $\pm\sigma_1$, the 95 percent confidence interval is $\pm 2\sigma_1$, etc.

These considerations hold not just for the individual parameters a_i, but also for any linear combination of them: If

$$b \equiv \sum_{k=1}^{M} c_i a_i = c \cdot a \tag{20.6}$$

then the 68 percent confidence interval on b is

$$\delta b = \pm\sqrt{c \cdot [C] \cdot c} \tag{20.7}$$

However, these simple, normal-sounding numerical relationships do *not* hold in the case $\nu > 1$. In particular, $\Delta\chi^2 = 1$ is not the boundary, nor does it project onto the boundary, of a 68.3 percent confidence region when $\nu > 1$. If you want to calculate not confidence intervals in one parameter, but confidence ellipses in two parameters jointly, or ellipsoids in three, or higher, then you must follow the following prescription for implementing Theorems C and D above:

1. Let ν be the number of fitted parameters whose joint confidence region you wish to display, $\nu \leq M$. Call these parameters the "parameters of interest."

2. Let p be the confidence limit desired, e.g., $p = 0.68$ or $p = 0.95$.

3. Find Δ (i.e., $\Delta\chi^2$) such that the probability of a chi-square variable with ν degrees of freedom being less than Δ is p. For some useful values of p and ν, Δ is given in the table.

4. Take the $M \times M$ covariance matrix $[C] = [\alpha]^{-1}$ of the chi-square fit. Copy the intersection of the ν rows and columns corresponding to the parameters of interest into a $\nu \times \nu$ matrix denoted $[C_{proj}]$.

5. Invert the matrix $[C_{proj}]$. (In the one-dimensional case this was just taking the reciprocal of the element C_{11}.)

6. The equation for the elliptical boundary of your desired confidence region in the ν-dimensional subspace of interest is

$$\Delta = \delta a' \cdot [C_{proj}]^{-1} \cdot \delta a' \tag{20.8}$$

where $\delta a'$ is the ν-dimensional vector of parameters of interest.

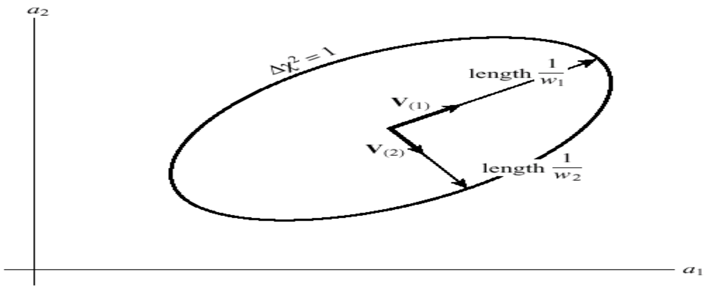


Figure 20.5: Relation of the confidence region ellipse $\Delta\chi^2 = 1$ to quantities computed by singular value decomposition. The vectors $V_{(i)}$ are unit vectors along the principal axes of the confidence region. The semi-axes have lengths equal to the reciprocal of the singular values w_i. If the axes are all scaled by some constant factor α, $\Delta\chi^2$ is scaled by the factor α^2.

If you are confused at this point, you may find it helpful to compare figure 20.4 and the accompanying table, considering the case $M = 2$ with $\nu = 1$ and $\nu = 2$. You should be able to verify the following statements: (i) The horizontal band between C and C' contains 99 percent of the probability distribution, so it is a confidence limit on a_2 alone at this level of confidence. (ii) Ditto the band between B and B' at the 90 percent confidence level. (iii) The dashed ellipse, labeled by $\Delta\chi^2 = 2.30$, contains 68.3 percent of the probability distribution, so it is a confidence region for a_1 and a_2 jointly, at this level of confidence.

20.2 Reading: The Vice of Gambling and the Virtue of Insurance by George Bernard Shaw

20.2.1 Introduction to the Reading

20.2.2 The Paper

Chapter 21

Correlation and Confidence

21.1 Contingency Table Analysis of Two Distributions

In this section, and the next two sections, we deal with *measures of association* for two distributions. The situation is this: Each data point has two or more different quantities associated with it, and we want to know whether knowledge of one quantity gives us any demonstrable advantage in predicting the value of another quantity. In many cases, one variable will be an "independent" or "control" variable, and another will be a "dependent" or "measured" variable. Then, we want to know if the latter variable *is* in fact dependent on or *associated* with the former variable. If it is, we want to have some quantitative measure of the strength of the association. One often hears this loosely stated as the question of whether two variables are *correlated* or *uncorrelated*, but we will reserve those terms for a particular kind of association (linear, or at least monotonic), as discussed below.

Notice that the different concepts of significance and strength appear: The association between two distributions may be very significant even if that association is weak - if the quantity of data is large enough.

It is useful to distinguish among some different kinds of variables, with different categories forming a loose hierarchy.

1. A variable is called *nominal* if its values are the members of some unordered set. For example, "state of residence" is a nominal variable that (in the U.S.) takes on one of 50 values; in astrophysics, "type of galaxy" is a nominal variable with the three values "spiral," "elliptical," and "irregular."

2. A variable is termed *ordinal* if its values are the members of a discrete, but ordered, set. Examples are: grade in school, planetary order from the Sun (Mercury = 1, Venus = 2, $\cdots$), number of offspring. There need not be any concept of "equal metric distance" between the values of an ordinal variable, only that they be intrinsically ordered.

3. We will call a variable *continuous* if its values are real numbers, as are times, distances, temperatures, etc. (Social scientists sometimes distinguish between *interval* and *ratio* continuous variables, but we do not find that distinction very compelling.)

Figure 21.1: Example of a contingency table for two nominal variables, here sex and color. The row and column marginals (totals) are shown. The variables are "nominal," i.e., the order in which their values are listed is arbitrary and does not affect the result of the contingency table analysis. If the ordering of values has some intrinsic meaning, then the variables are "ordinal" or "continuous," and correlation techniques can be utilized.

A continuous variable can always be made into an ordinal one by binning it into ranges. If we choose to ignore the ordering of the bins, then we can turn it a nominal variable. Nominal variables constitute the lowest type of the hierarchy, and therefore the most general. For example, a set of *several* continuous or ordinal variables can be turned, if crudely, into a single nominal variable, by coarsely binning each variable and then taking each distinct combination of bin assignments as a single nominal value. When multidimensional data are sparse, this is often the only sensible way to proceed.

The remainder of this section will deal with measures of association between *nominal* variables. For any pair of nominal variables, the data can be displayed as a *contingency table*, a table whose rows are labelled by the values of one nominal variable, whose columns are labelled by the values of the other nominal variable, and whose entries are nonnegative integers giving the number of observed events for each combination of row and column (see Figure 21.1). The analysis of association between nominal variables is thus called *contingency table analysis* or *crosstabulation analysis*. We will introduce two different approaches. The first approach, based on the chi-square statistic, does a good job of characterizing the significance of association, but is only so-so as a measure of the strength (principally because its numerical values have no very direct interpretations). The second approach, based on the information-theoretic concept of *entropy*, says nothing at all about the significance of association (use chi-square for that!), but is capable of very elegantly characterizing the strength of an association already known to be significant.

21.1.1 Measures of Association Based on Chi-Square

Some notation first: Let N_{ij} denote the number of events that occur with the first variable x taking on its i^{th} value, and the second variable y taking on its j^{th} value. Let N denote the total number of events, the sum of all the N_{ij}'s. Let $N_{i.} \leq$ denote the number of events for which the first variable x takes on its i^{th} value regardless of the value of y; $N_{.j}$ is the number of events with the j^{th} value of y regardless of x. So we have

$$N_{i.} = \sum_j N_{ij}, \qquad N_{.j} = \sum_i N_{ij}, \qquad N = \sum_i N_{i.} = \sum_j N_{.j} \tag{21.1}$$

$N_{.j}$ and $N_{i.}$ are sometimes called the *row and column totals* or *marginals*, but we will use these terms cautiously since we can never keep straight which are the rows and which are the columns!

The null hypothesis is that the two variables x and y have no association. In this case, the probability of a particular value of x given a particular value of y should be the same as the probability of that value of x regardless of y. Therefore, in the null hypothesis, the expected number for any N_{ij}, which we will denote n_{ij}, can be calculated from only the row and column totals,

$$\frac{n_{ij}}{N_{.j}} = \frac{N_{i.}}{N} \text{ which implies } n_{ij} = \frac{N_{i.}N_{.j}}{N} \tag{21.2}$$

Notice that if a column or row total is zero, then the expected number for all the entries in that column or row is also zero; in that case, the never-occurring bin of x or y should simply be removed from the analysis.

The chi-square statistic is now given by equation 15.8, which, in the present case, is summed over all entries in the table,

$$\chi^2 = \sum_{i,j} \frac{\left(N_{ij} - n_{ij}\right)^2}{n_{ij}} \tag{21.3}$$

The number of degrees of freedom is equal to the number of entries in the table (product of its row size and column size) minus the number of constraints that have arisen from our use of the data themselves to determine the n_{ij}. Each row total and column total is a constraint, except that this overcounts by one, since the total of the column totals and the total of the row totals both equal N, the total number of data points. Therefore, if the table is of size I by J, the number of degrees of freedom is $IJ - I - J + 1$. Equation 21.3, along with the chi-square probability function, now give the significance of an association between the variables x and y.

Suppose there is a significant association. How do we quantify its strength, so that (e.g.) we can compare the strength of one association with another? The idea here is to find some reparametrization of χ^2 which maps it into some convenient interval, like 0 to 1, where the result is not dependent on the quantity of data that we happen to sample, but rather depends only on the underlying population from which the data were drawn. There are several different ways of doing this. Two of the more common are called *Cramer's V* and the *contingency coefficient C*.

The formula for Cramer's V is

$$V = \sqrt{\frac{\chi^2}{N \min(I-1, J-1)}} \tag{21.4}$$

where I and J are again the numbers of rows and columns, and N is the total number of events. Cramer's V has the pleasant property that it lies between zero and one inclusive, equals zero when there is no association, and equals one only when the association is perfect: All the events in any row lie in one unique column, and vice versa. (In chess parlance, no two rooks, placed on a nonzero table entry, can capture each other.)

In the case of $I = J = 2$, Cramer's V is also referred to as the *phi statistic.*

The contingency coefficient C is defined as

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{21.5}$$

It also lies between zero and one, but (as is apparent from the formula) it can never achieve the upper limit. While it can be used to compare the strength of association of two tables with the same I and J, its upper limit depends on I and J. Therefore it can never be used to compare tables of different sizes.

The trouble with both Cramer's V and the contingency coefficient C is that, when they take on values in between their extremes, there is no very direct interpretation of what that value means. For example, you are in Las Vegas, and a friend tells you that there is a small, but significant, association between the color of a croupier's eyes and the occurrence of red and black on his roulette wheel. Cramer's V is about 0.028, your friend tells you. You know what the usual odds against you are (because of the green zero and double zero on the wheel). Is this association sufficient for you to make money? Don't ask us!

21.1.2 Measures of Association Based on Entropy

Consider the game of "twenty questions," where by repeated yes/no questions you try to eliminate all except one correct possibility for an unknown object. Better yet, consider a generalization of the game, where you are allowed to ask multiple-choice questions as well as binary (yes/no) ones. The categories in your multiple-choice questions are supposed to be mutually exclusive and exhaustive (as are "yes" and "no").

The value to you of an answer increases with the number of possibilities that it eliminates. More specifically, an answer that eliminates all except a fraction p of the remaining possibilities can be assigned a value $-\ln p$ (a positive number, since $p < 1$). The purpose of the logarithm is to make the value additive, since (e.g.) one question that eliminates all but 1/6 of the possibilities is considered as good as two questions that, in sequence, reduce the number by factors 1/2 and 1/3.

So that is the value of an answer; but what is the value of a question? If there are I possible answers to the question ($i = 1, 2, \cdots, I$) and the fraction of possibilities consistent with the i^{th} answer is p_i (with the sum of the p_i's equal to one), then the value of the question is the expectation value of the value of the answer, denoted H

$$H = -\sum_{i=1}^{I} p_i \ln p_i \tag{21.6}$$

In evaluating 21.6, note that

$$\lim_{p \to 0} p \ln p = 0 \tag{21.7}$$

The value H lies between 0 and $\ln I$. It is zero only when one of the p_i's is one, all the others zero: In this case, the question is valueless, since its answer is preordained. H takes on its maximum value when all the p_i's are equal, in which case the question is sure to eliminate all but a fraction $1/I$ of the remaining possibilities.

The value H is conventionally termed the *entropy* of the distribution given by the p_i's, a terminology borrowed from statistical physics.

So far we have said nothing about the association of two variables; but suppose we are deciding what question to ask next in the game and have to choose between two candidates, or possibly want to ask both in one order or another. Suppose that one question, x, has I possible answers, labeled by i, and that the other question, y, has J possible answers, labeled by j. Then the possible outcomes of asking both questions form a contingency table whose entries N_{ij}, when normalized by dividing by the total number of remaining possibilities N, give all the information about the p's. In particular, we can make contact with the notation 21.1 by identifying

$$\begin{aligned} p_{ij} &= \frac{N_{ij}}{N} && (21.8)\\ p_{i\cdot} &= \frac{N_{i\cdot}}{N} \text{ (outcomes of question } x \text{ alone)} && (21.9)\\ p_{\cdot j} &= \frac{N_{\cdot j}}{N} \text{ (outcomes of question } y \text{ alone)} && (21.10) \end{aligned}$$

The entropies of the questions x and y are, respectively,

$$H(x) = -\sum_i p_{i\cdot} \ln p_{i\cdot}, \qquad H(y) = -\sum_j p_{\cdot j} \ln p_{\cdot j} \tag{21.11}$$

The entropy of the two questions together is

$$H(x,y) = -\sum_{i,j} p_{ij} \ln p_{ij} \tag{21.12}$$

Now what is the entropy of the question y *given* x (that is, if x is asked first)? It is the expectation value over the answers to x of the entropy of the restricted y distribution that lies in a single column of the contingency table (corresponding to the x answer):

$$H(y|x) = -\sum_i p_{i\cdot} \sum_j \frac{p_{ij}}{p_{i\cdot}} \ln \frac{p_{ij}}{p_{i\cdot}} = -\sum_{i,j} p_{ij} \ln \frac{p_{ij}}{p_{i\cdot}} \tag{21.13}$$

Correspondingly, the entropy of x given y is

$$H(x|y) = -\sum_j p_{\cdot j} \sum_i \frac{p_{ij}}{p_{\cdot j}} \ln \frac{p_{ij}}{p_{\cdot j}} = -\sum_{i,j} p_{ij} \ln \frac{p_{ij}}{p_{\cdot j}} \tag{21.14}$$

We can readily prove that the entropy of y given x is never more than the entropy of y alone, i.e., that asking x first can only reduce the usefulness of asking y (in which case the two variables are *associated*!):

$$\begin{aligned} H(y|x) - H(y) &= -\sum_{i,j} p_{ij} \ln \frac{p_{ij}/p_{i\cdot}}{p_{\cdot j}} && (21.15)\\ &= \sum_{i,j} p_{ij} \ln \frac{p_{\cdot j}/p_{i\cdot}}{p_{ij}} && (21.16)\\ &\leq \sum_{i,j} p_{ij} \left(\frac{p_{\cdot j}/p_{i\cdot}}{p_{ij}} - 1 \right) && (21.17)\\ &= \sum_{i,j} p_{i\cdot} p_{\cdot j} - \sum_{i,j} p_{ij} && (21.18)\\ &= 0 && (21.19) \end{aligned}$$

where the inequality follows from the fact

$$\ln w \leq w - 1 \tag{21.20}$$

We now have everything we need to define a measure of the "dependency" of y on x, that is to say a measure of association. This measure is sometimes called the *uncertainty coefficient* of y. We will denote it as $U(y|x)$,

$$U(y|x) \equiv \frac{H(y) - H(y|x)}{H(y)} \tag{21.21}$$

This measure lies between zero and one, with the value 0 indicating that x and y have no association, the value 1 indicating that knowledge of x completely predicts y. For in-between values, $U(y|x)$ gives the fraction of y's entropy $H(y)$) that is lost if x is already known (i.e., that is redundant with the information in x). In our game of "twenty questions," $U(y|x)$ is the fractional loss in the utility of question y if question x is to be asked first.

If we wish to view x as the dependent variable, y as the independent one, then interchanging x and y we can of course define the dependency of x on y,

$$U(x|y) \equiv \frac{H(x) - H(x|y)}{H(x)} \tag{21.22}$$

If we want to treat x and y symmetrically, then the useful combination turns out to be

$$U(x|y) \equiv 2\left[\frac{H(x) + H(y) - H(x,y)}{H(x) + H(y)}\right] \tag{21.23}$$

If the two variables are completely independent, then $H(x,y) = H(x) + H(y)$, so 21.23 vanishes. If the two variables are completely dependent, then $H(x) = H(y) = H(x,y)$, so 21.22 equals unity. In fact, you can use the identities

$$H(x,y) = H(x) + H(y|x) = H(y) + H(x|y) \tag{21.24}$$

to show that

$$U(x,y) = \frac{H(x)U(x|y) + H(y)U(y|x)}{H(x) + H(y)} \tag{21.25}$$

i.e., that the symmetrical measure is just a weighted average of the two asymmetrical measures 21.21 and 21.22, weighted by the entropy of each variable separately.

21.2 Linear Correlation

We next turn to measures of association between variables that are ordinal or continuous, rather than nominal. Most widely used is the *linear correlation coefficient.* For pairs of quantities (x_i, y_i), $i = 1, 2, \cdots, N$, the linear correlation coefficient r (also called the *product-moment correlation coefficient*, or *Pearson's r*) is given by the formula

$$r = \frac{\sum\limits_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_i (x_i - \overline{x})^2}\sqrt{\sum\limits_i (y_i - \overline{y})^2}} \tag{21.26}$$

where, as usual, $\overline{x}$ is the mean of the x_i's, $\overline{y}$ is the mean of the y_i's. The value of r lies between -1 and 1, inclusive. It takes on a value of 1, termed "complete positive correlation," when the data points lie on a perfect straight line with positive slope, with x and y increasing together. The value 1 holds independent of the magnitude of the slope. If the data points lie on a perfect straight line with negative slope, y decreasing as x increases, then r has the value -1; this is called "complete negative correlation." A value of r near zero indicates that the variables x and y are *uncorrelated.*

When a correlation is known to be significant, r is one conventional way of summarizing its strength. In fact, the value of r can be translated into a statement about what residuals (root mean square deviations) are to be expected if the data are fitted to a straight line by the least-squares method. Unfortunately, r is a rather poor statistic for deciding *whether* an observed correlation is statistically significant, and/or whether one observed correlation is significantly stronger than another. The reason is that r is ignorant of the individual distributions of x and y, so there is no universal way to compute its distribution in the case of the null hypothesis.

About the only general statement that can be made is this: If the null hypothesis is that x and y are uncorrelated, and if the distributions for x and y each have enough convergent moments ("tails" die off sufficiently rapidly), and if N is large (typically > 500), then r is distributed approximately normally, with a mean of zero and a standard deviation of $1/\sqrt{N}$. In that case, the (double-sided) significance of the correlation, that is, the probability that $|r|$ should be larger than its observed value in the null hypothesis, is

$$erfc\left(\frac{|r|\sqrt{N}}{\sqrt{2}}\right) \tag{21.27}$$

where $erfc(x)$ is the complementary error function. A small value of 21.27 indicates that the two distributions are significantly correlated. (See below for a more accurate test.)

Most statistics books try to go beyond 21.27 and give additional statistical tests that can be made using r. In almost all cases, however, these tests are valid only for a very special class of hypotheses, namely that the distributions of x and y jointly form a *binormal* or *two-dimensional Gaussian* distribution around their mean values, with joint probability density

$$p(x,y)dxdy = \text{const.} \times \exp\left[-\frac{1}{2}\left(a_{11}x^2 - 2a_{12}xy + a_{22}y^2\right)\right] dxdy \tag{21.28}$$

where a_{11}, a_{12} and a_{22} are arbitrary constants. For this distribution r has the value

$$r = -\frac{a_{12}}{\sqrt{a_{11}a_{22}}} \tag{21.29}$$

There are occasions when 21.28 may be known to be a good model of the data. There may be other occasions when we are willing to take 21.28 as at least a rough and ready guess, since many two-dimensional distributions do resemble a binormal distribution, at least not too far out on their tails. In either situation, we can use 21.28 to go beyond 21.27 in any of several directions:

First, we can allow for the possibility that the number N of data points is not large. Here, it turns out that the statistic

$$t = r\sqrt{\frac{N-2}{1-r^2}} \tag{21.30}$$

is distributed in the null case (of no correlation) like Student's t-distribution with $\nu = N-2$ degrees of freedom, whose two-sided significance level is given by $1-A(t|\nu)$. As N becomes large, this significance and 21.27 become asymptotically the same, so that one never does worse by using 21.30, even if the binormal assumption is not well substantiated.

Second, when N is only moderately large ($\geq$ 10), we can compare whether the difference of two significantly nonzero r's, e.g., from different experiments, is itself significant. In other words, we can quantify whether a change in some control variable significantly alters an existing correlation between two other variables. This is done by using *Fisher's z-transformation* to associate each measured r with a corresponding z,

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{21.31}$$

Then, each z is approximately normally distributed with a mean value

$$\overline{z} = \frac{1}{2}\left(\ln\left(\frac{1+r_{true}}{1-r_{true}}\right) + \frac{r_{true}}{N-1}\right) \tag{21.32}$$

where r_{true} is the actual or population value of the correlation coefficient, and with a standard deviation

$$\sigma(z) \approx \frac{1}{\sqrt{N-3}} \tag{21.33}$$

Equations 21.32 and 21.33, when they are valid, give several useful statistical tests. For example, the significance level at which a measured value of r differs from some hypothesized value r_{true} is given by

$$erfc\left(\frac{|z-\overline{z}|\sqrt{N-3}}{\sqrt{2}}\right) \tag{21.34}$$

where z and $\overline{z}$ are given by 21.31 and 21.32, with small values of 21.34 indicating a significant difference. (Setting $\overline{z} = 0$ makes expression 21.34 a more accurate replacement for expression 21.27 above.) Similarly, the significance of a difference between two measured correlation coefficients r_1 and r_2 is

$$erfc\left(\frac{|z_1 - z_2|}{\sqrt{2}\sqrt{\frac{1}{N_1-3}+\frac{1}{N_2-3}}}\right) \tag{21.35}$$

where z_1 and z_2 are obtained from r_1 and r_2 using 21.31, and where N_1 and N_2 are, respectively, the number of data points in the measurement of r_1 and r_2.

All of the significances above are two-sided. If you wish to disprove the null hypothesis in favor of a one-sided hypothesis, such as that $r_1 > r_2$ (where the sense of the inequality was decided *a priori*), then (i) if your measured r_1 and r_2 have the *wrong* sense, you have failed to demonstrate your one-sided hypothesis, but (ii) if they have the right ordering, you can multiply the significances given above by 0.5, which makes them more significant.

But keep in mind: These interpretations of the r statistic can be completely meaningless if the joint probability distribution of your variables x and y is too different from a binormal distribution.

21.3 Nonparametric or Rank Correlation

It is precisely the uncertainty in interpreting the significance of the linear correlation coefficient r that leads us to the important concepts of *nonparametric* or *rank correlation.* As before, we are given N pairs of measurements (x_i, y_i), $i = 1, 2, \cdots, N$. Before, difficulties arose because we did not necessarily know the probability distribution function from which the x_i's or y_i's were drawn.

The key concept of nonparametric correlation is this: If we replace the value of each x_i by the value of its *rank* among all the other x_i's in the sample, that is, $1, 2, \cdots, N$, then the resulting list of numbers will be drawn from a perfectly known distribution function, namely uniformly from the integers between 1 and N , inclusive. Better than uniformly, in fact, since if the x_i's are all distinct, then each integer will occur precisely once. If some of the x_i's have identical values, it is conventional to assign to all these "ties" the mean of the ranks that they would have had if their values had been slightly different. This *midrank* will sometimes be an integer, sometimes a half-integer. In all cases the sum of all assigned ranks will be the same as the sum of the integers from 1 to N, namely $N(N + 1)/2$).

Of course we do exactly the same procedure for the y_i's, replacing each value by its rank among the other y_i's in the sample.

Now we are free to invent statistics for detecting correlation between uniform sets of integers between 1 and N, keeping in mind the possibility of ties in the ranks. There is, of course, some loss of information in replacing the original numbers by ranks. We could construct some rather artificial examples where a correlation could be detected parametrically (e.g., in the linear correlation coefficient r), but could not be detected nonparametrically. Such examples are very rare in real life, however, and the slight loss of information in ranking is a small price to pay for a very major advantage: When a correlation is demonstrated to be present nonparametrically, then it is really there! (That is, to a certainty level that depends on the significance chosen.) Nonparametric correlation is more robust than linear correlation, more resistant to unplanned defects in the data, in the same sort of sense that the median is more robust than the mean.

As always in statistics, some particular choices of a statistic have already been invented for us and consecrated, if not beatified, by popular use. We will discuss two, the *Spearman rank-order correlation coefficient* (r_s), and *Kendall's tau* (τ).

21.3.1 Spearman Rank-Order Correlation Coefficient

Let R_i be the rank of x_i among the other x's, S_i be the rank of y_i among the other y's, ties being assigned the appropriate midrank as described above. Then the rank-order correlation coefficient is defined to be the linear correlation coefficient of the ranks, namely,

$$r_s = \frac{\sum\limits_i \left(R_i - \overline{R}\right)\left(S_i - \overline{S}\right)}{\sqrt{\sum\limits_i \left(R_i - \overline{R}\right)^2}\sqrt{\sum\limits_i \left(S_i - \overline{S}\right)^2}} \tag{21.36}$$

The significance of a nonzero value of r_s is tested by computing

$$t = r_s\sqrt{\frac{N-2}{1-r_s^2}} \tag{21.37}$$

which is distributed approximately as Student's distribution with $N-2$ degrees of freedom. A key point is that this approximation does not depend on the original distribution of the x's and y's; it is always the same approximation, and always pretty good.

It turns out that r_s is closely related to another conventional measure of nonparametric correlation, the so-called *sum squared difference of ranks*, defined as

$$D = \sum i = 1^N \left(R_i - S_i\right)^2 \tag{21.38}$$

(This D is sometimes denoted D^{**}, where the asterisks are used to indicate that ties are treated by midranking.)

When there are no ties in the data, then the exact relation between D and r_s is

$$r_s = 1 - \frac{6D}{N^3 - N} \tag{21.39}$$

When there are ties, then the exact relation is slightly more complicated: Let f_k be the number of ties in the k^{th} group of ties among the R_i's, and let g_m be the number of ties in the m^{th} group of ties among the S_i's. Then it turns out that

$$r_s = \frac{1 - \frac{6}{N^3-N}\left[D + \frac{1}{12}\sum\limits_k \left(f_k^3 - f_k\right) + \frac{1}{12}\sum\limits_m \left(g_m^3 - g_m\right)\right]}{\left[1 - \frac{\sum\limits_k \left(f_k^3 - f_k\right)}{N^3-N}\right]^{1/2}\left[1 - \frac{\sum\limits_m \left(g_m^3 - g_m\right)}{N^3-N}\right]^{1/2}} \tag{21.40}$$

holds exactly. Notice that if all the f_k's and all the g_m's are equal to one, meaning that there are no ties, then equation 21.40 reduces to equation 21.39.

In 21.37 we gave a t-statistic that tests the significance of a nonzero r_s. It is also possible to test the significance of D directly. The expectation value of D in the null hypothesis of uncorrelated data sets is

$$\overline{D} = \frac{1}{6}\left(N^3 - N\right) - \frac{1}{12}\sum_k \left(f_k^3 - f_k\right) - \frac{1}{12}\sum_m \left(g_m^3 - g_m\right) \tag{21.41}$$

its variance is

$$Var(D) = \frac{(N-1)N^2(N+1)^2}{36}\left[1 - \frac{\sum\limits_k \left(f_k^3 - f_k\right)}{N^3 - N}\right]\left[1 - \frac{\sum\limits_m \left(g_m^3 - g_m\right)}{N^3 - N}\right] \tag{21.42}$$

and it is approximately normally distributed, so that the significance level is a complementary error function (cf. equation 21.27). Of course, 21.37 and 21.42 are not independent tests, but simply variants of the same test.

21.3.2 Kendall's Tau

Kendall's τ is even more nonparametric than Spearman's r_s or D. Instead of using the numerical difference of ranks, it uses only the relative ordering of ranks: higher in rank, lower in rank, or the same in rank. But in that case we don't even have to rank the data! Ranks will be higher, lower, or the same if and only if the values are larger, smaller, or equal, respectively. On balance, we prefer r_s as being the more straightforward nonparametric test, but both statistics are in general use. In fact, τ and τ are very strongly correlated and, in most applications, are effectively the same test.

To define τ, we start with the N data points (x_i, y_i). Now consider all $\frac{1}{2}N(N-1)$ *pairs* of data points, where a data point cannot be paired with itself, and where the points in either order count as one pair. We call a pair *concordant* if the relative ordering of the ranks of the two x's (or for that matter the two x's themselves) is the same as the relative ordering of the ranks of the two y's (or for that matter the two y's themselves). We call a pair *discordant* if the relative ordering of the ranks of the two x's is opposite from the relative ordering of the ranks of the two y's. If there is a tie in either the ranks of the two x's or the ranks of the two y's, then we don't call the pair either concordant or discordant. If the tie is in the x's, we will call the pair an "extra y pair." If the tie is in the y's, we will call the pair an "extra x pair." If the tie is in both the x's and the y's, we don't call the pair anything at all. Are you still with us?

Kendall's τ is now the following simple combination of these various counts:

$$\tau = \frac{\text{concordant} - \text{discordant}}{\sqrt{\text{concordant} + \text{discordant} + \text{extra-}y}\sqrt{\text{concordant} + \text{discordant} + \text{extra-}x}} \tag{21.43}$$

You can easily convince yourself that this must lie between 1 and -1, and that it takes on the extreme values only for complete rank agreement or complete rank reversal, respectively.

More important, Kendall has worked out, from the combinatorics, the approximate distribution of τ in the null hypothesis of no association between x and y. In this case τ is approximately normally distributed, with zero expectation value and a variance of

$$Var(\tau) = \frac{4N+10}{9N(N-1)} \tag{21.44}$$

Sometimes it happens that there are only a few possible values each for x and y. In that case, the data can be recorded as a contingency table that gives the number of data points for each contingency of x and y.

Spearman's rank-order correlation coefficient is not a very natural statistic under these circumstances, since it assigns to each x and y bin a not-very-meaningful midrank value and then totals up vast numbers of identical rank differences. Kendall's tau, on the other hand, with its simple counting, remains quite natural.

Note that Kendall's tau can be applied only to contingency tables where both variables are *ordinal*, i.e., well-ordered, and that it looks specifically for monotonic correlations, not for arbitrary associations. These two properties make it less general than the methods of contingency table analysis, which applied to *nominal*, i.e., unordered, variables and arbitrary associations.

21.4 Reading: The Probability of Induction by Charles Sanders Peirce

21.4.1 Introduction to the Reading

21.4.2 The Paper

Part III

Markov Processes

Chapter 22

Examples of Discrete Processes

22.1 The Population of Bremen

22.1.1 The Mathematical Model

Suppose that the Bremen population satisfies the statement that each year one tenth of the people living outside Bremen move into the state and two tenths of Bremenians move outside the state. We use two variables to denote this state of affairs, namely x_i and y_i to denote the population inside and outside Bremen during year i respectively. Thus we get the coupled recurrence relations

$$\begin{aligned} x_{i+1} &= 0.8x_i + 0.1y_i && (22.1) \\ y_{i+1} &= 0.2x_i + 0.9y_i && (22.2) \end{aligned}$$

Suppose further that initially 50,000 people live both inside and outside the state, i.e. $x_0 = y_0 = 50,000$. We can thus calculate the population in successive years using the above equations, the results are shown in table 22.1.

22.1.2 Solving for the Steady State: Iteration

As time passes, we expect that the population will settle down to a constant number. Of course, people will still move in and out but the number of people will remain constant over time. We observe from the table that in the beginning there was a large change in population but after some time the change gets less and less. The eventual equilibrium population is denoted by (x_∞, y_∞). The recurrence relations are a very simple example of a Markov process and the equilibrium population is called the *steady state response* of the process. It must be noted that a steady state response does not always exist but in this case, it does. Clearly,

$$(x_\infty, y_\infty) = \left(\lim_{i\to\infty} x_i, \lim_{i\to\infty} y_i\right) \qquad (22.3)$$

It is important also to note that the steady state response, if it exists, is independent of the initial state of the system as is illustrated in table 22.2 where the initial population distribution is 10,000 versus 90,000 but we reach the same distribution.

From the two initial conditions we observe that the steady state is approached from opposite sides. This can be used to compute upper and lower bounds on the

Year	Inside	Outside	Total
0	50,000	50,000	100,000
1	45,000	55,000	100,000
2	41,500	58,500	100,000
3	39,050	60,950	100,000
4	37,335	62,665	100,000
5	36,135	63,866	100,000
6	35,294	64,706	100,000
7	34,706	65,294	100,000
8	34,294	65,706	100,000
9	34,006	65,994	100,000
10	33,804	66,196	100,000

Table 22.1: The population of Bremen as a function of time with an even initial distribution.

Year	Inside	Outside	Total
0	10,000	90,000	100,000
1	17,000	83,000	100,000
2	21,900	78,100	100,000
3	25,330	74,670	100,000
4	27,731	72,269	100,000
5	29,412	70,588	100,000
6	30,588	69,412	100,000
7	31,412	68,588	100,000
8	31,988	68,012	100,000
9	32,392	67,608	100,000
10	32,674	67,326	100,000

Table 22.2: The population of Bremen as a function of time with a highly uneven initial distribution.

steady state and by iteration the steady state may be found to any accuracy desired. This requires many iterations however. It is generally possible to solve for the steady state response without this calculational aid even though it is more illustrative.

22.1.3 Solving for the Steady State: Transition Matrix

To solve the process, we construct a transition matrix by simply writing our recurrence in matrix form.

$$\begin{pmatrix} x_{i+1} \\ y_{i+1} \end{pmatrix} = \begin{pmatrix} 0.8 & 0.1 \\ 0.2 & 0.9 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} \tag{22.4}$$

The transition matrix (the one with the probabilistic coefficients) always has non-negative entries and its columns sum to one. The process converges to the steady state response at a geometric rate which is given by the sum of the diagonal entries (the trace) minus one. Here the rate is $r = 0.8 + 0.9 - 1 = 0.7$. We determine the system state after i iterations simply by

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} = A^i \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \tag{22.5}$$

where A is the transition matrix. The steady state response is then

$$\begin{pmatrix} x_\infty \\ y_\infty \end{pmatrix} = \lim_{i \to \infty} A^i \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \tag{22.6}$$

We can verify by actual calculation that this state is achieved after 27 iterations of our Bremen example (with equal initial population inside and outside) after which 33,334 people live inside the state. We can use the rate of convergence to predict the number of iterations needed achieve the steady state. This is simply done here

$$100,000(0.7)^i \leq 1 \tag{22.7}$$

in which we have multiplied the total population by the change in the i^{th} year and we requested that this is less than or equal to the accuracy we want (less than or equal to one person difference). This is trivially solved by logarithms and gives $i \geq 32$. Thus we should converge in 32 iterations. In our example we converged in 27 iterations because we chose an equally distributed initial population. This estimate applies for the worst case possible. In our situation the worst case scenario (in terms of the number of iterations to the steady state) is the one in which all the people live either inside or outside the state. For an actual case, the number of iterations will be less than or equal to the one calculated by the above method. Thus to compute the steady state response, we solve this problem and then take the appropriate matrix power of the transition matrix and multiply this by the initial state.

We must note that the steady state response does not depend on the initial state except in two cases. First, if the total population is zero, there is no convergence at all as the initial distribution can not change. Second, if the results are rounded (as they were above to the nearest whole person) we may get results differing by one unit depending on the initial distribution. If we start with the uneven distribution of 10,000 people living inside Bremen, we converge to 33,333 people living inside whereas if we start with 50,000 living inside we converge to 33,334 people living inside. This is a feature of rounding. The true steady state response is in general a real number. If a higher accuracy is required, the above method of calculating the worst case number of iterations will yield a larger number.

22.1.4 Solving for the Steady State: Linear Algebra

We may obtain the steady state another way. From the initial equation note,

$$\begin{aligned} x_{i+1} &= 0.8x_i + 0.1y_i && (22.8)\\ x_\infty &= 0.8x_\infty + 0.1y_\infty && (22.9)\\ 2x_\infty &= y_\infty && (22.10) \end{aligned}$$

We would have obtained the same result from the other equation. Also note that since $x_0+y_0 = 100,000$, we must have $x_\infty+y_\infty = 100,000$ and so $x_\infty \approx 33,333$. This procedure is very simple indeed but only for two variables. If the process becomes more complex, this is not so helpful.

22.1.5 Solving for the Steady State: Eigenvectors

The last method is the most complicated but also the best because once understood is very fast to apply. We find the eigenvalues of the transition matrix to be 0.7 and 1. The value 1 will always be an eigenvalue and the eigenvector it corresponds to is the steady state response. Another eigenvalue will be the convergence rate to the steady state response, in this case 0.7. We will give a slightly different method from above to get the eigenvector of interest to illustrate this route.

For each eigenvalue λ_1 and λ_2, we evaluate the expression $A - \lambda I$,

$$A - \lambda_1 I = \begin{pmatrix} -0.1 & 0.2 \\ 0.1 & -0.2 \end{pmatrix}, \qquad A - \lambda_2 I = \begin{pmatrix} 0.2 & 0.2 \\ 0.1 & 0.1 \end{pmatrix} \tag{22.11}$$

We now form a matrix S from these two matrices by taking one row from each and making them columns of S. It does not matter which rows we take as they are related by multiplication of a constant. Thus we choose,

$$S = \begin{pmatrix} -0.1 & 0.2 \\ 0.2 & 0.2 \end{pmatrix} \tag{22.12}$$

We use the rule

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{22.13}$$

to find the inverse of S,

$$S^{-1} = \begin{pmatrix} -4 & 4 \\ 4 & 2 \end{pmatrix} \tag{22.14}$$

It is now a theorem of linear algebra that we may write

$$A = S\Lambda S^{-1} \tag{22.15}$$

where Λ is the diagonal matrix of eigenvalues,

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 0.7 \end{pmatrix} \tag{22.16}$$

Using the formalism, we may evaluate

$$\begin{aligned}
A^n &= \left(S\Lambda S^{-1}\right)^n && (22.17)\\
&= \prod_{k=1}^{n} \left(S\Lambda S^{-1}\right) && (22.18)\\
&= \left(S\Lambda S^{-1}\right)\left(S\Lambda S^{-1}\right)\cdots\left(S\Lambda S^{-1}\right) && (22.19)\\
&= S\Lambda\left(S^{-1}S\right)\Lambda\left(S^{-1}\cdots S\right)\Lambda S^{-1} && (22.20)\\
&= S\Lambda^n S^{-1} && (22.21)\\
A^\infty &= S\Lambda^\infty S^{-1} && (22.22)\\
&= S\begin{pmatrix} 1 & 0 \\ 0 & 0.7 \end{pmatrix}^\infty S^{-1} && (22.23)\\
&= S\begin{pmatrix} 1^\infty & 0 \\ 0 & 0.7^\infty \end{pmatrix} S^{-1} && (22.24)\\
&= S\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} S^{-1} && (22.25)\\
\begin{pmatrix} x_\infty \\ y_\infty \end{pmatrix} &= S\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} S^{-1}\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} && (22.26)\\
&= \begin{pmatrix} -0.1 & 0.2 \\ 0.2 & 0.2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} -4 & 4 \\ 4 & 2 \end{pmatrix}\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} && (22.27)\\
&= (x_0 + y_0)\begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix} && (22.28)
\end{aligned}$$

Note that the constant $(x_0 + y_0)$ is just a scaling constant equal to the total population. Thus the steady state response may be obtained simply by multiplying the eigenvector corresponding to the eigenvalue of unity by the total population.

As the determination of eigenvectors is easy, all we have to do is find this eigenvector and multiply by the population. As such, this method is the best of the ones illustrated.

22.2 Bacterial Growth

Let the function $x(t)$, real-valued and continuous, denote the number of bacteria in a population at time t. To describe the growth of the population, we must formulate a model based on some postulated mechanism for the manner in which the number of bacteria can change. We assume (1) that at time t there are x bacteria in the population and (2) that the population can only increase in size and that the increase in the interval $(t, t + \Delta t)$ is proportional to the number of bacteria present at time t. Hence, we have the relation

$$\Delta x(t) = \lambda x(t)\Delta t \qquad \lambda > 0 \tag{22.29}$$

which leads to the differential equation

$$\frac{dx(t)}{dt} = \lambda x(t) \tag{22.30}$$

If we assume that $x(0) = x_0 > 0$, then the solution is

$$x(t) = x_0 e^{\lambda t} \tag{22.31}$$

In this simple model we have not made any assumption about the removal of bacteria from the population; hence it is clear on biological grounds, as well as from the solution, that as $t \to \infty$ the population size will go from x_0 to ∞. The distinguishing feature of the deterministic solution is that it tells us that, whenever the initial value x_0 is the same, the population size will always be the same for a given time $t > 0$.

We now consider the stochastic analogue of the above model. Let the integer-valued random variable $X(t)$ represent the number of bacteria in a population at time t, and let us assume that $X(0) = x_0 > 0$. In the stochastic approach we do not derive a functional equation for $X(t)$; instead, we attempt to find an expression for the probability that at time t the population size is equal to x. Hence, we seek $P_x(t) = \mathcal{P}(X(t) = x)$.

To formulate the stochastic model, we assume (1) that, if at time t there are $x > 0$ bacteria in the population, the probability that in the interval $(t, t + \Delta t)$ one bacterium will be added to the population is equal to $\lambda x \Delta t + o(\Delta t)$, $\lambda > 0$, and (2) that the probability of two or more bacteria being added to the population in $(t, t + \Delta t)$ is $o(\Delta t)$. These assumptions lead to the relation

$$P_x(t + \Delta t) = (1 - \lambda x \Delta t) P_x(t) + \lambda(x - 1) P_{x-1}(t) + o(\Delta t) \tag{22.32}$$

As $\Delta t \to 0$ we obtain the system of differential-difference equations

$$\frac{dP_x(t)}{dt} = -\lambda x P_x(t) + \lambda(x - 1) P_{x-1}(t) \qquad x = x_0, x_0 + 1, \cdots \tag{22.33}$$

Since we have assumed that $X(0) = x_0$, the above equation has to be solved with the initial conditions,

$$\begin{aligned} P_x(0) &= 1 \qquad \text{for } x = x_0 && (22.34) \\ &= 0 \qquad \text{otherwise} && (22.35) \end{aligned}$$

The solution is therefore given by

$$P_x(t) = P(X(t) = x) = \binom{x-1}{x-x_0} e^{-\lambda x_0 t} \left(1 - e^{-\lambda t}\right)^{x - x_0} \tag{22.36}$$

for $x \geq x_0$.

To compare the two models, we first observe that in the deterministic approach the population size was represented by a real-valued and continuous function of time, while in the stochastic approach we start by assuming that the random variable denoting the population size is integer-valued. An examination of the deterministic solution 22.31 shows that, for λ and x_0 fixed, we have associated with every value of t a real number $x(t)$. From 22.36 we see that, for λ and x_0 fixed, and for every pair (x, t), $x \geq x_0$, $t \geq 0$, there exists a number $P_x(t)$, $0 \leq P_x(t) \leq 1$, which is the probability that the random variable will assume the value x at time t. It is of interest to note that the deterministic model is a special case of a stochastic model, in the sense that it yields results which hold with probability one.

Consider the mean or expected population size. Let $m(t) = E(X(t))$, the expectation value. By definition,

$$m(t) = \sum_{x=0}^{\infty} x P_x(t) = x_0 e^{\lambda t} \tag{22.37}$$

Hence, we see that the expression for the mean population size 22.37 is the same as that for the population size 22.31 obtained from the deterministic model. In view of this correspondence, we can state that equation 22.30 describes the mean population size, while equation 22.33 takes into consideration random fluctuations. It is of interest to point out that this correspondence between the two models here considered does not hold in general; the deterministic solution is not always the same as the stochastic mean.

Chapter 23

The Markov Process

23.1 Markov Process

Before we give the definition of a Markov process, we will look at an example:

23.1.1 Bus Riders

Example 68 *Suppose that the bus ridership is studied in a city. After examining several years of data, it was found that 30% of the people who regularly ride on buses in a given year do not regularly ride the bus in the next year. Also it was found that 20% of the people who do not regularly ride the bus in that year, begin to ride the bus regularly the next year. If 5000 people ride the bus and 10,000 do not ride the bus in a given year, what is the distribution of riders/non-riders in the next year? In 2 years? In n years? First we will determine how many people will ride the bus next year. Of the people who currently ride the bus, 70% of them will continue to do so. Of the people who do not ride the bus, 20% of them will begin to ride the bus. Thus:*

$$5000(0.7)+10000(0.2) = \textit{The number of people who ride bus next year} = b1 \tag{23.1}$$

By the same argument as above, we see that:

$$5000(0.3)+10000(0.8) = \textit{The number of people who do not ride the bus next year} = b2 \tag{23.2}$$

This system of equations is equivalent to the matrix equation: $Mx = b$ where

$$M = \begin{pmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{pmatrix}, x = \begin{pmatrix} 5000 \\ 10000 \end{pmatrix} \textit{ and } b = \begin{pmatrix} b1 \\ b2 \end{pmatrix} \tag{23.3}$$

Note

$$b = \begin{pmatrix} 5500 \\ 9500 \end{pmatrix}. \tag{23.4}$$

For computing the result after 2 years, we just use the same matrix M, however we use b in place of x. Thus the distribution after 2 years is $Mb = M^2x$. In fact, after n years, the distribution is given by M^nx.

23.1.2 Definitions

The forgoing example is an example of a Markov process. Now for some formal definitions:

Definition 44 *A* stochastic process *is a sequence of events in which the outcome at any stage depends on some probability.*

Definition 45 *A* Markov process *is a stochastic process with the following properties:*

1. *The number of possible outcomes or states is finite.*
2. *The outcome at any stage depends only on the outcome of the previous stage.*
3. *The probabilities of any outcome become constant over time.*

If x_0 is a vector which represents the initial state of a system, then there is a matrix M such that the state of the system after one iteration is given by the vector Mx_0. Thus we get a chain of state vectors: x_0, Mx_0, M^2x_0, $\cdots$ where the state of the system after n iterations is given by M^nx_0. Such a chain is called a *Markov chain* and the matrix M is called a *transition matrix*.

The state vectors can be of one of two types: an absolute vector or a probability vector. An absolute vector is a vector where the entries give the actual number of objects in a give state, as in the first example. A *probability vector* is a vector where the entries give the percentage (or probability) of objects in a given state. We will take all of our state vectors to be probability vectors from now on. Note that the entries of a probability vector add up to 1. The main theorem on the third Markov processes concern property above, namely the notion that the probabilities become constant over time.

Theorem 18 *Let M be the transition matrix of a Markov process. Then there exists a vector x_s such that $Mx_s = x_s$. Moreover, if M^k has only positive entries for some k, then x_s is unique.*

The vector x_s is called a *steady-state vector*. The transition matrix of an n-state Markov process is an $n \times n$ matrix M where the i, j entry of M represents the probability that an object in state j transitions into state i, that is if $M = (m_{ij})$ and the states are S_1, S_2, $\cdots$, S_n then m_{ij} is the probability that an object in state S_j transitions to state S_i. What remains is to determine the steady-state vector. Notice that we have the chain of equivalences:

$$Mx_s = x_s \Rightarrow Mx_s - x_s = 0 \Rightarrow Mx_s - Ix_s = 0 \Rightarrow (M - I)x_s = 0 \Rightarrow x_s \in N(M - I) \tag{23.5}$$

Thus x_s is a vector in the nullspace of $M - I$. If M^k has all positive entries for some k, then $\dim(N(M - I)) = 1$ and any vector in $N(M - I)$ is just a scalar multiple of x_s. In particular if

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \tag{23.6}$$

is any non-zero vector in $N(M - I)$, then $x_s = \frac{1}{c}x$ where $c = x_1 + \cdots + x_n$.

Example 69 *A certain protein molecule can have three configurations which we denote as C_1, C_2 and C_3. Every second the protein molecule can make a transition from one configuration to another configuration with the following probabilities:*

$$C_1 \rightarrow C_2, P = 0.2 \qquad C_1 \rightarrow C_3, P = 0.5 \tag{23.7}$$

$$C_2 \rightarrow C_1, P = 0.3 \qquad C_2 \rightarrow C_3, P = 0.2 \tag{23.8}$$

$$C_3 \rightarrow C_1, P = 0.4 \qquad C_3 \rightarrow C_2, P = 0.2 \tag{23.9}$$

Find the transition matrix M and steady-state vector x_s for this Markov process. Recall that $M = (m_{ij})$ where m_{ij} is the probability of configuration C_j transitioning to C_i. Therefore

$$M = \begin{pmatrix} 0.3 & 0.3 & 0.4 \\ 0.2 & 0.5 & 0.2 \\ 0.5 & 0.2 & 0.4 \end{pmatrix} \tag{23.10}$$

and

$$M - I = \begin{pmatrix} -0.7 & 0.3 & 0.4 \\ 0.2 & -0.5 & 0.2 \\ 0.5 & 0.2 & -0.6 \end{pmatrix} \tag{23.11}$$

Now we compute a basis for $N(M-I)$ by putting $M-I$ into reduced echelon form:

$$U = \begin{pmatrix} 1 & 0 & -0.8966 \\ 0 & 1 & -0.7586 \\ 0 & 0 & 0 \end{pmatrix} \tag{23.12}$$

and we see that

$$x = \begin{pmatrix} 0.8966 \\ 0.7586 \\ 1 \end{pmatrix} \tag{23.13}$$

is the basis vector for $N(M-I)$. Consequently, $c = 2.6552$ and

$$x_s = \begin{pmatrix} 0.3377 \\ 0.2850 \\ 0.3766 \end{pmatrix} \tag{23.14}$$

is the steady-state vector of this process.

23.2 Random Walks

We take a little tour aside from Markov processes to look at random walks that we will need later. Intuitively a random walk is exactly that: A time-ordered series of states that are selected in accordance to some underlying probabilistic laws. An example is: Walk for two meters, turn in an arbitrary direction (all direction equally likely) and continue. Random walks are very important in many areas of scientific research and are used to simulate otherwise very complex deterministic processes in order to simplify them. An example is the Brownian motion of one type of substance in another (for example a drop of milk in a cup of coffee) in which the emersed substance thins out in the ambient substance according to a kind of random walk.

Definition 46 *Consider a family of independent and identically distributed random variables* $\{X_i\}$ *with* $P(X_n = 1) = p$ *and* $P(X_n = -1) = q = 1 - p$. *Furthermore let* $S_0 = 0$ *and* $S_n = X_1 + X_2 + \cdots + X_n$ *for* $n \geq 1$. *The values* $\{S_i\}$ *form a* simple random walk *that is* symmetrical *if, in addition,* $p = q = 1/2$.

By "simple" we mean that the value of the random variable is either 1 or -1; a general random walk does not restrict itself in this way. A random walk is *homogenous* in both time and space. This means

$$\begin{aligned} P(S_{m+n} = b|S_n = a) &= P(S_{m+r} = b|S_r = a) && (23.15) \\ P(S_{m+n} = a|S_n = 0) &= P(S_{m+n} = a + b|S_n = b) && (23.16) \end{aligned}$$

for all m, n, r, a and b. We will say that the random walk has returned to the origin if S_n takes the value zero, its initial value, at some $n \neq 0$.

The simple random walk returns to the origin certainly if and only if $p = 1/2$. In all other circumstances, the simple random walk may or may not return. In the case that $p = 1/2$, the mean time taken to returning is, however, infinite. Thus this result should be taken in the asymptotic convergence sense.

This can be extended to yield the result that the probability of visiting the point $\pm r$ is given by the following formulae

1. if $p = 1/2$, visiting $\pm r$ is certain.
2. if $p < 1/2$ and $r > 0$, $P(\text{visit } -r) = 1$ and $P(\text{visit } +r) = (p/q)^r$
3. if $p > 1/2$ and $r > 0$, $P(\text{visit } +r) = 1$ and $P(\text{visit } -r) = (q/p)^r$

If it is certain that we go to a certain value, then it is certain that we will do so again and again. Thus in the case $p = 1/2$, we reach every value infinitely often.

Example 70 *Jack and Jill play a game together. Jack initially has* $c - a$ *dollars and Jill has* a *dollars at the start. We assume that* $0 \leq a \leq c$. *At each play of the game, Jill wins or looses one dollar from or to Jack with probability* p *and* q *respectively. The game ends when either players fortune is zero. We are to take the role of a conservative friend of Jill's who wants to dissuade her from playing. Thus we wish to calculate the probability of her loosing her fortune.*

This is a random walk with $S_0 = a$ *as we are focusing on Jill and we seek the probability that* $S_n = 0$ *before* $S_n = c$ *as this corresponds to Jill loosing. Denote this desired probability by* p_a, *i.e.*

$$p_a = P(S_n = 0 \textit{ before } S_n = c|S_0 = a) = P(Loss|S_0 = a) \qquad (23.17)$$

Of course, $p_0 = 1$ *and* $p_c = 0$ *and so the only interesting case is when* $0 < a < c$.

$$\begin{aligned} p_a &= P(Loss \bigcap X_1 = +1|S_0 = a) + P(Loss \bigcap X_1 = -1|S_0 = a) && (23.18) \\ &= pP(Loss|X_1 = +1, S_0 = a) + qP(Loss|X_1 = -1, S_0 = a) && (23.19) \\ &= pP(Loss|S_0 = a + 1) + qP(Loss|S_0 = a - 1) && (23.20) \\ &= pp_{a+1} + qp_{a-1} && (23.21) \end{aligned}$$

having used time and space homogeneity of the random walk. The recurrence relation must be solved and this is typically done by a substitution $p_a = x^a$. *We obtain*

$$x^a = px^{a+1} + qx^{a-1} \qquad \text{for } 1 \leq a \leq c-1 \tag{23.22}$$

This has solutions $x = 1, q/p$. *When* $p \neq q$, *these solutions are different and the general solution of the recurrence relation is thus*

$$p_a = A + B\left(\frac{q}{p}\right)^a \tag{23.23}$$

and when $p = q = 1/2$ *the general solution is*

$$p_a = C + Da \tag{23.24}$$

In both cases, we have boundary conditions that thus allow the constants A, B, C *and* D *to be found and thus,*

$$p_a = 1 - \frac{a}{c}; \qquad p_a = \frac{x^c - x^a}{x^c - 1} \tag{23.25}$$

for the cases of $p = q = 1/2$ *and* $p \neq q$ *respectively where* $x = q/p$. *If we interchange the roles of the players we arrive at the conclusion that the probability of either person to loose is one, i.e. the game must end at some time.*

As the game is finite, denote by T_a *the time that the game lasts. Naturally* $T_0 = T_c = 0$ *and so we again take* $0 < a < c$. *Identically to above, we find*

$$T_a = 1 + pT_{a+1} + qT_{a-1} \tag{23.26}$$

We take μ_a *to be the expectation of* T_a *and so it obeys the same equation. Solving it in the same way as before, we obtain*

$$\mu_a = a(c-a), \qquad \mu_a = \frac{c - a - cp_a}{p - q} \tag{23.27}$$

respectively. Suppose that $p = q = 1/2$ *and Jill starts with one dollar and Jack starts with 99 dollars. Curiously the chance that Jill will win is 1/100 but the expectation of game length is 99 plays. This is curious because the chance of Jill loosing on first play is 0.5 and so on. This is an example of a case where the expectation value is not the same as the typical outcome of the situation. One must beware of such things in situations where it is less easy to see through it!*

If we set $c = \infty$, *Jill is playing an infinitely rich opponent. In this case, her loss is certain. This is the position of a gambler in a casino, for example.*

Let's record the random walk by considering the time as well as the location, i.e. we are going to record (n, S_n) at every step. This allows a graph to be produced with n on the horizontal axis as time. If a and b are integers and $n > 0$, we define $N(a, b, n)$ to be the number of paths from $(0, a)$ to (n, b) and $N^0(a, b, n)$ to be the number of such paths that cross the time axis, i.e. that contain a $(k, 0)$ for some $k > 0$. The important reflection principle states that if $a > 0$ and $b > 0$, we have $N^0(a, b, n) = N(-a, b, n)$. This principle is used to prove some interesting results.

Theorem 19 (Ballot Theorem) *There are two candidates for a political office, A gets x votes and B gets y votes with $x > y$. Assuming the votes are randomly mixed before being counted out, the probability that A is always ahead of B during the counting is $(x-y)/(x+y)$.*

Theorem 20 *In a symmetric simple random walk with $m \geq 1$, we have*

$$\begin{aligned} P(S_{2m}=0) &= P(S_1 \neq 0, S_2 \neq 0, \cdots, S_{2m} \neq 0) && (23.28) \\ &= P(S_1 \geq 0, S_2 \geq 0, \cdots, S_{2m} \geq 0) && (23.29) \\ & && (23.30) \end{aligned}$$

Chapter 24

Markov Chains

24.1 Recurrence

We first define a new concept.

Definition 47 *If for some $n \geq 0$, we have $P^n_{ij} > 0$ where P is the transition matrix, then we say that the state j is* accessible *from the state i. This is denoted by $i \rightarrow j$. If we have both $i \rightarrow j$ and $j \rightarrow i$, then the two states are said to* communicate.

The definition is pretty obvious, if a non-zero transition probability from i to j exists, then it is possible to transit and the future state is accessible. If transition is possible either way, they communicate. One may prove that communication is an equivalence relation (reflexive, transitive and symmetric) which is a neat thing to be.

The equivalence relation partitions the states into equivalence classes. A class is *closed* if it can not be left after it has been entered. If all states communicate (so that there is just one class) the chain is called *irreducible* and *reducible* otherwise.

Definition 48 *We denote*

$$f^{(1)}_{ij} = P(X_1 = j | X_0 = i) \tag{24.1}$$

and for $n > 1$

$$f^{(n)}_{ij} = P(X_1 \neq j, X_2 \neq j, \cdots, X_{n-1} \neq j, X_n = j | X_0 = i) \tag{24.2}$$

as the probability that given the first state i, we visit state j for the first time at time n. Thus

$$f_{ij} = \sum_n f^{(n)}_{ij} \tag{24.3}$$

is the probability that we ever visit j starting at i. If $f_{ii} < 1$, we say that i is transient *and otherwise* recurrent.

The message here is simple: If the likelihood that we never return to state i is non-zero, the state is a transient state that is, in some sense, unreliable. If it is certain that we will get back to i at some point, then the state is obviously recurrent.

As the future is independent of the past path, the probability of reaching i exactly N times is f_{ii}^N. For the recurrent state this probability is always unity and for a transient state this tends to zero as N gets large. This is encapsulated by the following theorem.

Theorem 21 *The state i is transient if and only if $\sum_n P_{ii}^n$ converges and the state is recurrent if and only if the same sum diverges. The properties of transience and recurrence belong to the whole equivalence class of states.*

We remember that return to the origin is only possible after an even number of steps. Hence the sequence $\{P_{ii}^n\}$ is alternatively zero and non-zero. If this occurs, we may or may not find that the non-zero terms converge. As the whole sequence does not converge, we need to deal with these cases.

Definition 49 *We define d_i to be the greatest common divisor of the sequence of n's when $P_{ii}^n > 0$. If $d_i = 1$, state i is* aperiodic *and is* periodic *with period d_i otherwise.*

Thus simple random walk states have period two. One can show that if i and j belong to the same equivalence class, then $d_i = d_j$. There are two *sufficient* conditions for aperiodicity: (1) some diagonal entry P_{ii} is non-zero and (2) there exist paths of 2 and 3 steps from state i to itself.

The periodic case is an annoyance and we want to get rid of it for further analysis. Let P be the transition matrix of an irreducible but periodic chain with period $d > 1$. For any state i, we put

$$C(i) = \{j : p_{ij}^{(dn)} > 0 \text{ for some } n > 0\} \tag{24.4}$$

It can be shown that the $C(i)$ are either identical or disjoint, i.e. they do not partially overlap. Consider the transition matrix $Q = P^d$. The chain with transition matrix Q is reducible with the d sets $C(i)$ as its equivalence classes. The original period chain is just like d aperiodic chains operating in parallel with a new timescale. Thus we can now focus on irreducible and aperiodic chains. The major result follows.

Theorem 22 *For all states i and j in an irreducible, aperiodic Markov chain, the following holds*

1. *If the chain is transient, $p_{ij}^{(n)} \to 0$.*
2. *If the chain is recurrent, $p_{ij}^{(n)} \to \pi_j$ where either*
   (a) *Every $\pi_j = 0$ (such a chain is called* null recurrent*), or*
   (b) *Every $\pi_j > 0$ with $\sum \pi_j = 1$ and π is the unique eigenvector of P corresponding to the eigenvalue one (such a chain is called* positive recurrent*).*
3. *If the chain is recurrent, let T_i represent the time to return to state i. Then $\mu_i = E(T_i) = 1/\pi_i$ with the understanding that $\mu_i = \infty$ if $\pi_i = 0$.*

This allows us to specify the long-term behavior of an irreducible aperiodic Markov chain. The initial probability state vector is x_0, the probability state after some n steps is x_n and we have from before

$$x_n = P^n x_0 \tag{24.5}$$

which allows us simply to iterate the whole thing. Now, in a transient or null recurrent chain $x_n \to 0$, i.e. the state tends to the zero vector. In different words, the probability that the n^{th} state is state j tends to zero for any particular state j. For a positive recurrent chain $x_n \to \pi$ where π is the probability vector from the theorem. Thus the probability that the n^{th} state is state j tends to $\pi_j > 0$ for any state j. The difference between null and transient chains is the frequency of visiting i. In a transient chain we may never make a visit and we cannot return infinitely often. In a null chain we are certain to make infinitely many visits although the mean time between visits is infinite. We may decide between the three alternatives using a criterion.

Theorem 23 (Forster's Criterion) *Consider an irreducible and aperiodic Markov chain with transition matrix P and state space $S = \{0, 1, 2, \cdots\}$. Let Q be P except that the row and column corresponding to state 0 are deleted. Then*

1. *The chain is transient if and only if the system $Qy = y$ has a bounded non-zero solution.*

2. *The chain is positive recurrent if and only if the system $\pi P = \pi$ has a solution π that is a probability vector.*

If the chain is recurrent, there is a strictly positive solution of $\pi P = \pi$ that is unique up to a multiplicative constant. This chain is positive or null depending on whether $\sum \pi_i$ is finite or infinite respectively.

24.2 Random Walk with Barrier

Consider a simple random walk with an *impenetrable barrier* at zero, i.e. a state space $\{0, 1, 2, \cdots\}$, generally with $P_{i,i+1} = p$ and $P_{i,i-1} = q$, except that $P_{00} = q$. This is clearly irreducible and aperiodic when $0 < p < 1$. The system $Qy = y$ gives $py_2 = y_1$ and, for $n \geq 1$, we have $qy_n + py_{n+2} = y_{n+1}$. By induction, we see that $y_n = py_1(1 - (q/p)^n)/(p - q)$ if $p \neq q$, or $y_n = ny_1$ if $p = q$. This can be non-zero and bounded if and only if $p > q$: The condition for transience. The system $\pi P = \pi$ reads $(\pi_0 + \pi_1)q = \pi_0$, $p\pi_{n-1} + q\pi_{n+1} = \pi_n$ for $n \geq 1$, with unique solution $\pi_n = \pi_0(p/q)^n$. We have π a probability vector, if and only if $p < q$: The condition for positive recurrence. When $p = q$, $\sum \pi_n = \infty$ if $\pi_0 \neq 0$, and the chain is null recurrent.

In summary, the chain is transient, and $X_n \to +\infty$ when $p > 1/2$; positive recurrent, and $P(X_n = k) \to p^k/(q^{k-1}(q-p))$ when $p < 1/2$; and null recurrent when $p = 1/2$. These broad conclusions will accord with your intuition, if you consider what happens, on average, at each step.

To estimate how often a transient state is ever visited, or a recurrent state is visited up to time T, we can use indicator functions. Define $I_{jn} = 1$ if $X_n = j$, and $I_{jn} = 0$ otherwise, so that

$$Y(j) = \sum_{n=0}^{T} I_{jn} \tag{24.6}$$

is the total number of times state j is visited up to time T. Since

$$P(I_{jn} = 1) = \sum_i P(X_n = j | X_0 = i) P(X_0 = i) = \sum_i P(X_0 = i) P_{ij}^{(n)}, \tag{24.7}$$

we have

$$E(Y(j)) = \sum_{n=0}^{T} P(I_{jn} = 1) = \sum_{n=0}^{T} \sum_i P(X_0 = i) P_{ij}^{(n)} = \sum_i P(X_0 = i) \sum_{n=0}^{T} P_{ij}^{(n)}. \tag{24.8}$$

For example, if we definitely start in state i, so that $P(X_0 = i) = 1$ and $P(X_0 = k) = 0$ for $k \neq i$, the mean number of visits to j up to time T is

$$\sum_{n=0}^{T} P_{ij}^{(n)} = \left(\sum_{n=0}^{T} P^n \right)_{ij}. \tag{24.9}$$

Recall the gamblers example with Jack and Jill from last lecture. As a Markov chain, the states are $\{0, 1, \cdots, c\}$, and the non-zero transition probabilities are $P_{00} = P_{cc} = 1$, with $P_{i,i-1} = q$ and $P_{i,i+1} = p$ for $1 \leq i \leq c-1$. These are three classes: $\{0\}$, $\{c\}$, both recurrent, and $\{1, 2, \cdots, c-1\}$, clearly transient.

It is convenient to write the transition matrix P with the states in a non-standard order: Let

$$P = \begin{array}{r} 1 \\ 2 \\ 3 \\ \vdots \\ c-1 \\ 0 \\ c \end{array} \begin{pmatrix} 0 & p & 0 & \cdots & q & 0 \\ q & 0 & p & \cdots & 0 & 0 \\ 0 & q & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & p \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix} = \begin{pmatrix} Q & A \\ 0 & I \end{pmatrix} \tag{24.10}$$

where Q is a $(c-1) \times (c-1)$ matrix.

Because of the structure of P, the powers $\{P^n\}$ have a similar structure, with

$$P^n = \begin{pmatrix} Q^n & A_n \\ 0 & I \end{pmatrix} \tag{24.11}$$

for some $(c-1) \times 2$ matrix A_n. Hence

$$\sum_{n=0}^{T} P^n = \begin{pmatrix} \sum\limits_{n=0}^{T} Q^n & \sum\limits_{n=0}^{T} A_n \\ 0 & (n+1)I \end{pmatrix} \tag{24.12}$$

We are interested in the expression

$$S_T = \sum_{n=0}^{T} Q^n. \tag{24.13}$$

Now

$$(I-Q)S_T = \sum_{n=0}^{T} Q^n - \sum_{n=0}^{T} Q^{n+1} = I - Q^{T+1}. \tag{24.14}$$

But $Q^n \to 0$ as $n \to \infty$, since the states $\{1, 2, \cdots, c-1\}$ are transient, so $(I-Q)S_T \to I$, and hence $S_T \to (I-Q)^{-1}$. This shows that in the gamblers ruin problem, the mean total number of visits to state j, starting from state i, is $((I-Q)^{-1})_{ij}$, if $1 \leq i, j \leq c-1$.

For example, if $c = 4$ then

$$Q = \begin{pmatrix} 0 & p & 0 \\ q & 0 & p \\ 0 & q & 0 \end{pmatrix}, \tag{24.15}$$

and so

$$(I-Q)^{-1} = \frac{1}{p^2+q^2} \begin{pmatrix} 1-pq & p & p^2 \\ q & 1 & p \\ q^2 & q & 1-pq \end{pmatrix}. \tag{24.16}$$

Starting with unit amount, the mean number of times over the whole game that we possess exactly three units is the $(1,3)$ entry $p^2/(p^2+q^2)$.

In this fashion, we can find the mean total number of visits to any transient state of a general Markov chain. For a recurrent state, this mean number is either zero (if we cannot reach it from our starting point) or infinite.

A systematic way to assess the long-term behavior of a Markov chain with transition matrix P, i.e. the fate of X_n, conditional on $X_0 = i$, might proceed as follows.

1. Find the classes and establish which, if any, are closed. Find the period of each closed class. If the chain has no closed classes, all states are transient and $P(X_n = j) \to 0$ for all j.

2. For each closed aperiodic class C, determine whether it is transient, null or positive. In the first two cases, $P(X_n = j) \to 0$ for all $j \in C$, otherwise $P(X_n = j | C \text{ ever entered}) \to \pi_j > 0$ for all $j \in C$.

3. For each closed class C of period $d > 1$, let P_0 be that part of P that describes transitions among the states of C alone, and let $Q = P_0^d$. For the transition matrix Q, C splits into d aperiodic subclasses, each to be treated as in the above case. Ascertain the order in which these subclasses are visited.

4. Denote the closed classes by C_1, C_2, $\cdots$, and write R = Rest of the states. Write $x_{ij} = P(\text{Eventually enter } C_j | X_0 = i)$. Considering one step,

$$x_{ij} = \sum_{k \in R} P_{ik} x_{kj} + \sum_{k \in C_j} P_{ik}, \qquad i \in R, j = 1, 2, \cdots \tag{24.17}$$

from which the $\{x_{ij}\}$ are to be found, and then use the above two steps.

Chapter 25

Queuing Theory

Please see sections 8.1 and 8.2 in the attached photocopies please for this chapter.

Chapter 26

Brownian Motion

Please see sections 8.3 and 8.4 in the attached photocopies please for this chapter.

Chapter 27

Markov Random Fields

27.1 Markov Random Fields

A *Markov random field* is a stochastic process defined on a two-dimensional set, i.e. a region of the plane, that has a Markov property. Let $\{X_k\}$ be a Markov chain taking values in a finite set. Then we may show that the distribution of X_n conditional on the values at all other time points, depends on the values at the time points $n-1$ and $n+1$ only. If we call these time points the *neighbors* of n, the distribution depends only on the neighborhood of X_n.

Let Z be the set of integers and let $S \in Z^2$ be a finite rectangular two-dimensional lattice of integer points. Typically we will let

$$S = \{0, 1, \cdots, n-1\} \times \{0, 1, \cdots, m-1\} \tag{27.1}$$

for some n and m, so that S contains nm points. These points are often called *sites*. We must now define what the neighborhood of a point is going to be. This definition is up to us and depends on the use to which we wish to put the model. We will adopt some restrictions however: (1) a site must not neighbor itself and (2) the neighborhood property is symmetric. We will write $s \sim t$ if the two sites $s, t \in S$ are neighbors. Some common neighborhood structures are shown in figure 27.1. If s is a site, we define the *neighborhood* N_s of s as the set of all neighbors,

$$N_s = \{t \in S : t \sim s\} . \tag{27.2}$$

And so figure 27.1 shows the neighborhood of the middle site for two different structures. We note that in these structures, special care must be taken at the edge of the lattice S, since site located there have smaller neighborhoods. One may get around these difficulties by identifying the borders with each other, i.e. if we leave on one side, we enter on the opposing side. This technique is called *periodic boundary conditions* and is very important in applied mathematics.

We consider a Markov random field $\{X(s)\}_{s\in S}$ defined on S, i.e. a collection $X(s)$ of random variables indexed by sites in S. These random variables are assumed to take their values in a finite set $\mathcal{X}$, the state space. Examples of $\mathcal{X}$ that we will use is $\mathcal{X} = \{-1, +1\}$ and $\mathcal{X} = \{1, 2, \cdots, r\}$. The set $\mathcal{X}^S$ is the set of elements of the form $x = \{x(s)\}_{s\in S}$ with $x(s) \in \mathcal{X}$ for each s. An element of $\mathcal{X}^S$ will often be referred to as a *configuration* of the Markov random field. In addition, we will often write

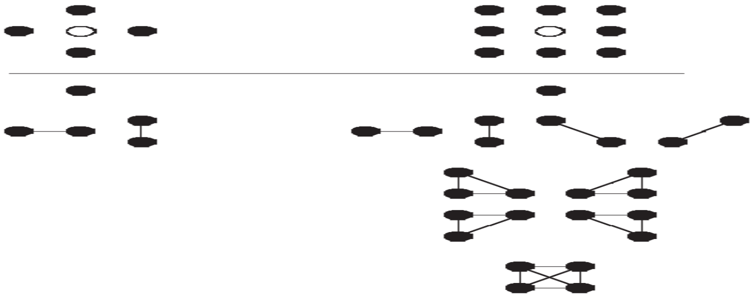


Figure 27.1: Two different neighborhood structures and their corresponding cliques. Left: Four closest points are neighbors (top) and the cliques (bottom). Right: Eight closest points are neighbors (top) and the cliques (bottom. The concept of cliques will be defined below.

X for $\{X(s)\}_{s\in S}$ and think of X as a random variable with values in $\mathcal{X}^S$, the set of configurations. Letting $|S|$ denote the number of elements of S and similarly for $\mathcal{X}$, the number of elements of the configuration space $\mathcal{X}^S$ is $|\mathcal{X}|^{|S|}$ and it is hence often extremely large. For example, if $\mathcal{X} = \{-1, +1\}$ and S is a lattice of size 128×128, its size is 2^{128^2}. If A is a subset of S, we write $X(A)$ for $\{X(s)\}_{s\in A}$, i.e. the collection of random variables on A, and similarly for a particular configuration $x = \{x(s)\}_{s\in S}$. We also recall that the symbol $\setminus$ denotes set-difference. For example, $S \setminus \{s\}$ is the set of sites in S except s. We also write this difference as $S \setminus s$.

We now say that the random field X is a *Markov random field* (MRF) on S with respect to the given neighborhood structure if

$$P\left(X(s) = x(s)|X(S\setminus s) = x(S\setminus s)\right) = P\left(X(s) = x(s)|X(N_s) = x(N_s)\right) \quad (27.3)$$

for all sites $s \in S$ and all configurations $x \in \mathcal{X}^S$. That is, the distribution of $X(s)$, given all other sites, depends on the realized values in its neighborhood only. These conditional distributions are often called the *local specification* of the MRF.

27.1.1 The Ising Model

Let $\mathcal{X} = \{-1, +1\}$ and the neighborhood structure be the left one in figure 27.1 and the local specification be

$$P\left(X(s) = x(s)|X(N_s) = x(N_s)\right) = \frac{\exp\left(\beta \sum\limits_{t\in N_s} x(s)x(t)\right)}{2\exp\left(\beta \sum\limits_{t\in N_s} x(t)\right)} \quad (27.4)$$

for some real number β. Note that the denominator does not depends on $x(s)$ and is only a normalizing factor to make the right hand side a proper distribution, i.e. summing to unity. This model is called the *Ising model* after the German physicist Ising who invented it to explain ferromagnetism.

The sum in the exponent is positive if $x(s)$ has the same sign as the majority of its neighbors. Hence, if $\beta > 0$ the sites interact such that configurations x with many neighbors of the same sign will have large probabilities. On the contrary, if $\beta < 0$ configurations with many neighbors having opposite signs will have large probabilities.

27.1.2 The Potts Model

We only require $\mathcal{X}$ to be finite, the neighborhood structure to be one of the two illustrated in figure 27.1 and the local specification to be

$$P\left(X(s) = x(s)|X(N_s) = x(N_s)\right) = \frac{\exp\left(\beta\#\left\{t \in N_s|x(t) \neq x(s)\right\}\right)}{\sum\limits_{i\in\mathcal{X}} \exp\left(\beta\#\left\{t \in N_s|x(t) \neq i\right\}\right)} \tag{27.5}$$

for some real number β. As before, the denominator does not depend on $x(s)$ and is only a normalizing factor. This model is called the *Potts model.* We note that $\#\left\{t \in N_s|x(t) \neq x(s)\right\}$ is the number of neighbors of s that have values different from $x(s)$. Hence, if $\beta > 0$ this model gives large probabilities to configurations x in which there are many neighbors with different values. If $\beta < 0$, the model works the opposite way.

27.2 Cliques and the Gibbs Distribution

So far we have talked about the local specification of an MRF, but we could also be interested in a corresponding distribution on $\mathcal{X}^S$, that is, in the probabilities of various configurations x. We will denote this distribution by π,

$$\pi(x) = P(X = x) = P(X(s) = x(s) \text{ for all } s \in S) \tag{27.6}$$

for any configuration $x \in \mathcal{X}^S$. In defining such a distribution, it turns out to be convenient to introduce a concept called *clique.* Cliques are particular subsets of sites in S, defined in the following way: (1) Any single site s is a clique and (2) any subset $C \subset S$ of more than one site is a clique if all pairs of sites in C are neighbors.

What cliques look like depends on the neighborhood system. Figure 27.1 shows what cliques there are for the two neighborhood systems displayed therein. Note that these schematic cliques should be moved around over the lattice to find all subsets of sites that fit with the given pattern. For example, in the left neighborhood system of figure 27.1, pairs (s, t) of sites such that s is located immediately above t are different cliques, but of the same form.

Now assume that for each clique C there is a function $V_C|\mathcal{X}^S \to R$. That is, V_C maps a configuration x into a real number. Moreover, V_C must not depend on sites other than those in C; we could write this as $V_C = V_C(X(C))$. A probability mass function (pmf), or distribution, π on the configuration space $\mathcal{X}^S$ of the form

$$\pi(x) = Z^{-1} \exp\left(\beta \sum_C V_C(x)\right) \tag{27.7}$$

is called a *Gibbs distribution.* Here the sum runs over all cliques C. The normalizing constant Z is given by

$$Z = \sum_{x\in\mathcal{X}^S} \exp\left(\sum_C V_C(x)\right) \tag{27.8}$$

and is generally infeasible to compute as the outer sum runs over a very large set. The importance of Gibbs distributions is made clear from the following facts: (1) Any random field with a distribution π that is a Gibbs distribution is a Markov random field with respect to the neighborhood system governing the cliques and (2) any random field that is Markov with respect to a given neighborhoods system has a distribution π that is a Gibbs distribution generated by the corresponding cliques. Hence we may say that Markov random fields and Gibbs distributions are equivalent. It is not too difficult to verify the first fact, while the second is much more difficult and is known as the Hammersley-Clifford theorem (it does require one mild condition in addition to the field being an MRF). Let us consider some specific examples of Gibbs distributions.

27.2.1 The Ising Model continued

The cliques in the Ising model are of the form given in figure 27.1, and we define the functions V_C by

$$V_{\{s\}}(x) = \alpha x(s) \tag{27.9}$$

for and singleton clique s and

$$V_{\{s,t\}}(x) = \beta x(s)x(t) \tag{27.10}$$

for any pair (s, t) of neighbors. The Gibbs distribution is then

$$\pi(x) = Z(\alpha, \beta)^{-1} \exp\left(\alpha \sum_{s \in S} x(s) + \beta \sum_{s \sim t} x(s)x(t) + \right) \tag{27.11}$$

Note that as we sum over cliques, only one of the pairs (s, t) and (t, s) are accounted for in the sum. It is not difficult to check that this distribution yields the local specification 27.4, with $\alpha = 0$. In this model, α controls the fractions of -1's and +1's in the configurations (with $\alpha > 0$ large probabilities are given to configurations with many +1's) while β controls the interaction between neighbors as previously described. Note that the normalizing constant $Z(\alpha, \beta)$ depends on the parameters α and β.

27.2.2 The Potts Model continued

The Potts model has a Gibbs distribution given by

$$\pi(x) = Z(\beta)^{-1} \exp\left(\beta \sum_{s \sim t} I(x(s) \neq x(t))\right) \tag{27.12}$$

where $I(x(s) \neq x(t))$ is an indicator function being equal to one if $x(s) \neq x(t)$ and zero otherwise. Hence, the sum counts pairs of neighbors that have values which do not agree. Again, only one of the pairs (s, t) and (t, s) are accounted for in the sum. This Gibbs distribution yields the local specification 27.5.

27.3 Markov Chain Monte Carlo Simulation

As we have seen above, the distribution π of an MRF is complex and it is defined on a set of enormous size. Hence, it does not easily lend itself to simulation. Instead, simulation of MRFs is typically done in an iterative manner, modifying one site at a time, thereby gradually building a simulated replication. We will now describe this process in detail.

Let s be a site of the lattice S on which the MRF $\{X(s)\}_{s\in S}$ is defined, and, as above, let N_s be its neighborhood. Then, for each i in the state space $\mathcal{X}$, the conditional probability $P(X(s)=i|X(N_s))$ is provided by the local specification of the field. Let us now simulate the site $X(s)$ according to these probabilities. That is, we update $X(s)$ according to its conditional distribution given the rest of the field — a distribution that in turn only depends on values in the neighborhood of s. This simulation is typically easy. For example, in the Ising model, we flip a coin that shows -1 or +1 with probabilities proportional to $\exp\left(\beta\sum_{t\in N_s} x(s)x(t)\right)$. Of course, we must divide these exponentials by their sum as to make them proper probabilities.

Returning to the general setting, if we consider the whole configuration $x\in\mathcal{X}^S$, we can write this simulation rule as a transition probability matrix P_s. This is a matrix of size $|\mathcal{X}|^{|S|}\times|\mathcal{X}|^{|S|}$; hence it is finite but very large. Fortunately, we never have to work explicitly with this matrix. Although P_s is a large matrix, most of its elements are zero. Indeed, the only non-zero transition probabilities are those corresponding to transitions in which no other sites than $x(s)$ are (possibly) modified.

The key feature of P_s is that π, the distribution of the MRF X, is a stationary distribution of this matrix. To verify this, we will show that the local balance equations are satisfied. Doing this, we only need to consider pairs $x=((x(s),X(S\setminus s))$ and $x'=((x'(s),X(S\setminus s))$ of configurations that only differ at the site s, since transitions between any other pairs of configurations have zero probability. For a pair as above, the transition probability of moving from x to x' is $P(X(s)=x'(s)|X(S\setminus s)=x(S\setminus s))$, and similarly in the other direction. The local balance equation thus reads

$$\pi(x)P(X(s)=x'(s)|X(S\setminus s)=x(S\setminus s))=\pi(x')P(X(s)=x(s)|X(S\setminus s)=x(S\setminus s)) \tag{27.13}$$

Expanding the conditional probability we find that the above equation is equivalent to

$$\pi(x)\frac{\pi(x')}{P(X(S\setminus s)=x(S\setminus s))}=\pi(x')\frac{\pi(x)}{P(X(S\setminus s)=x(S\setminus s))'} \tag{27.14}$$

which is trivially true.

We can now envision how we can simulate the MRF. Imagine a Markov chain $\{X_n\}_{n=0}^{\infty}$ of configurations of the MRF with transition probability matrix P_s. In other words, the state space of this Markov chain is the configuration space $\mathcal{X}^S$, and it moves from one configuration to another by updating $X(s)$ according to its conditional distribution given the rest of the field. This Markov chain has a stationary distribution π, but its distribution does not converge to π because it is not irreducible. Indeed, since the chain can only modify the MRF at the single site

s, it cannot move from any configuration x to any other configuration x'. Obviously, a chain that is irreducible must have the ability to modify the MRF at any site. We accomplish this by forming a new transition probability matrix

$$P = \prod_{s \in S} P_s \tag{27.15}$$

In other words, the action of P is that we visit all sites s of S (in some order), and when we visit s we modify its value by simulating $X(s)$ from its conditional distribution give the rest of the MRF. Since $\pi P_s = \pi$ for each s, it is plain that $\pi P = \pi$; that is, the distribution π of the MRF is a stationary distribution of P. Moreover, it is not difficult to check that P is irreducible (since it can modify the current configuration at all sites) and aperiodic (since it can also choose not to modify the current configuration at all). Hence, a Markov chain $\{X_n\}$ with state space $\mathcal{X}^S$ and with transition probability matrix P will converge in distribution to π. Thus we may take the initial configuration X_0 arbitrary, run the chain for a long time and then obtain a simulated configuration whose distribution is roughly π. For how long we need to run the chain in order to make this approximation good depends on the model, and is a difficult question to address in general. A full turn of the simulation scheme, in which each site is visited once, is often called a *sweep* of the algorithm.

27.3.1 The Ising Model continued

We shall now in detail describe how this simulation approach applies to the Ising model, with $\alpha = 0$. The extension to include the parameter α is straightforward. We can write the simulation scheme algorithmically as follows:

1. Traverse all sites $s \in S$ in some order, carrying out steps 2 to 5 at each site.
2. Compute $w = \sum_{t \sim s} x(t)$.
3. Set $p_- = e^{-\beta w}$ and $p_+ = e^{\beta w}$.
4. Normalize these numbers to probabilities by setting $q = p_- + p_+$ and then $p_- = p_-/q$ and $p_+ = p_+/q$.
5. Set $x(s) = -1$ with probability p_- and $x(s) = +1$ with probability p_+. This can be accomplished by drawing a random variable u, uniformly distributed on $(0, 1)$, and setting $x(s) = -1$ if $u \leq p_-$ and $x(s) = +1$ otherwise.

The simulation scheme derived above is called the *Gibbs sampler*. Figures 27.2 and 27.3 show simulations of the Ising model using a Gibbs sampler that visits sites row wise. The value of β is 0.4 in figure 27.2 and 0.4 in figure 27.3. The figures show the simulated replications after 0, 3, 10, 25, 100 and 250 sweeps of the Gibbs sampler. The replication after 0 iterations is the initial replication and is identical in both figures. We clearly see that $\beta = 0.6$ gives much more dependence in the model, compared to $\beta = 0.4$, manifested through larger regions of the same color. Also plotted in these figures are the fraction of cliques of equal sign and the fraction of sites with positive sign, as a function of the number of iterations. Again we see that $\beta = 0.6$ yields a larger fraction of cliques with equal sign, as expected, but we also see that it takes more iterations for the Gibbs sampler to reach a steady state in this respect. When $\beta = 0.4$ it appears that this curve flattens out after about 100 iterations, but when $\beta = 0.6$ it is not obvious that convergence has occurred even after 250 iterations. This indicates that the Markov chain $\{X_n\}$ that constitutes the Gibbs sampler has a slower rate of convergence when $\beta = 0.6$. This is indeed true, and is caused by the larger degree of spatial dependence in this case.

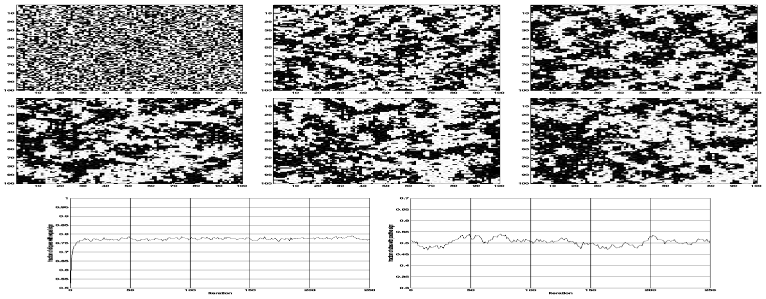


Figure 27.2: Simulation of the Ising model with $\beta = 0.4$ on a 100×100 lattice. Top left to middle right: replications after 0, 3, 10, 25, 100 and 250 iterations of the Gibbs sampler. Bottom: fraction of cliques with equal signs (left) and fraction of sites with positive sign (right) as a function of the number of iterations.

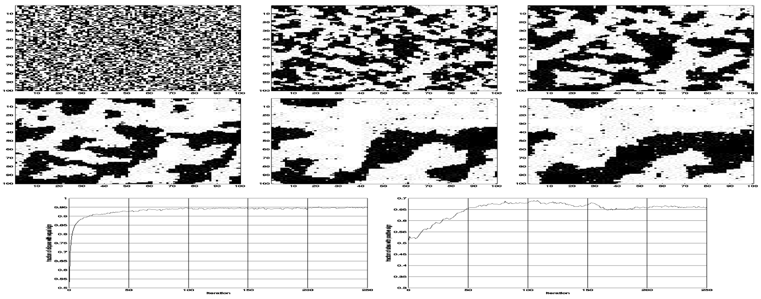


Figure 27.3: Simulation of the Ising model with $\beta = 0.6$ on a 100×100 lattice. Top left to middle right: replications after 0, 3, 10, 25, 100 and 250 iterations of the Gibbs sampler. Bottom: fraction of cliques with equal signs (left) and fraction of sites with positive sign (right) as a function of the number of iterations.

We can construct variants to the Gibbs sampler, for example by not visiting the site cyclically in a prescribed order, but rather selecting a site at random, updating it, then selecting another one at random, updating it, and so on. This scheme works equally well and corresponds to the transition probability matrix

$$P = \frac{1}{|S|} \sum_{s \in S} P_s \tag{27.16}$$

The Gibbs sampler is a particular case of what is called *Markov chain Monte Carlo* simulation methods. The general idea of these methods is to construct a Markov chain that has a prescribed stationary distribution. The Gibbs sampler is a special case of the *Metropolis-Hastings* sampler, a framework on which almost all MCMC algorithms are based.

Note that there are two "Markov concepts" involved here. First as in "Markov random field" (a random variable $X \in \mathcal{X}^S$), and then as in "Markov chain" (a sequence of configurations). These two Markov properties are independent in the sense that we could perfectly well construct a Gibbs sampler (which is a Markov chain) for *any* random field, but the Markov property of the field is extremely helpful in that the conditional distributions involved only depend on a small neighborhood and are hence easily computed.

27.4 Parameter Estimation: Pseudo-likelihood

Like most other statistical models, MRFs usually include one or more parameters. For example, in the Ising and Potts models above, there is a parameter β. In applications involving data, the values of these parameters are usually unknown, and one then wants to estimate them from the data. When it comes to MRFs, this task is far from simple. For example, consider the principle of maximum likelihood in the Ising or Potts models. Maximum likelihood then amounts to maximizing the probability $\pi(x) = \pi(x; \beta)$ over β, where x is the observed configuration. Now, β is not only found in the exponent of the Gibbs distribution, but also in the normalizing constant $Z = Z(\beta)$. Also noted above, this constant is generally infeasible to compute, and likewise it is usually unknown how it depends on β. Thus, we cannot apply maximum likelihood in a straightforward manner. It is common to attack this problem by massive Monte Carlo simulations, in which one in one way or another tries to approximate $Z(\beta)$ by simulation. There are other approaches, however, that are statistically less efficient but whose computational requirements are more moderate.

One such method is *pseudo-likelihood*. The function that one sets out to maximize is the the product, over $s \in S$, of all the local specifications $P(X(s) = x(s)|X(S \setminus s) = x(S \setminus s))$. On the log-scale we obtain the log-pseudo-likelihood function,

$$\log PL(\beta; x) = \sum_{s \in S} \log P(X(s) = x(x)|X(S \setminus s) = x(S \setminus s)). \tag{27.17}$$

The maximizer of this function is called the *maximum likelihood estimator* (MPLE). Since the local specification does not involve the normalizing constant $Z(\beta)$, the pseudo-likelihood function is free of this constant as well. This is the main advantage of this approach.

Figure 27.4 shows the log-pseudo-likelihood function for the Ising model replications displayed in figures 27.2 and 27.3, obtained after 250 iteration of the Gibbs sampler. We that there are maximas around 0.4 and 0.6, the respective true values of β, but we also see that the maximum is much more flat when $\beta = 0.6$. In statistical terms this implies that the variance of the MPLE is large, and this is in line with the rule of thumb that the MPLE is inefficient when there is a large degree of spatial dependence.

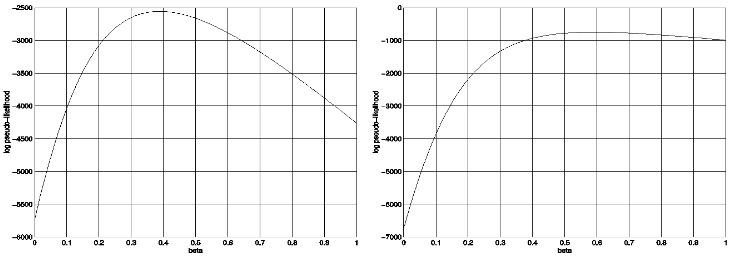


Figure 27.4: The log-pseudo-likelihood as a function of β for the replications obtained after 250 steps for figures 27.2 (left) and 27.3 (right).

Part IV

Assignments

Chapter 28

Homework

28.1 Grading Policy

28.1.1 Grades

You will be graded by being assigned points. There are 500 points in total for homeworks and 500 points for the project (see below for description). There will be opportunity for extra credit as explained below. Your total points will be divided by 1000 and your percentage compared with the following table to determine your IUB grade:

Percentage Score	Letter Grade	IUB official point grade
100 - 97	A+	1.00
96 - 93	A	1.33
92 - 90	A-	1.67
89 - 85	B+	2.00
84 - 80	B	2.33
79 - 75	B-	2.67
74 - 70	C+	3.00
69 - 65	C	3.33
64 - 60	C-	3.67
59 - 57	D+	4.00
56 - 53	D	4.33
52 - 50	D-	4.67
< 50	F	5.00

Anyone who scores over 100% will obtain a 1.00 as there is no better grade available. As we simply add up points, the extra credit points remedy any loss of points regardless of where.

28.1.2 Extra Credit

If you find mistakes or make corrections to the lecture script (including the readings) you may gain extra credit. For each change *suggested and implemented* you will get extra credit points *if you were the first person to suggest it.* The following table gives the points available.

Suggestion	Points
Grammatical error	0.5
Mathematical error in formula	1.0
Argumentative error	5.0
Improvement of explanation	5.0
Missing information	3.0

28.1.3 Attendance

Attendance to lectures is required. This will be checked by asking you to complete the reading assigned for each lecture in advance and to write down on a piece of paper at least two points that you observed, liked, found questionable or otherwise noteworthy. These papers are to be submitted *by yourself at the start of the relevant lecture.* Only one paper will be accepted per person and there will be no exceptions whatsoever regarding this or the time of submission. The last six lectures do not have reading associated with them and here the notes are to be written on the lecture script.

Note that the purpose is manyfold. I want to make sure that you do the reading, engage with it and come to class. As such this appears to be the most economical way to enforce this. I am certainly not a fan of rules but some guidelines have to be set. As there are 27 such submissions possible and, of course, the usual problems of illness and family emergency prevents submission of some of them, I want at least 20 of these papers submitted by each student. If you submit less than 20, your total point score will be multiplied by a factor. The following table gives the factors.

Submissions	Factor
20 - 27	1.00
15 - 19	0.80
10 - 14	0.70
0 - 9	0.60

Please also note that these notes can be very brief. Two or three sentences per point are completely sufficient. Do not write an essay but merely jot down some thoughts.

28.2 Homework

28.2.1 General Policy

Homeworks are due each fortnight at the start of the class starting with the third week (the dates are written in the titles of the homeworks below) and are to be written on paper and not submitted by email. The points available for each problem are written into the following problem sheets.

28.2.2 Homework 1 (19.02.2004): Basic Probability and Random Variables (75 Points)

1. A biased die was thrown 100 times and gave the following results:

   Score 1 2 3 4 5 6

   Number 17 21 15 10 21 16 Total 100

   Make the best possible estimate that the sum of the scores of the next two throws will be at least 4. **Solution:** 0.9 to one decimal place. [5]

2. Trains at the same platform at Baker Street underground station go alternately on the Circle line and the Metropolitan line. A traveller who always wishes to take a Metropolitan line train, and who can be assumed to arrive on the platform at random, finds that 403 times out of 500 the next train due is a circle line train. How can this be reconciled with the fact that the trains alternate? **Solution:** Uneven intervals between trains. [5]

3. What is the probability that if the letters of the word MANAGEMENT are arranged in random order, then the vowels will be separated. **Solution:** 1/6 [5]

4. In a certain school, the probability that a person studying German also studies Physics is 1/4, whereas the probability that someone studying Physics also studies German is 1/5. The probability that a person chosen at random studies neither is 1/3. Calculate the probability that a person chosen at random studies both Physics and German.

   Could it be said that Physics and German were independent choices at that school? **Solution:** 1/12; No. [10]

5. Smith and Wesson fight a duel with guns. The probability that Smith kills Wesson on any shot is 1/4 and that Wesson kills Smith is 1/3. Find the probability that just one of them is killed after one shot each if (a) both fire simultaneously (b) Smith fires first.

   If both fire simultaneously, what is the probability that both are still alive after the second round?

   What is Smith's probability of survival if they duel to the death, each time firing simultaneously? **Solution:** 5/12; 1/2; 1/4; 1/3. [15]

6. A batch of fifty articles contains three which are defective. The articles are drawn in succession (without replacement) from the batch and tested. Show that the chance that the first defective met will be the rth article drawn is $\frac{(50-r)(49-r)}{39200}$. [10]

7. A card is missing from a pack of 52 cards. If this is the only information you have, what is the probability that the missing card is a spade? The pack is well shuffled and the first card is removed and proves to be a spade. What would your assessment of the probability that the missing card is a spade be now? The card removed is now replaced and the pack shuffled. The top card again proves to be a spade. What is your assessment of the probability now? **Solution:** 1/4; 4/17; 48/217. [10]

8. A certain rare disease from which one in ten thousand of the population suffers is diagnosed by a test which reveals the presence of the disease in 95% of the cases of those tested who actually have the disease. However, it also incorrectly yields a positive reaction in 1% of the cases of those who are not suffering from the disease. If a person selected at random from the population shows a positive reaction what is the probability that he is actually suffering from the [15] disease? **Solution:** 95/10094.

28.2.3 Homework 2 (04.03.2004): Estimators, Binomial and Poisson Distributions (75 Points)

1. A player pays a certain sum of money to spin two coins. For two heads he receives back 10 peso, for two tails he receives 2 peso, for a head and a tail he receives nothing. In all four cases he forfeits his stake money. What should [5] the stake money be for the game to be fair? **Solution:** 3 peso.

2. If the probability that a man aged 60 will survive another year is 0.9, what premium should he be charged for a life insurance policy of $1000? (If he [5] survives the year he receives no money back.) **Solution:** $ 100.

3. Two dice are thrown in one "turn", each turn costing 5 peso. If a prize of 40 peso is given for a double 6, and a prize of 20 peso for any other double (together in both cases with the stake money), determine the expected loss to [5] a player who plays the game 100 times. **Solution:** 27.5 peso.

4. The following table gives the number of children in each of 360 families.

No. of children	0	1	2	3	4	5	6	7	8
No. of families	38	91	108	76	39	5	2	0	1

(28.1)

Calculate the mean and standard deviation of the number of children per [10] family. **Solution:** 2.02; 1.28.

5. Two dice are thrown together and the scores added. What is the chance that the total score exceeds 8? Find the mean and standard deviation of the total [10] score. What is the standard deviation of the score for a single die? **Solution:** 5/18; 7; 2.415; 1.71.

6. (This problem is known as the St. Petersburg Paradox). A coin is spun. If a head is obtained first time you are paid $1; If you get a tail followed by a head you receive $2; for two tails followed by a head $4, the next prize being $8 and so on. Show that, however much you are prepared to pay to play the game your expected profit will be positive. Criticize any assumptions you have made and indicate what further knowledge you would require before offering a more realistic "fair price" for the game. If the banker against whom you are playing starts with a capital of $100, what would be a fair price for you to [10] offer him before playing the game? **Solution:** $ 4 favors the player and $ 4.50 favors the banker.

7. A tetrahedral die has its faces colored red, green, yellow and blue. If a group of 8 such dice are thrown, calculate the probabilities of 0, 1, 2, 3, 4 red faces being seen. How many times would you expect to see only 2 red faces if the experiment of throwing the 8 dice was repeated 200 times? [5]

8. A binomial distribution has mean 18 and standard deviation 3. Calculate the probability of 18 successes. [5]

9. In a quality control laboratory, samples of size 60 were examined and the number of defectives counted. Over thousands of trials, the number of defectives never exceeded 14 nor was less than 2. Assuming unchanged probability over the testing period, what was the approximate percentage defective? [5]

10. If telephone calls come into an exchange at an average rate of 70 per hour, find the probability of there being 0,1,2 calls in a period of 2 minutes. What is the probability that there are no calls in a period of 7 minutes. **Solution:** 0.097; 0.226; 0.264; 0.0003. [5]

11. After World War II, an area of South London was divided into 576 squares each of 0.25 square mile area. The number of flying bomb hits in each square was recorded. The results are given in the table below. Compare this with a Poisson model with the same mean. **Solution:** 228, 211, 98, 30, 7, 1. Good correspondence. [5]

Number of hits	0	1	2	3	4	$\geq$ 5
Number of squares	229	211	93	35	7	1

12. A book has errors on 403 out of 480 pages. Assuming that the errors follow a Poisson distribution, find the expected number of pages with just one error. **Solution:** 141. [5]

28.2.4 Homework 3 (18.03.2004): Chi-Squared and Normal Distributions with Central Limits (75 Points)

1. Find the mean, median, mode and variance for the random variable with the pdf.

$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{3}{4}x\,(2-x) & 0 \leq x \leq 2 \\ 0 & x > 2 \end{cases} \tag{28.2}$$

Solution: 1; 1; 1; 1/5. [7]

2. A probability density function of a random variable X is defined as follows:

$$f(x) = \begin{cases} x\,(x-1)\,(x-2) & 0 \leq x < 1 \\ \lambda & 1 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases} \tag{28.3}$$

where λ is a suitable constant. Calculate the expectation value μ of x. What is the probability that $x \leq \mu$? **Solution:** 97/60; 77/160. [7]

3. The line AB has length 10cm. An interval of length 2cm is marked at random on the line, the positions of the interval being uniformly distributed. What is the probability that the interval will contain the midpoint of AB? **Solution:** 1/4 [5]

4. [5] For N(1.6, 2) find a. $P\left(|x| \leq 1.4\right)$ b. P(x is negative) c. P(x $\geq$ 3.2) **Solution:** 0.3934; 0.2119; 0.2119.

5. For N(0, 1) find a value z for which a. P(Z $\geq$ z) = 0.5199 b. P(Z $\leq$ z) = 0.942 c. P(Z $\leq$ z) = 0.3665 **Solution:** -0.05; 1.57; -0.34. [5]

6. [5] A normal distribution $N\left(2, \sigma\right)$ is such that P(X $\leq$ 2.8) = 0.7123. Find σ. **Solution:** 1.43

7. [5] A normal distribution $N\left(\mu, 3\right)$ is such that P(X $\geq$ 1.4) = 0.6179. Find μ. **Solution:** 2.3

8. The IQ's of 500 students are assumed to be normally distributed with mean 105 and standard deviation 12. How many students may be expected:

   (a) to have an IQ greater than 140;

   (b) to have an IQ less than 90;

   (c) to have an IQ between 100 and 110.

   [6] **Solution:** 1; 53; 160.

9. How many degrees of freedom are required for a χ^2 test, for testing data against the following distributions? Where parameters are known independently of the data they are given.

   (a) Binomial, 8 cells, p = 5/8

   (b) Binomial, 10 cells.

   (c) Normal, 12 cells

   (d) Normal, 7 cells, $\mu = 4.6$

   (e) Normal, 15 cells, $\mu = 2.8$, $\sigma = 5.2$

   (f) Poisson, 6 cells, $\lambda = 2.13$

   (g) Poisson 8 cells.

   [10] **Solution:** 7; 8; 9; 5; 14; 5; 6.

10. Use tables to look up the following values. State whether the results are significant at the 5% or 1% levels, or whether they seem too good to be true.

    (a) $\chi^2_4 = 9.60$

    (b) $\chi^2_{11} = 2.51$

(c) $\chi^2_{20} = 26.52$

(d) $\chi^2_{12} = 36.04$

Solution: yes, significant at 5% but not at 1%; too good to be true; not significant; significant at 1%. [10]

11. One hundred observations of a variable x were as listed:

x	0	1	2	3	4	5 or more	
Frequency	8	25	32	14	5	16	Total 100

(28.4)

Is it possible that these observations came from a Poisson distribution with mean 2? **Solution:** $\chi^2_5 = 27.6$, almost certainly not. [10]

28.2.5 Homework 4 (01.04.2004): Sampling and Testing (75 Points)

1. You are to test if a particular coin is fair or not. Suppose your decision rule is to accept the hypothesis of fairness if the number of heads in a single sample of 100 throws is somewhere between 40 and 60 (inclusive) and to reject the hypothesis otherwise.

   (a) Determine the specificity of this decision rule. [5]

   (b) If there are (a) 53 heads or (b) 60 heads in the 100 throws, what conclusions would you draw? [4]

   (c) It is possible that the conclusions of the previous question are not actually correct? Explain. [3]

   (d) Write down a decision rule to test the hypothesis that a particular coin is fair if we take a sample of 64 throws of the coin and use significance levels of (a) 0.05 and (b) 0.01. [7]

   (e) Take a 1 Euro coin and perform 64 throws of it while recording them. Subsequently use the above decision rules to determine whether it is fair or not. Do not take your neighbors numbers but make the effort yourself. If the class has 50 students and each student does this on a different coin, what can we do with the collective data to make a more certain decision? [10]

Solution: Specificity = 0.0358. We must accept the hypothesis in both cases without distinction; we can only add that in the case of 60 heads, only one more head would have forced us to reject the hypothesis. It is possible that we are wrong about this conclusion, this is the meaning of type II error. We design the rule by the z-score method so that z has to be $-1.96 < z < 1.96$ and $-2.58 < z < 2.58$ in the two cases that translates to a number of heads in the ranges $[24.16, 39.84] = [25, 39]$ and $[21.68, 42.32] = [22, 42]$ where the rounded numbers are to be understood as inclusive. General wisdom has it that the 1 Euro coin is not fair; this has already produced significant problems at events like football matches. The increased sample of 50 coins and 50 times more throws decreases the error by the square-root; a desirable effect that allows us to draw the conclusion at a higher confidence limit.

2. An experiment in extrasensory perception (ESP) is performed in which an individual is asked to declare the color of a playing card randomly drawn from a well-shuffled deck of 50 cards by an individual in another room. It is unknown to the tested individual how many red and black cards are in the deck.

   (a) Supposing that the individual correctly identifies 32 cards (of the 50), test the hypothesis at the 0.05 and 0.01 levels. [10]

   (b) The *p-value* corresponding to a statistical test is defined to be the smallest significance level at which the null hypothesis is rejected. Find the p-value for this case. [7]

   Solution: We choose a one-tailed test since we are not concerned with the individual reaching low numbers but only high ones. We test relative to the standard that no powers of ESP means the person chooses at random and thus the likelihood of choosing any one color is 0.5. The z-scores are 1.645 and 2.33 for the two significance levels and the z-score of the measurement is 1.98 and so we accept at 0.05 level and reject at 0.01 level the hypothesis that the person has powers of ESP. The p-value is the probability on the standard normal distribution that z is larger than or equal to 1.98 which is 2.39%. Thus we conclude that the probability that the conclusion that this individual has ESP has a chance of 2.39% of being wrong; in short, it is likely that this individual does have ESP.

3. Using the chi-squared distribution and test, test the following hypotheses.

   (a) A coin is tested for fairness and thrown 200 times; there were 115 heads and 85 tails. Test the hypothesis of fairness at the 0.05 and 0.01 significance levels [7]

   (b) A pot contains a large number of marbles of red, orange, yellow and green colors. We draw out 12 of these marbles at random and get 2 reds, 5 oranges, 4 yellows and 1 green. The hypothesis is that the pot contains an equal proportion of marbles of all four colors. [11]

   (c) 320 families with 5 children are surveyed and the sex of the children asked for. The results are shown in table 28.1 and the hypothesis is that men and women have equal birth probability. [11]

   Solution: The coin has $\chi^2 = 4.5$ and the problem has one degree of freedom. The critical values for χ^2 at the 0.05 and 0.01 significance levels for a single degree of freedom are 3.84 and 6.63 and thus we reject it at the 0.05 level and accept it at the 0.01 level. We would have gotten the same answer using z-score methods as the experimental z-score is 2.12 and so larger than 1.96 but less than 2.58.

   With the urn, we would expect to see three of each kind in a sample of 12. However as the expected numbers are less than five, the chi-squared distribution test is not applicable. Thus we must combine the categories. We shall thus test whether red-green is equally proportioned to orange-yellow. Thus

b's and g's	5-0	4-1	3-2	2-3	1-4	0-5	Total
families	18	56	110	88	40	8	320

Table 28.1: The child distribution over 320 families of 5 children. (b = boys, g = girls)

$\chi^2 = 3$ and there is one degree of freedom. From the same critical values, we must accept the hypothesis at the 0.05 level.

We use the binomial distribution to compute the probabilities of each kind of family if the birth probability is 0.5 for each sex. The result is 1, 5, 10, 10, 5, 1 out of 32 in order of the table. Thus $\chi^2 = 12.0$ and there are five degrees of freedom and we reject the hypothesis at the 0.05 level but not at the 0.01 level. We therefore conclude that the birth probabilities are in all likelihood not equal.

28.2.6 Homework 5 (22.04.2004): Maximum Likelihood and Correlation (100 Points)

1. A mobile phone company conducted a survey of mobile ownership among different age groups. The results for 1000 households are shown in table 28.2. Test the hypothesis that the proportions of mobile ownership are the same for the different age groups (use contingency tables for chi-squared). Briefly comment on the sampling strategy of this survey in the light of these results. [11]
   Solution: We adopt the estimate of 25% per age group as a reasonable guess at the same proportion in each group. Based on this we compute $\chi^2 = 14.3$ using contingency tables. There are 3 degrees of freedom. At the 0.05 significance level the critical χ^3 is 7.81 and thus we reject the null hypothesis that the proportions are equal. Choosing the same number of people in each age group was decidedly unwise as there are many more people in the population with ages in 25-54 than in 18-24 and so this is not very meaningful, i.e. the group 18-24 is given more weight than it deserves.

2. Using linear least square methods fit a straight line to the data given in table 28.3. **Solution:** NEED DATA WITH ERRORS IN BOTH COORDINATES. [11]

3. Using three data sets, you are asked to fit lines and make certain predictions.

   (a) Farm real estate values in the USA is given in table 28.4. Find the least squares straight line for this data. Estimate the value of farm real estate in 1988 and 1996 and compare your values to the true values of 626.8 and 859.7 billion US Dollars. [11]

   (b) Purchasing power of the US Dollar is given in table 28.5. Find the least square straight line. Predict the power for 1998 assuming that the trend continues. [10]

Mobiles	18-24	25-54	55-64	≥65	Total
Yes	50	80	70	50	250
No	200	170	180	200	750
Total	250	250	250	250	1000

Table 28.2: Mobile phone distribution over age groups.

x	1	2	3	4	5	6	7	8	9	10
Δx	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
y	1.1	1.5	2.1	2.4	3.0	3.7	3.8	4.4	5.2	5.6
Δy	0.1	0.2	0.2	0.1	0.1	0.2	0.1	0.2	0.2	0.1

Table 28.3: Data for straight line fit with errors in both coordinates.

(c) The population of the USA is given in table 28.6. Find the least squares straight line as well as the least squares parabola to this data and comment on which fits best. Use both models to predict the population in the year 2000. [10]

Solution: All computations are basic. For the farms we find $y = 23.061x - 45222.914$ and the predictions are 621.8 and 806.27. For the purchasing power we find $y = -0.0289x + 0.989$ and the prediction is 0.556. For the population we obtain $y = 12x + 155$ and $y = -0.178x^2 + 13.6x + 153$ where we have used x to number the years starting with 0 for 1950 and going up to 9 for 1995. The sum of the squares of the residuals for the line is 30.024 and for the parabola 13.289 and thus the parabola is a better fit. We predict 275 and 271.2 respectively.

4. The number of students in a class of 100 who got certain grades in a mathematics and a physics exam are given in table 28.7. How correlated are these grades? Find a numerical measure of correlation and discuss. [10] **Solution:** The coefficient of linear correlation $r = 0.7686$ and we conclude that given one grade we can predict the other grade with a good accuracy. If we are interested in the rough placement of a student's abilities, one of the two tests would be enough as they measure 77% of the identical ability, i.e. doing the second test only supplies 23% additional information and this is not worth the effort.

Year	1989	1990	1991	1992	1993	1994	1995
Value	660.0	671.4	688.0	695.5	717.1	759.2	807.0

Table 28.4: Data giving the total value of USA farm real estate in billions of US Dollars. U.S. Department of Agriculture, Economic Research Service.

Year	1983	1984	1985	1986	1987	1988	1989
Price	1.003	0.961	0.928	0.913	0.880	0.846	0.807
Year	1990	1991	1992	1993	1994	1995	1996
Price	0.766	0.734	0.713	0.692	0.675	0.656	0.638

Table 28.5: Data giving the purchasing power of the US Dollar as measured by consumer prices according to the U.S. Bureau of Labor Statistics, Survey of Current Business.

Year	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995
Population	152	166	181	194	205	216	228	238	250	263

Table 28.6: Data giving the population of the USA in millions. U.S. Bureau of Census.

Physics				Math			
	40–49	50–59	60–69	70–79	80–89	90–99	Total
90–99				2	4	4	10
80–89			1	4	6	5	16
70–79			5	10	8	1	24
60–69	1	4	9	5	2		21
50–59	3	6	6	2			17
40–49	3	5	4				12
Total	7	15	25	23	20	10	100

Table 28.7: Number of people of a class of 100 who receive a certain grade on a physics and a mathematics exam.

28.2.7 Homework 6 (06.05.2004): Markov Processes (100 Points)

1. The Daley-Kendall (1965) model for how a rumour might spread in a closed, homogenously mixing population splits the members into three classes: (a) ignorants, who do not know the rumor, (b) spreaders, who know the rumor and are actively spreading it and (c) stiflers, who know the rumor but have stopped spreading it. When a spreader A meets any person B, there is a chance A will start to tell the rumor. If B is an ignorant, B learns the rumor and becomes a spreader while A continues to be a spreader. But if B knows the rumor, either as a spreader or as a stifler, B tells A that the rumor is stale news, and *both* become stiflers. Set up this process as a Markov chain and analyze it in parallel with the epidemic model (example 8.8 in the photocopies) as far [33] as you can. Is there a threshold phenomenon? **Solution:** Keeping as close to the epidemic model as we can, let $X(t)$, $Y(t)$ and $Z(t) = N - X(t) - Y(t)$ be the numbers of ignorants, spreaders and stiflers. Given $(X(t), Y(t)) = (x, y)$, the non-zero transition rates out of that state are to $(x-1, y+1)$ at rate βxy, to $(x, y-2)$ at rate $\beta y(y-1)/2$ and to $(x, y-1)$ at rate $\beta y(N-x-y)$. Here β is related to the rate of mixing and the likelihood a spreader begins to tell the rumor; there is thus no threshold. For an exact analysis via random walks, the transition rates are clear. Taking the deterministic equations as with epidemics, $dx/dt = -\beta xy$ and $dy/dt = \beta(xy - y(y-1) - y(N-x-y))$, giving $dy/dx = (N-2x-1)/x$. Taking $y = 1$ when $x = N-1$, integrate to obtain $y = (N-1)\ln(x/(N-1)) - 2x + 2N - 1$. The epidemic is over when $y = 0$; take N large, and write $u = x/(N-1)$. This leads to $2u = \ln(u) + 2$, approximately, i.e. $u = 0.2032$. The rumor is expected to spread to about 80% of the population.

2. Consider a town of 30,000 families which have been divided by the local chamber of commerce for planning purposes into three economic brackets (or classes): lower, middle, and upper. A builder wants to know the future populations of these brackets so they can decide what types (read: cost) of houses to build. The city also wants this information to help decide other issues like taxes and social services. The past year the flow between these populations was the following:

   (a) 20% of the lower move into the middle.
   (b) 10% of the middle move back to the lower.
   (c) 10% of the middle move to the upper.
   (d) 15% of the upper move down to the middle.
   (e) 5% of the lower move directly into the upper.
   (f) 4% of the upper move down to the lower.

   Last year there were 12,000 lower, 10,000 middle, and 8,000 upper income families. Write down the transition matrix for a Markov process for this example and calculate the steady state of the system and the minimum number of years [34] required before the steady state is reached. **Solution:** The transition matrix is

$$P = \begin{pmatrix} 0.81 & 0.15 & 0.04 \\ 0.10 & 0.80 & 0.10 \\ 0.05 & 0.20 & 0.75 \end{pmatrix} \tag{28.5}$$

We calculate the eigenvectors corresponding the unity eigenvalue and multiply it by the total population of 30,000 to obtain the steady state response of (,). Simple iteration shows that this is reached after .... years.

3. A gas station has four pumps. Cars arrive at random at an average rate of three per minute and service times are exponential with a mean of two minutes. Give the parameters for the queue to which this corresponds; what happens in the long run?

   Suppose the forecourt has room for six cars only, in addition to any at the pumps. Find the equilibrium distribution and estimate how many cars per hour drive on, finding the forecourt full.

   Each car that drives on represents lost profit. Consider the relative merits of (a) extending the forecourt to take two more cars; (b) installing one more pump, which reduces the forecourt capacity to five waiting cars. **Solution:** [33] Per hour, the queue is $M(180)/M(30)/4$. Here $k\mu = 120 < 180 = \lambda$, and so the queue would grow without bound. With the given limit, it is $M/M/4/10$, so equation 8.10 (in the photocopies) implies $\pi_n = \lambda^n \pi_0/(n!\mu^n) = 6^n\pi_0/n!$ when $n \leq 4$, and $\pi_n = \pi_4(3/2)^{n-4}$ for $4 \leq n \leq 10$. This yields $\pi_0 = 1/1798.28$, hence $\pi_1, \cdots, \pi_{10}$. $P(\text{Forecourt full}) = \pi_{10} = P(\text{Arrival drives on})$, so mean number lost per hour is $180\pi_{10} \approx 61.6$.

   (a) leads to $M/M/4/12$, from which we find $\pi_{12} = 0.3371$, and (b) is $M/M/5/10$, now $\pi_n = 6^n\pi_0/n!$ for $n \leq 5$, and $\pi_n = \pi_5(6/5)^{n-5}$ for $5 \leq n \leq 10$. Then $\pi_{10} = 0.2126$, so the respective forecourt losses are about 60.7 and 38.3. Thus an extra pump is much better than extending the forecourt.

Chapter 29

Projects

In addition to the homework, you are asked to complete a project. The average of all homework grades will count for fifty percent of the final grade whereas the other fifty percent will come from the project. The project takes the form of a written report about a topic on the list below. The report must be typed and submitted by email. The LaTeXformat is much preferred but MS Word will also be accepted. There is an absolute upper limit on the length of the report of 5000 words (data and equations *not* counted) but please do not feel that you have to write almost that many words. If you can make your point in significantly less words, that is very good.

The report must contain but is not limited to:

1. An introduction in which it is made clear what the task (hypothesis) is, why it is interesting/useful and how you have chosen to go about it,

2. A body in which you give the data used for the report and any calculations made,

3. A discussion in which the data and the results are discussed with respect to the task given in the introduction,

4. A conclusion in which you sum up what was done and

5. A bibliography in which you reference your sources.

Every report must contain some data *obtained from a reliable source.* The fact that the source is reliable must be argued in the report. Should you prefer to collect the data yourself, that is great and, in this case, you need not argue for your own reliability. In addition, the data must be *significant* and *representative.* Both of these properties must be addressed in the report.

It is estimated that this project will not require more than a weekend's worth of your time. You are encouraged to *start work early so that you might distribute your load (from this course and others) over the semester.* The report is due in *at latest two weeks before the end of the semester.*

A list of possible projects follows. If you have an idea which is not listed, please contact me to see if it would be suitable.

1. Investigate Zipf's law of least effort in one of the following circumstances:

(a) Words in books (any language is fine for this)

(b) Population of cities

(c) Website hit counts

(d) Salaries of individuals

(e) Market capitalization of public listed companies

(f) Size and/or Strength of natural disasters (earthquakes, hurricanes, etc.)

(g) Prizes of consumer goods

(h) Flares on the sun

(i) Anything else you can think of

Note that the law is effectively a social scientific law and thus not so precise. It says that humans will do whatever they have to do with the least effort possible. Thus (and this "thus" requires quite some arguing) if we have a list of items that occur with a given frequency (number of occurrences of that item divided by the total number of occurrences) sorted in order of decreasing frequency, then the rank r and frequency f of this list should be related by $f = br^a$ where a and b are some constants to be found. Note that, in spirit, this law is identical to the "action principle" at the basis of physics.

2. Argue for or against the following hypotheses (note that some are deliberately provoking and do *not* necessarily reflect my own opinion)

   (a) The air-plane is the safest method of transportation.

   (b) If no radically new technologies are found, the world will be overpopulated in year x meaning that we will no longer be able to grow enough food to feed the population. The year must be found.

   (c) Global warming due to greenhouse gasses is a reality.

   (d) The most commonly owned electronic device on the planet is the television.

   (e) The rich get richer and the poor get poorer.

   (f) Globalization is a negative trend in economics from the point of view of the average individual.

   (g) History is written by the victors.

   Note that most of these hypotheses were vague and in your report, you must define the words so as for them to become precise statements! Also note that while these topics are of everyday relevance, the report should discuss them from the standpoint of the data available using probabilistic and statistical methods.

3. Choose a game of chance (Blackjack, Poker, Dice, Roulette, etc.) and prove that the house will win in the long run by deriving the probability that a player will loose at any stage of said game. Demonstrate a strategy of "cheating" by which the player may increase his chances of winning so substantially that

playing the game will become a realistic source of profit for him (for example by counting cards in Blackjack or Poker).

In this report you must first explicitly define the game, find the probability that a player will loose and argue that in the long run, the house profits. Then you are to give complete details of a method of winning and find the new probability of loosing and argue that thus the player will profit.

4. Apply the theory of Markov processes to the study of financial indices. Choose an index (Dow Jones, DAX, etc. or also a consumer index) and define it and the method of its calculation in detail. Derive a Markov process model for this index by making economical assumptions that must be argued for. Using data up to a certain arbitrary date available on the Internet for this index, calculate the necessary model parameters. Use the data after the chosen date to check the prediction of the Markov process and show for what period of time your model is accurate. As a benchmark, a model would be considered good (for the purposes of this report) if it is accurate to within 5% for at least one week (for a market index) in which no dramatic political event disturbed the evolution of the index.