PROBABILITY THEORY FOR STATISTICAL METHODS

'On this occasion, I must take notice to such of my Readers as are well versed in Vulgar Arithmetick, that it would not be difficult for them to make themselves Masters, not only of all the practical rules in this Book, but also of more useful Discoveries, if they would take the Small Pains of being acquainted with the bare notation of Algebra.'

A. DE MOIVRE

Preface to *The Doctrine of Chances*, 1718

PROBABILITY THEORY FOR STATISTICAL METHODS

BY

F. N. DAVID

CAMBRIDGE
AT THE UNIVERSITY PRESS
1951

PUBLISHED BY
THE SYNDICS OF THE CAMBRIDGE UNIVERSITY PRESS

London Office: Bentley House, N.W.1
American Branch: New York

Agents for Canada, India, and Pakistan: Macmillan

First Edition 1949
Reprinted 1951

Printed in Great Britain at the University Press, Cambridge
(Brooke Crutchley, University Printer)

To

PROFESSOR JERZY NEYMAN

PREFACE

For some years it has been the privilege of the writer to give lectures on the Calculus of Probability to supplement courses of lectures by others on elementary statistical methods. The basis of all statistical methods is probability theory, but the teacher of mathematical statistics is concerned more with the application of fundamental probability theorems than with their proof. It is thus a convenience for both teachers and writers of textbooks on statistical methods to assume the proof of certain theorems, or at least to direct the student to a place where their proof may be found, in order that there shall be the minimum divergence from the main theme.

This treatise sets out to state and prove in elementary mathematical language those propositions and theorems of the calculus of probability which have been found useful for students of elementary statistics. It is not intended as a comprehensive treatise for the mathematics graduate; the reader has been envisaged as a student with Inter.B.Sc. mathematics who wishes to teach himself statistical methods and who is desirous of supplementing his reading. With this end in view the mathematical argument has often been set out very fully and it has always been kept as simple as possible. Such theorems as do not appear to have a direct application in statistics have not been considered and an attempt has been made at each and every stage to give practical examples. In a few cases, towards the end of the book, when it has been thought that a rigorous proof of a theorem would be beyond the scope of the reader's mathematics, I have been content to state the theorem and to leave it at that.

The student is to be pardoned if he obtains from the elementary algebra textbooks the idea that workers in the probability field are concerned entirely with the laying of odds, the tossing of dice or halfpennies, or the placing of persons at a dinner table. All these are undoubtedly useful in everyday life as occasion arises but they are rarely encountered in statistical practice. Hence, while I have not scrupled to use these illustrations in my turn, as

soon as possible I have tried to give examples which might be met with in any piece of statistical analysis.

There is nothing new under the sun and although the elementary calculus of probability has extended vastly in mathematical rigour it has not advanced much in scope since the publication of *Théorie des Probabilités* by Laplace in 1812. The serious student who wishes to extend his reading beyond the range of this present book could do worse than to plod his way patiently through this monumental work. By so doing he will find how much that is thought of as modern had already been treated in a very general way by Laplace.

It is a pleasure to acknowledge my indebtedness to my colleague, Mr N. L. Johnson, who read the manuscript of this book and who made many useful suggestions. I must thank my colleague, Mrs M. Merrington for help in proof reading, and the University Press, Cambridge, for the uniform excellence of their type setting. Old students of this department cannot but be aware that many of the ideas expressed here have been derived from my teacher and one-time colleague, Professor J. Neyman, now of the University of California. It has been impossible to make full acknowledgement and it is to him therefore that I would dedicate this book. Nevertheless, just as probability is, ultimately, the expression of the result of a complex of many factors on one's own mind, so this book represents the synthesis of different and often opposing ideas. In brief, while many people have given me ideas the interpretation and possible distortion of them are peculiarly mine.

F. N. DAVID

DEPARTMENT OF STATISTICS
UNIVERSITY COLLEGE, LONDON

CONTENTS

CHAPTER I

FUNDAMENTAL IDEAS

It has become customary in recent years to open expositions of probability or statistical theory by setting out those philosophical notions which appear to the author to underlie the foundations of his mathematical argument. Unfortunately 'as many men, so many opinions' and the unhappy student of the subject is often left bogged in the mire of philosophical disquisitions which do not lead to any satisfactory conclusion and which are not essential for the actual development of the theory. This does not imply, however, that they are not necessary. It is true that it is possible to build up a mathematical theory of probability which can be sufficient in itself and in which a probability need only be represented by a symbol. If the building of such a framework were all that was required then speculations and theories would be unprofitable, for there can be no reality in mathematical theory except in so far as it is related to the real world by means of its premises and its conclusions. Since the theory of probability attempts to express something about the real world it is clear that mathematics alone are not sufficient and the student needs must try to understand what is the meaning and purpose of the logical processes through which his mathematical theory leads him; for the anomalous position obtains to-day in which there are many points of view of how to define a probability, and as many more interpretations of the results of applying probability theory to observational data, but the actual calculations concerned are agreed on by all.

Fundamentally the term *probable* can only apply to the state of mind of the person who uses the word. To make the statement that an event is probable is to express the result of the impact of a complex of factors on one's own mind, and the word *probable* in this case will mean something different for each particular individual; whether the statement is made as a result of numerical calculation or as a result of a number of vague general impressions is immaterial. The mathematical theory of probability is concerned, however, with building a bridge, however inadequate it

may seem, between the sharply defined but artificial country of mathematical logic and the nebulous shadowy country of what is often termed the real world. And, descriptive theory being at present in a comparatively undeveloped state, it does not seem possible to measure probability in terms of the strength of the expectation of the individual. Hence, while the student should never lose sight of the fact that his interpretation of figures will undoubtedly be coloured by his own personal impressions and prejudices, we shall restrict the meanings of *probable* and *probability* to the admittedly narrow path of numbers, and we shall consider that the work of the probabilist is complete when a numerical conclusion has been reached. We shall thus transfer a probability from the subjective to the objective field.

In making such a transference, however, we do not escape the main question 'What do we mean by a probability?' although we possibly make it a little easier to answer. In the field of statistical analysis there would seem to be two definitions which are most often used, neither of which is logically satisfying. The first of these theories we may call the mathematical theory of arrangements and the second the frequency theory. Since it is by the light of these two theories that probabilities are generally interpreted, it will be useful to consider each of these in a little detail. The mathematical theory of arrangements possibly is as old as gaming and cardplaying; certainly the idea of probability defined in such a way was no secret to Abram de Moivre (*Doctrine of Chances*, 1718), and it is in fact the definition which everyone would tend automatically to make. For example, suppose that the probability of throwing a six with an ordinary six-sided die is required. It is a natural proceeding to state that there are six sides, that there is a six stamped on only one side and that the probability of throwing a six is therefore 1/6.

The probability of an event happening is, in a general way, then, the ratio of the number of ways in which the event *may* happen, divided by the total number of ways in which the event may or may not happen. As a further illustration we may consider the throwing of a coin into the air. When the coin falls there are two possible alternatives, the head may be uppermost or the tail. If the probability of throwing a head is required then out of the two possible alternatives there is only one way in which

a head may be obtained and the required probability, according to this theory, is therefore $\frac{1}{2}$.

So far there is perhaps little need for comment. A mathematical set of arrangements Ω is postulated; from the set Ω the subset ω of arrangements in which the event may happen is picked out, and the ratio of the subset to the complete set of arrangements is the defined probability. This probability is exact in the mathematical sense, just as is the constant π, but it is without meaning. The statistician now takes over and states that in saying the probability of an event is p, where p is the ratio of favourable to total arrangements, it will mean that if the experiment is carried out on a very large number of occasions, under exactly similar conditions, then the ratio of the number of times on which the event actually happened to the total number of times the trial was made will be approximately equal to p, and this ratio will tend more closely to be p as the number of trials is increased.

It is the definition of probability by mathematical arrangements and its justification by means of repeated trials which we usually accept for lack of anything better, but it is well to realize that the interpretation of probability defined in this way is open to objection. We have said that probability theory is an attempt to bridge the gap which lies between mathematical theory and observational reality. It follows therefore that any justification of the theory should be based on this so-called reality whereas it quite obviously is not. In no series of trials in the so-called real world can experiments be made under exactly similar conditions. If the trials are made by one person then they must differ in at least one aspect, time, and if they are carried out at the same instant of time then they must differ in that they are performed by entirely different people. It is certain that the conditions as stated can never obtain in the real world and that in a strict logical sense therefore the bridge between theory and practice seems impossible to construct. It must, however, be stated that in practice, provided care is taken in the experimental conditions, and provided the number of experiments is not large enough for the effect of wear in the experimental apparatus to become apparent, the ratio of the number of successes to the total number of trials does approximate to that stated in the mathematical

model, and it is considered that this is a justification of the mathematical framework.

The writer feels, however, that the mathematical model should be subjected to closer scrutiny. Suppose that we imagine a mathematician who has spent all his life in one room and who has had no opportunity to observe the real world outside. Such a mathematician could build a mathematical theory of probability just as could another mathematician who had contact with reality, and he would be able to state the probability of throwing a six with a die or of a head with a halfpenny simply by idealizing the die or halfpenny which had been described to him. For any mathematical argument and conclusion follows logically from the premises on which it is based, and both our mathematicians' premises will be based on the simplification of the die or the halfpenny.

This much is certain, but what is not so certain is that these premises would be the same for the two persons. The imaginary mathematician might postulate a set of arrangements in which a weight of one was given to one, a weight of two to two, and so on making a total set of 21. From this he could go on to postulate that the probability of a six was 6/21, and within the narrow framework of his own postulates he would be correct. Or he might postulate three cases for the halfpenny, one for heads, one for tails, and one for the occasion on which the halfpenny stands upright. Again, whatever conclusions he drew regarding the probability of heads would be correct with reference to the set of arrangements which he had postulated.

The mathematician of the real world would certainly act differently in that he would have no hesitation in stating that the total number of arrangements for the die would be 6 and for the halfpenny 2; but why would he do so? It could only be because previous experience had taught him, either from a study of applied mathematics or from gaming itself, that one side of the die was as likely to turn up as any other. In other words, unconsciously perhaps, he would choose for his fundamental set of arrangements that set in which the alternatives were equally probable, and in so doing he would be guilty of circular reasoning.

That the mathematical theory of arrangements will always lead to such an impasse appears more or less inevitable. Suppose for example that the real mathematician is confronted with a

problem in which he had no means of being able to foretell the answer, a state of affairs which is more frequent than not in statistical practice. He must therefore build his fundamental set of arrangements as best he may, but since the acid test of theory is experiment he would undoubtedly carry out experiments to test the adequacy of his mathematical set-up. Now if experiment showed that one of his alternatives occurs with twice or thrice the frequency which he had allowed, he would then alter his fundamental set so that his theoretical probability and the probability as estimated from practice were more or less in agreement. And, in so doing, he would again be arguing in a circular way and his theoretical definition would have little validity when applied to a practical problem.

The writer would suggest therefore that although the mathematical theory of arrangements is exact on the theoretical side, it is inadequate when the link between theory and practice is attempted and that the stumbling block of circular reasoning which lay in the path of Laplace and subsequent nineteenth-century writers has not really been eliminated. Some probabilists have shown themselves to be aware of this and have attempted a definition of probability not very different from that which we have already given as the connecting bridge between the mathematical theory of arrangements and observation. The frequency definition is commonly given as follows: 'If in a series of independent and absolutely identical trials of number n the event E is found to occur on m occasions, the probability of E happening is defined as the limit of the ratio m/n as n becomes very large.' We have already noted the objections which might be raised against this definition in that the conditions 'absolutely identical' are impossible to satisfy and that as n increases the ratio does not tend to a definite limit for the effect of wear on the apparatus is not inconsiderable. In practice this frequency definition does seem to work over a limited range, but it is difficult to fit into a mathematical scheme, and is therefore skirted rather warily by mathematicians and statisticians alike.

A definition along the lines put forward for other statistical parameters by J. Neyman and E. S. Pearson might seem to hold a promise of more validity, although the pitfalls to be met in pursuing such a course are many. We may begin with the idea

that there is a population of events, and that there exists a population parameter, constant and fixed, which describes these events, and which we may call a population probability. Such a probability would bear no resemblance to the probabilities which we have just discussed. For example, if we wished to know the population probability of a Chelsea Pensioner who attains the age of 70 in 1945, dying before he reaches the age of 71, we could by waiting until the end of 1946 find out how many Chelsea Pensioners attaining age 70 in 1945 had died before reaching age 71, and define our population probability as just that proportion of Pensioners who had died out of the total number exposed to risk.

Hence a population probability is just the ratio of the number of times which the event has happened divided by the total number of events if the population is finite. There are, however, many populations in statistics which are not capable of being enumerated in this way. For instance the population of the tossing of a die or of the throwing of halfpennies will never be completed. Nevertheless we shall postulate that these populations can be described by a constant parameter, or a population probability, which experience has shown to be equal to a certain value, and which, if an infinite population were capable of being enumerated, would be equal to the proportion of successes in the total population.

Following along the lines of statistical practice we have therefore an unknown population parameter, our population probability p, which we postulate exists, and which we desire to estimate. We perform a series of experiments and from these experiments we derive an estimate of p. For example, we might throw a die n times and count the number of times, x, that a six fell uppermost. It is clear that the mean value of x/n in repeated sampling will be approximately equal to p, and if it were possible to carry out an infinite series of trials, in each of which the die was thrown n times, the mean value would be exactly equal to p.

Let us turn for illustration to the case of our (hypothetical) Chelsea Pensioners. It is desired to know the probability of a Pensioner who attains age 70 in the year 1945 dying before he is 71 years of age. As has been pointed out, it would be possible to wait until the end of 1946 and calculate the exact proportion of

Pensioners who have died before reaching 71. This is equivalent to stating that the desired population probability will exist at the end of the calendar year but that at the present time it does not exist because it is not known which of the individuals possesses the required characteristics of dying or not dying. In order therefore to estimate the probability of a Pensioner attaining age 70 in the year 1945 dying before he is 71 years of age it will be necessary to postulate a hypothetical population of Pensioners attaining age 70 in 1945 of whom a fixed proportion may be expected to die before reaching 71. If we now choose a number of other years in which conditions of living were reasonably the same, and calculate the proportion of Pensioners who satisfied the required conditions, we may regard the proportions thus obtained as estimates of the unknown population parameter and from combining them we may make an estimate of which can be stated to lie within certain limits.

Now, if we take this estimated value of p as the probability that a Pensioner attaining age 70 in the year 1945 will die before he reaches the age of 71, we shall not expect the value of p actually calculated from the 1945–6 figures to have necessarily the same value. For the 1945–6 figure will be the exact value for that particular year, but will itself also be an estimate of the chances of death in any year for a Pensioner of the stated age. Hence if we alter our question a little and ask what are the chances of death before reaching the age of 71 for a Pensioner who attains age 70 in the year Y, then the addition of the 1945–6 data should give increased accuracy to our estimate of the unknown probability and should enable closer limits to be obtained for this probability provided all the known causes which might cause fluctuations are controlled for each of the years considered.

This control of causes of variation is important and it may be well to digress a little from the main theme in order to consider what is meant by it. In any set of figures, whether obtained directly by experiment or collected from records, there will be variations both between the individual sets and from some values which might be expected by hypothesis. It is a commonplace to state that before the collection of material is begun, no matter what the material may be, all known causes of variation should be eliminated or at least controlled. Such known causes are often

spoken of as assignable causes because we are able to state definitely that they would cause variation unless they were controlled. Usually in statistical method the aim of the compiler of figures is to eliminate such sources of variation, but there are occasions on which assignable variation is too small to influence results, or it would cost too much in time and money to eliminate it. In such cases the assignable variation remains in the material but it is necessary to remember it when interpreting any numerical results.

Thus in discussing the chances of death of a Chelsea Pensioner it is clear that we should use for our estimate of probability only those figures relating to other Pensioners of the same age and that for each year studied we should take care that as far as was possible all other conditions were the same. By so doing we should ensure that each set of figures gave an estimate of the same unknown hypothetical population probability.

After all the assignable causes have been controlled the result of any one experiment or collection of figures is still subject to variation from causes which we do not know about and therefore cannot control. It is these unassignable causes, or as they are more often called, random errors, which create the need for the concept of probability. A penny is tossed in the air. If care is taken in the spin of the coin imparted by the finger and thumb, and if a soft cushioned surface is chosen for the penny to fall upon, then it is often possible to determine beforehand whether it will fall with the head or the tail uppermost. That is to say by controlling certain sources of variation the fall of the penny can be predicted.

On the other hand, if no care is exercised in the tossing and in arranging the fall of the penny it is not possible to predict which way up it will fall even if the experiment is carried out a number of times. Random errors due to unassigned causes determine whether the penny shall fall head or tail uppermost and, as Borel has written, 'the laws of chance know neither conscience nor memory'. All that we can know is that if we carry out a series of trials and estimate a probability from each, these estimates will vary about an unknown population probability, that these estimates can be used to estimate this unknown population parameter and the limits within which it may lie, and that previous

experience has shown that if the errors are really random then the mean value of estimates from a number of series of trials will approximate to a constant number.

What we have postulated therefore for our definition of probability is that the population probability is an unknown proportion which it is desired to estimate. This parameter may only be estimated by a series of experiments in each of which the same assignable causes are controlled and in each of which, as far as is possible, no new cause of variation is allowed to enter. The result of each experiment may be regarded as an estimate of this unknown proportion and the pooling of these estimates enables prediction to be made for any new set of experiments it is desired to carry out. We have not postulated absolutely identical conditions for each of a set of experiments. Experience has shown this to be unnecessary and provided the more obvious sources of variation are controlled the different probability estimates will vary about the unknown true value. We may discuss the nature and size of this variation at a later stage.

The advantages to be gained by defining a probability in this way would seem to the writer to be many. For example, such a definition will fit more closely into statistical practice than does (say) the mathematical theory of arrangements. It is rare in statistical practice to be able to state the alternatives of equal weight, such as one is able to do with the six faces of a die; in fact generally it is necessary to find a probability by evaluating the ratio of the number of successes to the total number of trials. Under the scheme which we have just set out this would be recognized for what it is; an estimate of the unknown probability and an estimate which will undoubtedly be different from that which will be obtained when more evidence renders another calculation possible. Further, if it is shown by experiment that several alternatives are approximately equally possible as in the case of the six faces of the die or the two sides of a penny, then there appears to be no reason why a mathematical model based on equi-probable alternatives should not be constructed if it is so desired. But the mathematical model can only be established after experiment has shown its possible construction, although such a construction will be valuable in certain mathematical applications.

The interpretation of a probability will follow directly from the definition which we have given. If as the result of calculations we have found that the probability of a given event is p, then we should say that if a series of experiments were carried out in which all assignable causes of variation were controlled and no further large assignable or unassignable causes of variation were permitted to intervene, then the mean value in repeated experiments of the proportion of times in which the event occurred, will be approximately equal to p. As an illustration, suppose that the results of experiments tell us that the probability of a tank crossing a minefield without detonating a mine is 0·86. This would mean that the average number of tanks crossing unscathed out of every 100 attempting to cross the minefield would be 86. We should not be able to say which tank would be blown up, nor what the exact proportion of tanks crossing unscathed in any given 100 would be, but we should feel sure that the average proportion of successes, if a series of trials could be made, would approximate to 0·86.

We began by stating that by producing a numerical probability the work of the probabilist should be considered as finished and that of the interpreter begun and perhaps the illustration last used of tanks crossing a minefield may be helpful in emphasizing what is meant by this. To tell a worker in pure probability theory that the chance of a tank crossing a minefield unscathed is 0·86 would be to convey very little information to him unless he also happened to be a tank specialist. He may perhaps reply 'Good' or 'Not so good' according to whether he regarded 0·86 as a high or not so high number, but almost certainly his reply would be as the result of the impact of the pure number on his mind and little else.

On the other hand, to tell a general in charge of armoured troops that the probability was 0·86 would provoke an instant response. If he had plenty of tanks and if advantage was to be gained by a swift crossing then he might regard 0·14 as an acceptable risk and order his armour to attempt the crossing. On the other hand if tanks were few and pursuit not profitable then he might regard 0·14 as not acceptable. In this case, as in every other interpretation of probability, the numerical result is only one of many factors which have to be taken into account in reaching

a decision. Generally these other factors are not capable of numerical expression or they would have been included in the probability calculation. For the experimentalist turned statistician it is often possible for the role of probabilist and interpreter to be combined, for only the person who has collected the material can know its exact worth in interpretation; but the professional statistician, *per se*, may only calculate a numerical probability and must perforce leave the interpretation of probability in the shape of decision for action in the hands of someone else.

REFERENCES AND FURTHER READING

The definition of probability according to the mathematical theory of arrangements may be found in almost any text-book of *Higher Algebra for Schools*. The question of an interpretation of such a probability is not often discussed.

An excellent discussion of the definition and interpretation of probabilities is given by H. Levy and L. Roth, *Elements of Probability*, Chapters I, II and III. These writers choose the definition of probability according to the theory of mathematical arrangements. No detailed statistical exposition of their results is given.

The student may benefit from a reading of R. v. Mises, *Probability, Statistics and Truth*, which will introduce him to the justifications which may be advanced in support of the frequency theory definition.

Different points of view, not discussed here, will be found in J. M. Keynes, *A Treatise on Probability* and H. Jeffreys, *Theory of Probability*.

CHAPTER II

PRELIMINARY DEFINITIONS AND THEOREMS

In the previous chapter we have been concerned with the definition and interpretation of what is meant by a probability. In this and succeeding chapters the objective will be the setting out of the definitions and rules whereby we may build up a theory for the addition and multiplication of probabilities. The actual theory will be mathematical and there will be no need to interpret it in the terms of the world of observation until the logical processes of the mathematics are complete and a numerical answer is reached. It is useful to note that the theory which we shall set out will be applicable to all numerical theories of probability; in fact the only difference between any of the numerical theories of probability will lie in the definition of what is meant by a probability and the interpretation of statistical calculations by the help of such a probability.

We begin by defining the Fundamental Probability Set. The fundamental probability set, written F.P.S. for short, will be just that set of individuals or units from which the probability is calculated. Thus if it is necessary to estimate a probability from a series of n observations, then these n observations will be the F.P.S. Or if sufficient experimental evidence is available to justify the setting up of a mathematical model in the shape of the mathematical theory of arrangements, then the F.P.S. would be the total number of arrangements specified by the theory. In the case of a die the F.P.S. given by the mathematical theory of arrangements would have six members, but if experiment had shown that the die was biased in some way, and it was necessary to estimate a probability from the observations, then the F.P.S. would contain the total number of throws of the die which were recorded. It is unnecessary to labour the point unduly, but it should be noted that the elasticity of the definition does, as we have stated above, render the subsequent theory independent of whatever definition of probability is used.

In order to keep the theory in a generalized form we shall speak of elements of the F.P.S. possessing or not possessing a certain property when it is desired to calculate a probability. For example, in calculating the probability of throwing a two with one throw of a die we may speak of an element of the F.P.S. possessing the property of being a two, or in calculating the probability of an event we may speak of an element of the F.P.S. possessing the property of happening or not happening. A definite notation will be adopted for this and we shall write $P\{E \in \omega \mid \Omega\}$ to stand for the words 'the probability that elements of the subset ω possess the property E referred to a fundamental probability set Ω'. This will invariably be abbreviated to $P\{E\}$. As stated previously this probability, or strictly estimate of probability, will be the ratio of the number of elements of the F.P.S. possessing the property E (i.e. the subset ω) to the total number of elements of which the F.P.S. is composed (i.e. the F.P.S. Ω).

It is one of the weaknesses of probability theory that it has, possibly through its intimate connexion with everyday life, taken certain everyday words and used them in a specialized sense. This is confusing in that it is not always possible to avoid using the words in both their specialized and everyday meanings. As far as is possible we shall attempt to confine the words to their specialized meaning only.

DEFINITION. Two properties, E_1 and E_2, are said to be 'mutually exclusive' or 'incompatible' if no element of the F.P.S. of E_1 and E_2 may possess both the properties E_1 and E_2.

The definition is immediately extended to k properties.

The definition of mutually exclusive or incompatible properties is thus seen to follow along common-sense lines. For example, if it was desired to calculate the probability that out of a given number of persons chosen at random ten (say) would have blue eyes, and nine (say) would have brown eyes, then in stating that the property of possessing a pair of blue eyes was incompatible with the property of possessing a pair of brown eyes, we should merely be expressing the obvious.

DEFINITION. $E_1, E_2, \ldots, E_k$, are said to be the 'only possible' properties if each element of the F.P.S. must possess one of these properties, or at least one.

THEOREM. If E_1 and E_2 are mutually exclusive and at the same time the only possible properties, then

$$P\{E_1\} + P\{E_2\} = 1.$$

Let the F.P.S. be composed of n elements of which n_1 possess the property E_1 and n_2 possess the property E_2. Since E_1 and E_2 are mutually exclusive no element of the F.P.S. may possess both the properties E_1 and E_2. We have therefore by definition,

$$P\{E_1\} = n_1/n, \quad P\{E_2\} = n_2/n.$$

Further, since E_1 and E_2 are the only possible properties, each element of the F.P.S. must then possess either E_1 or E_2, from which it follows that $n_1 + n_2 = n$ and that

$$P\{E_1\} + P\{E_2\} = 1.$$

Extension of theorem to k properties. An extension of the above theorem for k mutually exclusive and only possible properties may easily be made by following along the same lines of argument. If there are k mutually exclusive and only possible properties, $E_1, E_2, \ldots, E_k$, then

$$\sum_{i=1}^{k} P\{E_i\} = P\{E_1\} + P\{E_2\} + \ldots + P\{E_k\} = 1.$$

Definition of logical sum. Assume that the F.P.S. is composed of elements some of which possess the property E_1, or the property $E_2, \ldots$ or the property E_k. The logical sum, E_0, of any number of these different properties will be a property which consists of an element of the F.P.S. possessing any one of these properties, or at least one. This may be written

$$E_0 = E_1 + E_2 + \ldots.$$

Definition of logical product. Assume that the F.P.S. is composed of elements some of which possess the property E_1, or the property $E_2, \ldots$ or the property E_k. The logical product, E', of any number, m, of these different properties will be a property which consists of an element of the F.P.S. possessing all m properties. Thus

$$E' = E_1 E_2 \ldots.$$

These definitions may be illustrated by means of the following theorem.

THEOREM. The logical sum of two properties E_1 and E_2 is given by

$$P\{E_0\} = P\{E_1+E_2\} = P\{E_1\}+P\{E_2\}-P\{E_1E_2\}$$
$$= P\{E_1\}+P\{E_2\}-P\{E'\}.$$

Let the F.P.S. consist of n elements, n_1 of which possess the property E_1, n_2 possess the property E_2, n_{12} possess both the properties E_1 and E_2, and n_0 of which possess neither E_1 nor E_2. The proof of the theorem then follows directly from definition.

$$P\{E_1+E_2\} = \frac{n_1+n_2+n_{12}}{n}; \quad P\{E_1\} = \frac{n_1+n_{12}}{n};$$

$$P\{E_2\} = \frac{n_2+n_{12}}{n}; \qquad P\{E_1E_2\} = \frac{n_{12}}{n};$$

and the result follows.

COROLLARY. If E_1 and E_2 are mutually exclusive then

$$P\{E_1+E_2\} = P\{E_1\}+P\{E_2\}.$$

For if E_1 and E_2 are mutually exclusive then no element of the F.P.S. may possess both the properties E_1 and E_2. This means that $n_{12} = 0$ and therefore that $P\{E_1E_2\} = 0$.

Similarly for k mutually exclusive properties

$$P\left\{\sum_{i=1}^{k} E_i\right\} = \sum_{i=1}^{k} P\{E_i\}.$$

Exercise. Find an expression for the logical sum of three properties E_1, E_2 and E_3, i.e. find $P\{E_1+E_2+E_3\}$ and show how the expression is simplified if the properties are assumed to be mutually exclusive.

Numerical Examples

(1) Given that the probability of throwing a head with a single toss of a coin is constant and equal to 1/2, if two identical coins are thrown simultaneously once what is the probability of obtaining (*a*) two heads, (*b*) one head and one tail, (*c*) two tails?

If it is given that the probability of throwing a head with a single toss of a coin is constant and equal to 1/2, it follows that the probability of throwing a tail is also constant and equal to $\frac{1}{2}$ since we may assume that the properties of being head or tail are the only possible and they are obviously mutually exclusive. We may therefore construct a F.P.S. containing two elements of equal

weight; the first of these would possess the property of being a head (H) and the second would possess the property of being a tail (T). Hence for two coins we should have the following possible alternatives

$$H_1H_2, \quad H_1T_2, \quad T_1H_2, \quad T_1T_2,$$

i.e. a F.P.S. of four elements and the probability of two heads, one head and one tail, two tails will be $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ respectively.

(2) The probability that any given side will fall uppermost when an ordinary six-sided die is tossed is constant and equal to 1/6. What is the probability that when a die is tossed a 2 or 3 will fall uppermost? There are various ways in which this problem may be attempted. The setting up of a mathematical model along the lines of the previous problem would give a probability of 1/3. As an alternative we may note that the property of being a 2 is mutually exclusive of the property of being a 3 and that the logical sum of the probabilities of a 2 or a 3 will therefore be $\frac{1}{6}+\frac{1}{6}=\frac{1}{3}$.

(3) Two dice are tossed simultaneously. If the probability that any one side of either die will fall uppermost is constant and equal to 1/6, what is the probability that there will be a 2 uppermost on one and a 3 on the other? The probability that there will be a 2 on the first die and a 3 on the second will be the logical product of the two probabilities, i.e.

$$P\{2 \text{ on 1st and 3 on 2nd}\} = \tfrac{1}{6}\cdot\tfrac{1}{6} = \tfrac{1}{36}.$$

The problem does not, however, specify any order to the dice and it would be possible to get a 3 on the first and a 2 on the second. The required probability is therefore 1/18.

(4) Conditions as for (3). What is the probability that the two numbers will add up to 5? *Answer:* 1/9.

(5) n halfpennies are tossed simultaneously. If for each coin the probability that it will fall with head uppermost is constant and equal to $\frac{1}{2}$, what is the probability that k out of the n coins will fall with head uppermost?

Suppose that the coins are numbered and that the first k fall with head uppermost and the second $n-k$ with tail uppermost. The probability of this is

$$(\tfrac{1}{2})^k(\tfrac{1}{2})^{n-k} = (\tfrac{1}{2})^n.$$

But no order was specified regarding the k heads and they may be spread in any manner between the n coins. The number of ways in which k heads and $n-k$ tails can be arranged is

$$n!/k!(n-k)!$$

and the required probability is therefore

$$(\tfrac{1}{2})^n . n!/k!(n-k)!.$$

(6) Three halfpennies are tossed one after the other. If the constant probability of getting a head is $\frac{1}{2}$ for each coin, what is the joint probability that the first will be a head, the second a tail, and the third a head?

(7) Four dice are tossed simultaneously. If the constant probability of getting any number uppermost on any one die is 1/6, what is the probability that the sum of the numbers on the four dice is 12?

(8) Five cards are drawn from a pack of 52. If the constant probability of drawing any one card is 1/52, what is the probability that these five cards will contain (*a*) just one ace, (*b*) at least one ace? If the probability is constant then a F.P.S. can be set up consisting of 52 equally likely alternatives. The number of ways in which 5 cards can be drawn from 52 if all cards are equally likely is

$$52!/5!\,47!.$$

From this number it is necessary to pick out the number of sets of 5 cards, one card of which is an ace. This is perhaps most easily done by first withdrawing the 4 aces from the pack. The number of ways in which 4 cards may be drawn from 48 will be

$$48!/4!\,44!.$$

To each of these sets of 4 cards one ace must be added and this may be done in 4 ways. Hence the total number of ways in which 5 cards may be drawn, one of which is an ace, is

$$4.48!/4!\,44!,$$

and the required probability of drawing 5 cards, one of which is an ace, is

$$4.\frac{48!}{4!\,44!}\Big/\frac{52!}{5!\,47!}.$$

A similar argument may be used for the probability of obtaining at least one ace. In this problem the 5 cards may contain just

one ace, or two aces, or three or four. The required probability will therefore be

$$\frac{5!\,47!}{52!}\left[4\,\frac{48!}{4!\,44!}+\frac{4.3}{1.2}\,\frac{48!}{3!\,45!}+\frac{4.3.2}{1.2.3}\,\frac{48!}{2!\,46!}+\frac{48!}{1!\,47!}\right].$$

Examples based on games of chance tend to be somewhat artificial in that the answer when obtained is without interest, for few persons are prepared nowadays to wager large sums of money on the fall of a die. Nevertheless it was from a study of the gaming tables that the earliest developments of the theory were made and such problems are of value if the student learns the elements of combinatory theory from their study.

Definition of relative probability. The relative probability of a property E_2, given a property E_1, will be defined as the probability of a property E_2 referred to the set of individuals of the F.P.S. possessing the property E_1. This will be written as

$$P\{E_2 \mid E_1\}.$$

The notation may be translated into words exactly as before with the addition of the word 'given' represented by the upright stroke |.

THEOREM. Whatever the two properties E_1 and E_2,

$$P\{E_1E_2\} = P\{E_1\}\,P\{E_2 \mid E_1\} = P\{E_2\}\,P\{E_1 \mid E_2\}.$$

Let the F.P.S. be composed of n elements of which n_1 possess the property E_1, n_2 possess the property E_2, n_{12} possess both the properties E_1 and E_2, n_0 possess neither E_1 nor E_2. By definition,

$$P\{E_1E_2\} = \frac{n_{12}}{n};\quad P\{E_1\} = \frac{n_1+n_{12}}{n};\quad P\{E_2 \mid E_1\} = \frac{n_{12}}{n_1+n_{12}}$$

and the proof of the theorem follows. The second equality follows in a similar way.

THEOREM. Whatever the properties $E_1, E_2, \ldots, E_k$,

$$P\{E_1E_2\ldots E_k\} = P\{E_1\}\,P\{E_2 \mid E_1\} \times P\{E_3 \mid E_1E_2\}\ldots P\{E_k \mid E_1E_2\ldots E_{k-1}\}.$$

The proof follows immediately from repeated application of the preceding theorem.

Numerical Example

Three bags A, B and C are filled with balls identical in size and weight. Bag A contains M balls of which m_1 are stamped with the number 1, m_2 with the number 2, and $m_1+m_2=M$. Bag B contains N_1 balls, n_1 of which are white and N_1-n_1 of which are black, while Bag C contains N_2 balls, n_2 of which are white and N_2-n_2 black. A ball is drawn from A. If it bears the number 1 a ball is drawn from B; if it bears the number 2 a ball is drawn from C. Assuming that the probability of drawing an individual ball from any bag is constant and equal to the reciprocal of the number of balls in the bag, what is the probability that, if a ball is drawn from A and then from B or C, the second ball is white? Describe in detail the F.P.S. to which this probability may refer.

It is possible to state the required probability immediately. Using the notions of logical sum and logical product we may say at once that the probability that the second ball is white is

$$\frac{m_1}{M}\frac{n_1}{N_1}+\frac{m_2}{M}\frac{n_2}{N_2}.$$

It is required, however, to discuss the possible mathematical model which may be set up for the calculation of this probability under the given assumption of all balls within one bag being equally likely to be drawn. One mathematical model would be the enumeration of all possible pairs of balls which might be drawn. These will be

$1W$	$1B$	$2W$	$2B$
a	b	c	d

where a, b, c and d are the number of these pairs. It follows that

$$P\{1\}=\frac{m_1}{M}=\frac{a+b}{a+b+c+d}, \quad P\{2\}=\frac{m_2}{M}=\frac{c+d}{a+b+c+d},$$

$$P\{W \mid 1\}=\frac{n_1}{N_1}=\frac{a}{a+b}, \qquad P\{W \mid 2\}=\frac{n_2}{N_2}=\frac{c}{c+d}.$$

The probability of getting a white ball, given 1 and 2, will therefore be

$$P\{W \mid 1 \text{ and } 2\}=\frac{a+c}{a+b+c+d}=\frac{m_1 n_1}{MN_1}+\frac{m_2 n_2}{MN_2}$$

and a solution of the equations for a, b, c and d will give a complete enumeration of the F.P.S.

Definition of independence. The property E_1 is independent of the property E_2 if

$$P\{E_1\} = P\{E_1 \mid E_2\}.$$

THEOREM. If the property E_1 is independent of the property E_2 then the property E_2 is independent of the property E_1.

From the definition of independence, if E_1 is independent of E_2 then

$$P\{E_1\} = P\{E_1 \mid E_2\}.$$

It is required to show that E_2 is independent of E_1, i.e.

$$P\{E_2\} = P\{E_2 \mid E_1\}.$$

The result follows immediately from a consideration of the logical product of E_1 and E_2.

$$P\{E_1 E_2\} = P\{E_1\} P\{E_2 \mid E_1\} = P\{E_2\} P\{E_1 \mid E_2\}.$$

Using the fact that E_1 is independent of E_2 the converse follows.

Example. If E_1 and E_2 are mutually exclusive, can they be independent? The answer to the question follows directly from the definitions. E_1 and E_2 are mutually exclusive if no element of the F.P.S. possesses both the properties E_1 and E_2. That is to say

$$P\{E_1 \mid E_2\} = 0 = P\{E_2 \mid E_1\}.$$

The condition for independence has just been stated as

$$P\{E_1\} = P\{E_1 \mid E_2\}, \quad P\{E_2\} = P\{E_2 \mid E_1\}.$$

Hence E_1 and E_2 can only be mutually exclusive and independent if

$$P\{E_1\} = 0 = P\{E_2\},$$

which is absurd. It follows therefore that E_1 and E_2 cannot be both mutually exclusive and independent.

Example. Consider three properties E_1, E_2 and E_3. Given (i) the F.P.S. is finite, (ii) E_1 is independent of E_2, (iii) E_1 is independent of E_3, (iv) E_1 is independent of $E_2 E_3$.

Prove that E_1 is also independent of $E_2 + E_3$.

Again the solution of the problem follows directly from the definitions previously given. Let the F.P.S. be composed of n elements, n_1 of which possess the property E_1, n_2 the property E_2, n_3 the property E_3, n_{12} both the properties E_1 and E_2, n_{23} both

the properties E_2 and E_3, n_{31} both the properties E_3 and E_1, n_{123} all three properties E_1, E_2, E_3, and n_0 possess none of these properties.

From the given conditions

$$P\{E_1\} = P\{E_1 \mid E_2\} = P\{E_1 \mid E_3\} = P\{E_1 \mid E_2 E_3\} = \alpha \text{ (say)}.$$

Substituting for these probabilities we have

$$\alpha = \frac{n_1 + n_{12} + n_{31} + n_{123}}{n} = \frac{n_{12} + n_{123}}{n_2 + n_{12} + n_{23} + n_{123}}$$

$$= \frac{n_{31} + n_{123}}{n_3 + n_{31} + n_{23} + n_{123}} = \frac{n_{123}}{n_{23} + n_{123}}.$$

A solution of these equations and a simple rearrangement shows that

$$\alpha = \frac{n_{12} + n_{31} + n_{123}}{n_2 + n_3 + n_{31} + n_{23} + n_{123}} = P\{E_1 \mid E_2 + E_3\}$$

and since

$$\alpha = P\{E_1\}$$

the result follows. It will be noticed that no mention was made of independence (or otherwise) between E_2 and E_3 and the result will hold therefore whether these two properties are independent or not.

These then are the preliminary definitions and theorems which are necessary for the development of probability theory. All will be used in the exposition which follows although we shall not necessarily restate each theorem or definition at the time at which it is used. It should be the aim of the reader so to familiarize himself with the concepts that reference back to this chapter becomes unnecessary. Further, each and every stage of a calculation of any probability, no matter how trivial, should be followed by an interpretation of what the probability means, or would mean if experiments were carried out. Only by such repetition can the theory of probability acquire for the reader both meaning and sense.

REFERENCES AND READING

There would appear to be little argument possible about the theorems outlined in this chapter. The reader will notice that all proofs are based on the assumption that the F.P.S. is finite. The propositions can be shown also to hold for an infinite F.P.S. but appeal would need to be made

to the theory of sets in order to justify the proofs. The reader must perforce accept the propositions as true for all cases.

If further examples are required they may be found in many algebra text-books or in the chapter headed 'Probability' in W. A. Whitworth, *Choice and Chance*, whose examples and illustrations are exhaustive.

It will be useful for the reader of these books to add to both the question and answer the words necessary before a probability can be calculated or interpreted. For example, the favourite problem of the probability of drawing balls from an urn is incalculable unless it is assumed that all balls have an equal probability of being drawn, and so on.

CHAPTER III

THE BINOMIAL THEOREM IN PROBABILITY

Following the definitions and theorems for the addition and multiplication of probabilities which have been set out in the previous chapter it would be possible to solve any problem in elementary probability; for it is not possible to conceive a problem which could not be solved ultimately by its reduction to first principles. Nevertheless, from the basic raw material of our subject as represented by these first principles, it is possible to fashion tools which add not only to our appreciation of its applications but which also seem greatly to extend our knowledge. Intrinsically the application of the binomial theorem to probability described below is just a rapid method for the calculation of probabilities by the joint application of the elements of probability theory and combinatorial analysis. Nevertheless, because of its utility in modern probability and statistical theory, it will be advantageous to consider the use of the theorem in some detail.

The problem in which the binomial theorem is most frequently employed is sometimes referred to as the problem of repeated trials. This arises from the fact that it generally presupposes a series of repeated trials in each of which the probability of an event occurring is constant; it is required to state the probability of a given number of successes in a total of repeated trials. In order to prove the result it is not necessary to assume that this probability, constant from trial to trial, is also a known probability. We shall, however, begin by an illustration in which the probability of a single event occurring is assumed as known.

Let us consider the case of the tossing of a halfpenny when it is known that the constant probability of a head is $\frac{1}{2}$ in a single trial. If the halfpenny is tossed twice then, as we have seen earlier, an appropriate mathematical model will be

$$H_1H_2, \quad H_1T_2, \quad T_1H_2, \quad T_1T_2,$$

with the probabilities of $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$ of obtaining two heads, one head

and one tail, or two tails respectively with two spins of the coin. If the halfpenny is tossed three times in succession the appropriate alternatives will be

$$H_1H_2H_3,\quad H_1H_2T_3,\quad H_1T_2H_3,\quad T_1H_2H_3,$$
$$H_1T_2T_3,\quad T_1H_2T_3,\quad T_1T_2H_3,\quad T_1T_2T_3.$$

The three tosses can give the possible results of three heads, two heads and one tail, one head and two tails or three tails with corresponding probabilities $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$; but already the enumeration of the possible alternatives is becoming cumbersome and leaves the way open for errors. It is clear that for cases in which ten or more tosses were made, the appeal to first principles, although still a possibility, would need considerable calculation, whereas, as will be shown, the required probabilities may be obtained immediately by an application of the binomial theorem.

THEOREM. If the probability of the success of the event E in a single trial is constant and equal to p, then the probabilities of k successes (for $k = 0, 1, 2, \ldots, n$), in n independent trials are given by the successive terms of the expansion of

$$(q+p)^n,$$

where $q = 1-p$.

Let $P_{n.k}$ denote the probability that an event, the constant probability of the occurrence of which in a single trial is p, will happen exactly k times in n trials. In order to prove the theorem, therefore, it is necessary to prove the identity

$$(q+px)^n \equiv P_{n.0}+P_{n.1}x+\ldots+P_{n.k}x^k+\ldots+P_{n.n}x^n.$$

There are many ways in which this may be done but possibly one of the simplest is by induction, dividing the proof into three parts.

(1) The identity is true for $n = 1$, i.e.

$$q+px \equiv P_{1.0}+P_{1.1}x.$$

This follows directly from definition. $P_{1.0}$ is the probability that in a single trial the event will not happen, that is $P_{1.0}$ equals q. Similarly $P_{1.1}$ is the probability that in a single trial the event will happen. Hence $P_{1.1}$ must be equal to p.

(2) Assume that the identity is true for m and prove that if it is true for m it is true for $m+1$. The assumption is therefore that

$$(q+px)^m \equiv P_{m.0}+P_{m.1}x+\ldots+P_{m.k}x^k+\ldots+P_{m.m}x^m.$$

Multiply each side by $(q+px)$ and collect the coefficients.

$$(q+px)^{m+1} \equiv q.P_{m.0} + x(q.P_{m.1} + p.P_{m.0}) + \ldots + x^k(q.P_{m.k} + p.P_{m.k-1}) + \ldots + x^m p.P_{m.m}.$$

Consider the different coefficients separately.

$q.P_{m.0}$ = the probability that an event will not happen at all in m trials multiplied by the probability that it will not happen in a further single trial.

Hence $$q.P_{m.0} = P_{m+1.0}.$$

Similarly it may be argued that

$$p.P_{m.m} = P_{m+1.m+1}.$$

It will be sufficient for the other coefficients to consider a typical term, say the coefficient of x^k.

$q.P_{m.k} + p.P_{m.k-1}$ = the probability that an event will happen exactly k times in m trials multiplied by the probability that it will not happen in one further trial, plus the probability that it will happen $k-1$ times in the first m trials multiplied by the probability that it will happen in one further trial.

It is clear therefore that

$$q.P_{m.k} + p.P_{m.k-1} = P_{m+1.k}.$$

(3) It has been shown that if the identity is true for m it is also true for $m+1$. It has also been shown that the identity is true for n equal to unity. Hence if it is true for n equal to one it is true for n equal to two and so on universally. Writing x equal to unity we have

$$(q+p)^n = P_{n.0} + P_{n.1} + \ldots + P_{n.k} + \ldots + P_{n.n}$$

and the theorem is proved.

The function $(q+px)^n$ is sometimes called the generating function of the probabilities $P_{n.k}$.

Example. If the constant probability of obtaining a head with a single throw of a halfpenny is $\frac{1}{2}$, what is the probability that in twelve tosses of the coin there will be nine heads?

The answer will be the coefficient of x^9 in the expansion of $(\frac{1}{2}+\frac{1}{2}x)^{12}$, that is it will be

$$\frac{12!}{9!\,3!}\left(\frac{1}{2}\right)^{12} = \frac{55}{1024}.$$

This result may be compared with that of Example (5) of Chapter II. (Page 16.)

Example. If the constant probability of obtaining a two uppermost with a single throw of a die is 1/6, what is the probability of obtaining 3 twos in 8 throws of the die?

Answer: $\dfrac{8!}{3!\,5!}\left(\dfrac{1}{6}\right)^3\left(\dfrac{5}{6}\right)^5$.

Example. (*The problem of points.*) Jones and Brown are playing a set of games. Jones requires s games to win and Brown requires t. If the chances of Jones winning a single game is p, find the probability that he wins the set. This problem is a version of the famous 'problem of points' which has engaged the attention of many writers on classical probability. The solution follows directly from an application of the binomial theorem but its artificiality should be recognized. It appears doubtful whether skill at games can ever be expressed in terms of chances of winning and, further, it would seem that when chances of winning are spoken of with regard to a set of games something in the nature of a subjective probability is meant and not purely the objective probability of number. However in spite of its artificiality the problem is not without interest.

If Jones' chances of winning a single game is p, then Brown's must be $q = 1-p$, because either Jones or Brown must win the single game. Suppose Jones takes $s+r$ games to win the set. In order to do this he must win the last game and $s-1$ of the preceding $s+r-1$ games. The probability that Jones will win $s-1$ out of $s+r-1$ games, when the constant probability that he wins a single game is p, may be written down directly from the binomial theorem. It will be

$$\frac{(s+r-1)!}{(s-1)!\,r!}\,p^{s-1}q^r.$$

It follows then that the probability that Jones wins the set in $s+r$ games will be the probability that he wins the last game

multiplied by the probability that he wins $s-1$ out of the other $s+r-1$ games. This is

$$\frac{(s+r-1)!}{(s-1)!\,r!}p^s q^r.$$

Now Jones may win in exactly s games, or $s+1$ games, or $s+2$ games, ... or $s+t-1$ games. The probability that he wins the set is therefore obtained by letting r take values $0, 1, 2, \ldots, (t-1)$ successively and summing, i.e. Jones' chance is

$$p^s + sp^s q + \frac{s(s+1)}{2!}p^s q^2 + \ldots + \frac{(s+t-2)!}{(s-1)!\,(t-1)!}p^s q^{t-1}.$$

An interesting algebraic identity may be obtained by considering Brown's chance of winning and its relation to Jones' chance.

Thus far we have treated binomial applications in which the constant probability of a single event is known. It is, however, not usual in statistical problems that this probability should be known and in the majority of cases it is necessary to estimate it from the data provided. We shall return to this point at length at a later stage but an example here will serve as further illustration of the application of the binomial theorem.

Example. The report of the dean of a Cambridge college showed the following figures:

Subject	Number of students examined	Number of honour grades	Number of failures
Mathematics	466	162	38
Music	22	11	0
All subjects	—	38 %	5·4 %

What is the probability that

(1) in selecting 466 students at random one would obtain as few honour grades as were obtained in mathematics, and as many failures?

(2) in selecting 22 students at random one would obtain no failures, as in music, and 11 or more honour grades?

The percentage of students obtaining honour grades in all subjects is 38. Further, the problem states that 466 students are to be selected at random. It would seem therefore that the appropriate value for the probability of obtaining an honour grade is the proportion of honours students within the population, i.e. 0·38. Similarly the probability for failure may be taken

as 0·054. The answers will be, from direct application of the theorem,

$$(1)\ \frac{466!}{162!\,304!}(0{\cdot}38)^{162}(0{\cdot}62)^{304},\quad \frac{466!}{38!\,428!}(0{\cdot}054)^{38}(0{\cdot}946)^{428}.$$

$$(2)\ \frac{22!}{0!\,22!}(0{\cdot}054)^{0}(0{\cdot}946)^{22}.$$

The '11 or more' honour grades requires us to calculate the probabilities of obtaining 11, 12, ..., 22 honour grades. This will be

$$\frac{22!}{11!\,11!}(0{\cdot}38)^{11}(0{\cdot}62)^{11}+\frac{22!}{12!\,10!}(0{\cdot}38)^{12}(0{\cdot}62)^{10}$$
$$+\frac{22!}{13!\,9!}(0{\cdot}38)^{13}(0{\cdot}62)^{9}+\ldots+\frac{22!}{22!\,0!}(0{\cdot}38)^{22}.$$

In the foregoing example we have treated a problem in which a sample is drawn randomly from a population and, in so doing, may appear to have diverged a little from the problem of repeated trials. If we consider the procedure in detail, however, it will be obvious that the problem of sampling from an infinite population, or sampling with replacement from a finite population, is identical with the problem of repeated trials. In the problem of repeated trials we consider an event E which has a constant probability of happening in a single trial. If n trials are performed, then the binomial theorem enables us to calculate the probability that the event will happen exactly k (say) times in these n trials without specifying anything about the order of these k successes; in fact the k successes are supposed to occur in an entirely random way. Now consider the drawing of a sample at random* from a population. For illustration we may imagine the population to consist of a box containing disks. If a disk is chosen at random from the box it will mean that any one disk has the same chance of being chosen as any other disk, and provided the disk is returned to the box after each drawing, the probability of choosing any one disk will be constant from trial to trial. Hence if a disk

* A sample randomly drawn from a population is commonly spoken of as a 'random sample'. We shall follow common usage by writing of 'random samples' but it is necessary to remember that the adjective 'random' should apply to the method of drawing the sample and not to the sample itself.

is taken out and returned ten times, we may say that we have selected a 'random sample' of the disks, but we might also say that we had made repeated trials which were ten in number. In the case of the Cambridge students, provided each student in the population was given the chance of being chosen more than once (i.e. provided the disk is returned to the box), the choosing of 466 students at random is equivalent to making repeated trials 466 in number, the probability for choosing an honours student being constant from trial to trial.

Example. A population is composed of equal numbers of red and white disks. These disks are identical in all respects except for colour. A disk is chosen at random and replaced 8 times. What is the probability that this sample of 8 will be made up of 0, 1, 2, ..., 8 red disks?

We may assume that the probability of obtaining a red disk at a single trial is $\frac{1}{2}$. It is stated that 8 trials are made. The required probabilities are therefore given by successive terms of the binomial

$$(\tfrac{1}{2}+\tfrac{1}{2})^8.$$

Example. In the 8 offspring from the mating of a hybrid (**Aa**) and a recessive (**aa**), 7 were observed to be hybrids and one recessive. Is this result exceptional?

Following the genetical hypotheses (see Chapter VIII) it is clear that the only offspring from the mating **Aa** × **aa** can be hybrids (**Aa**) and recessives (**aa**) and that these may be expected to occur in equal numbers. In other words, the probability of obtaining a hybrid is $\frac{1}{2}$ and the hybrid and recessive are the only possible properties. It follows that the probability of obtaining 7 or more than 7 hybrid offspring from such a mating will be

$$2\left[\frac{8!}{7!\,1!}\left(\frac{1}{2}\right)^8+\frac{8!}{8!\,0!}\left(\frac{1}{2}\right)^8\right]=0{\cdot}07.$$

We interpret this probability by stating that, on the average, seven times in 100 we should expect to obtain 7 or 8 hybrids in a family of 8 and we cannot therefore consider that the reported 7 hybrids are exceptional in number.

Let us return now to a study of the binomial probabilities generated from $(q+p)^n$. This series of probabilities will only be symmetrical for $p = q = \frac{1}{2}$. For p greater than $\frac{1}{2}$ the largest term

will be towards the right of the distribution, for it is to be expected that if the probability of success in a single trial is large then the probability of obtaining a high proportion of successes would also tend to be large. The position of the largest term, or of the two largest terms, may be found by means of two simple inequalities.

It has been shown that

$$(q+p)^n = P_{n.0}+P_{n.1}+\ldots+P_{n.k}+\ldots+P_{n.n}.$$

Let the largest term be when $k = k_0$. It will follow that

$$P_{n.k_0-1} \leqslant P_{n.k_0} > P_{n.k_0+1}.$$

Consider first the left-hand inequality. Substituting for $P_{n.k_0}$ and $P_{n.k_0-1}$ we have

$$\frac{n!}{(k_0-1)!\,(n-k_0+1)!}\,p^{k_0-1}q^{n-k_0+1} \leqslant \frac{n!}{k_0!\,(n-k_0)!}\,p^{k_0}q^{n-k_0},$$

from which it follows that

$$(n+1)\,p \geqslant k_0.$$

Similarly it may be shown from the right-hand inequality that

$$k_0 > (n+1)\,p-1,$$

so that

$$(n+1)\,p \geqslant k_0 > (n+1)\,p-1.$$

The largest term of the binomial series may therefore be found quickly for it is the term corresponding to the integer which satisfies this inequality.

Example. Find the largest term in the expansion of $(\frac{1}{2}+\frac{1}{2})^8$. Here $p = \frac{1}{2}$ and $n = 8$ and we have

$$4{\cdot}5 \geqslant k_0 > 3{\cdot}5.$$

This means that the largest term of the expansion will be when $k_0 = 4$. It may be pointed out, however, that if the largest term is when $k_0 = 4$ this will imply that the *fifth* term of the series is the greatest, for the number of successes may be **0**, 1, 2, 3, 4, ..., 8.

Exercise. Find the largest term of the following binomials.

(1) $(\frac{1}{2}+\frac{1}{2})^{10}$, $(\frac{1}{3}+\frac{2}{3})^{10}$, $(\frac{1}{5}+\frac{4}{5})^{10}$.

(2) $(\frac{1}{2}+\frac{1}{2})^{20}$, $(\frac{1}{3}+\frac{2}{3})^{20}$, $(\frac{1}{5}+\frac{4}{5})^{20}$.

Calculate the distributions and thus check your results.

This exercise shows clearly the way in which a change in p and in n alters the shape of the binomial distribution. Generally,

however, we shall not be concerned with the term of greatest probability, for it is not of much utility in statistical theory. We shall consider the two other collective characters, the mean and the standard deviation of the distribution.

A knowledge of the moments of the binomial distribution is necessary for several reasons. First, by studying these moments and the derived collective characters β_1 and β_2 an idea of the shape of the distribution can be arrived at just as quickly as from a study of the most probable term, and possibly more accurately. Secondly, in cases where it is necessary to approximate to the binomial by some frequency curve it will be necessary to know at least the first two moments. Thirdly, in fitting the binomial series to a set of observations it is usually necessary to estimate both n and p. This may be done most quickly by equating the mean and standard deviation of the binomial series to those calculated from the observations. We shall therefore extend the framework of our theory and derive the moments of the binomial series.

THEOREM. If the probability of the success of an event E in a single trial is constant and equal to p, then the theoretical distribution of the probabilities of obtaining $0, 1, 2, \ldots$ successes in n trials will have the following first four moments:

(1) Mean $= \mu_1' = np$, (2) Variance $= \mu_2 = npq$,

(3) $\mu_3 = npq(q-p)$, (4) $\mu_4 = npq[1+3pq(n-2)]$,

where $q+p = 1$.*

The calculation of these theoretical moments may be exactly paralleled in the reader's mind by the calculation of the moments of any given frequency distribution, such as is worked out early on in statistical practice.

Regarding the probabilities as frequencies we have

$$\text{Mean} = \mu_1' = \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^k q^{n-k},$$

where k may be compared with the distance from the arbitrary origin and $\frac{n!}{k!(n-k)!} p^k q^{n-k}$ may be compared with the group

* It is assumed that the student will have read enough statistics to be familiar with the notation for moments. Briefly μ_k' indicates the kth moment about any arbitrary origin, and μ_k indicates the kth moment about the mean.

frequency. There is no need here to divide by the sum of the 'frequencies' because in this case it is unity. If the term np is taken outside the summation sign it will be seen that the terms inside are simply another binomial series with $n-1$ as index instead of n, and are accordingly equal to unity. Hence

$$\text{Mean} = \mu_1' = np.$$

The variance follows in the same way.

$$\mu_2' = \sum_{k=0}^{n} k^2 \frac{n!}{k!(n-k)!} p^k q^{n-k} = \sum_{k=0}^{n} [k(k-1)+k] \frac{n!}{k!(n-k)!} p^k q^{n-k},$$

whence $$\mu_2' = n(n-1)p^2 + np.$$

Applying the correction to μ_2' in order to obtain the variance μ_2, that is the second moment about the mean, we have

$$\mu_2 = \mu_2' - \mu_1'^2 = npq.$$

Similarly

$$\begin{aligned}\mu_3' &= \sum_{k=0}^{n} k^3 \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &= \sum_{k=0}^{n} [k(k-1)(k-2) + 3k(k-1) + k] \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &= n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np.\end{aligned}$$

and

$$\begin{aligned}\mu_4' &= \sum_{k=0}^{n} k^4 \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &= \sum_{k=0}^{n} [k(k-1)(k-2)(k-3) + 6k(k-1)(k-2) \\ &\quad + 7k(k-1) + k] \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &= n(n-1)(n-2)(n-3)p^4 \\ &\quad + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np.\end{aligned}$$

It will be necessary to convert μ_3' and μ_4' from the arbitrary origin to the mean. These corrections are

$$\begin{aligned}\mu_3 &= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^2, \\ \mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4\end{aligned}$$

and using these relations easy algebra gives

$$\begin{aligned}\mu_3 &= npq(q-p), \\ \mu_4 &= npq[1 + 3pq(n-2)],\end{aligned}$$

which proves the theorem.

As a corollary it is straightforward to show that

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{1}{npq} - \frac{4}{n}; \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{1-6pq}{npq}; \quad \beta_2 - \beta_1 = 3 - \frac{2}{n}.$$

Example. In 103 litters of 4 mice the number of litters which contained 0, 1, 2, 3, 4 females were noted. The figures are given in the table below:

Number of female mice	0	1	2	3	4	Total
Number of litters	8	32	34	24	5	103

(1) If the chance of obtaining a female in a single trial is assumed constant, estimate this constant but unknown probability.

(2) If the size of the litter (4) had not been given, how could it be estimated from the data?

(3) How could the assumption that the chance of obtaining a female in a single trial is constant be tested?

(1) The mean of the observations given is equal to

$$\tfrac{1}{103}[8.0 + 32.1 + 34.2 + 24.3 + 5.4] = 1{\cdot}864.$$

If it is assumed therefore that the number of litters in each class divided by 103 is an estimate of the binomial probability in each class we shall have

$$np = 1{\cdot}864$$

whence, since n is given equal to 4, it follows that p is equal to 0·466.

(2) If n is not given but must be estimated from the data it will be necessary to calculate the variance of the given observations. This is equal to 1·030. We have, therefore, equating the theoretical variance of the binomial to that of the observational data, that

$$npq = 1{\cdot}030.$$

Dividing by the relationship

$$np = 1{\cdot}864,$$

we have $q = 0{\cdot}553$, $p = 0{\cdot}447$ and n accordingly approximately equal to 4. Since in this case n must be an integer we should have no hesitation in estimating the litter size as 4 and readjusting the probability accordingly.

(3) There are several ways in which the adequacy of the initial assumptions may be tested. The simplest one perhaps is to calculate the frequencies as given by the terms of the theoretical binomial

$$103(0{\cdot}534+0{\cdot}466)^4$$

and compare with the actual frequencies.

Observed frequency	8	32	34	24	5	103
Theoretical frequency	8	29	38	23	5	103

There is no need for a further test here to find out whether the theoretical hypothesis as given by the binomial adequately describes the observational data. The agreement between theory and observation is good and we may say that there is no reason why the probability should not be assumed constant.

Exercise. A cast of 12 dice was made 26,306 times and the frequency of dice with 5 or 6 points uppermost was recorded. W. F. R. Weldon found the following distribution:

Number of dice with 5 or 6 points	0	1	2	3	4	5	6	7	8	9	10	11	12
Observed frequency	185	1149	3265	5475	6114	5194	3067	1331	403	105	14	4	0

Check whether the mathematical model whereby all sides of a die may be assumed equi-probable is suitable for this set of observations and calculate the theoretical frequencies appropriate to the binomial hypothesis.

Exercise. A wooden target is divided into 1000 squares. Shots are fired at the target, the aiming being supposedly random within the area of the target. The distribution of the number of shots in any one square is the following:

Number of shots (k) within a square	0	1	2	3	4	5	6	7	8	9	10	11
Number of squares with k shots	0	1	4	10	89	190	212	204	193	79	16	2

How could the hypothesis that the aiming was random within the target area be tested from these figures?

REFERENCES AND READING

W. Whitworth, *Choice and Chance*, again provides many examples necessitating the calculation of a binomial probability for their solution.

J. V. Uspensky, *Introduction to Mathematical Probability*, also has a variety of examples but with a strong mathematical bias to them.

Almost any statistical text-book may be consulted for practical examples.

The student might profitably read G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, for statistical examples.

CHAPTER IV

EVALUATION OF BINOMIAL PROBABILITIES

It will not have escaped the notice of the reader that the evaluation of binomial probabilities if n, the number of trials, be large, will lead to a certain amount of somewhat tedious arithmetic. The evaluation of a single value of $P_{n.k}$ may be carried out fairly quickly, provided tables of log-factorials are available, but it is rare that a single probability is required. More often than not in statistical practice we are concerned with finding the probability of obtaining a number greater than or less than a given number and it becomes necessary to evaluate the sum of a number of binomial probabilities. In this case the calculations are frequently lengthy and allow much scope for error. It is worth while therefore to study such methods as there are for the evaluation of such a sum, and to consider what approximations to the binomial series have been made. We shall do this in several stages.

1. Relation between the Binomial Series and the Incomplete B-function Ratio

The complete B-function may be defined as

$$\mathrm{B}(s,r) = \int_0^1 x^{s-1}(1-x)^{r-1}\,dx$$

and the incomplete B-function as

$$\mathrm{B}_t(s,r) = \int_0^t x^{s-1}(1-x)^{r-1}\,dx \quad \text{for} \quad 0<t<1.$$

Provided s and r are integers the complete B-function may also be expressed in terms of complete Γ-functions and thence in the ratio of factorials, viz.:

$$\mathrm{B}(s,r) = \frac{\Gamma(s)\,\Gamma(r)}{\Gamma(s+r)} = \frac{(s-1)!\,(r-1)!}{(s+r-1)!} = \mathrm{B}(r,s).$$

The incomplete B-function ratio is the ratio of the incomplete B-function to the complete B-function. In statistical practice it is commonly written

$$I_t(s,r) = \frac{\mathrm{B}_t(s,r)}{\mathrm{B}(s,r)} = \int_0^t x^{s-1}(1-x)^{r-1}\,dx \Big/ \int_0^1 x^{s-1}(1-x)^{r-1}\,dx.$$

We begin by considering the incomplete B-function ratio, $I_p(k, n-k+1)$, where n, k and p have their usual meanings.

From the above definition

$$I_p(k, n-k+1) = \frac{n!}{(k-1)!\,(n-k)!}\int_0^p x^{k-1}(1-x)^{n-k}\,dx.$$

The function may easily be evaluated by integrating by parts:

$$\int_0^p x^{k-1}(1-x)^{n-k}\,dx = \frac{p^k(1-p)^{n-k}}{k} + \frac{(n-k)}{k(k+1)}p^{k+1}(1-p)^{n-k-1} + \ldots + \frac{(k-1)!\,(n-k)!}{n!}p^n.$$

Writing $q = 1-p$ it follows that

$$I_p(k, n-k+1) = \frac{n!}{k!(n-k)!}p^k q^{n-k} + \frac{n!}{(k+1)!\,(n-k-1)!}p^{k+1}q^{n-k-1} + \ldots + p^n,$$

which is equivalent to saying that

$$I_p(k, n-k+1) = P_{n.k} + P_{n.k+1} + \ldots + P_{n.n} = \sum_{r=k}^{r=n} P_{n.r} = P\{r \geqslant k\}.$$

It will be recognized therefore that, provided the incomplete B-function ratio is tabled, the sum of any number of binomial terms may be obtained. For example, if the sum of a number of terms from k_1 to k_2 is required, then this is the difference of two incomplete B-function ratios, and so on.

Tables of the incomplete B-function ratio were prepared in the Biometric Laboratory, University College, London, and edited by Karl Pearson. The function $I_t(s,r)$ is tabled for r over the range 0 to 50, and the values of s for each particular value of r extend from $s = r$ to $s = 50$. The argument of t is in hundredths. Thus in the table entries it is not possible immediately to extract the incomplete B-function ratio if $r > s$. This omission was, however, deliberate in order to cut down the length of the table, since

a simple mathematical transformation is all that is required in such cases. We have noted that the incomplete B-function ratio is defined as

$$I_t(s,r) = \frac{(s+r-1)!}{(s-1)!\,(r-1)!}\int_0^t x^{s-1}(1-x)^{r-1}\,dx.$$

Write $1-x=y$ and we have

$$I_t(s,r) = \frac{(s+r-1)!}{(s-1)!\,(r-1)!}\left[\int_0^1 y^{r-1}(1-y)^{s-1}\,dy - \int_0^{1-t} y^{r-1}(1-y)^{s-1}\,dy\right],$$

that is

$$I_t(s,r) = 1 - I_{1-t}(r,s).$$

Hence the incomplete B-function ratio can be determined for any value of r and s between 0 and 50. This would suggest that the largest value for n would also be 50, and this would be so for a complete enumeration of the binomial series. If, however, the sum of terms is required for some $k > np$ it will be seen that it is possible for n to be greater than 50 and that the extreme case which may be evaluated will be for $k = 50$ and $n = 99$.

In such tables we have therefore a weapon of great utility by which much laborious calculation may be avoided, provided that the n and k of the binomial series are sufficiently small to fall within the range of the tables. The answer obtained from these tables is exact. For values of n falling outside the range of the tables it will be necessary to use an approximation. Such approximations which do not involve more calculation than the working out of the binomial terms themselves are not valid for n small but may be used freely for n greater than 50.

2. Approximation to the Binomial Series using a Hypergeometric Series

It has been pointed out by Uspensky* that an approximation to the binomial series, using properties of the hypergeometric series, was put forward by Markoff and that this approximation

* The analysis of this section follows directly along the same lines developed independently by J. V. Uspensky and J. Müller. (See references at the end of the chapter.) My proof follows closely that given by Uspensky, although I have inserted more detail, because I feel that in general his proof cannot be improved upon.

has not received the recognition which is undoubtedly its due. This is, possibly, owing to the fact that the normal curve is well tabulated and the effort involved in approximating to a sum of binomial probabilities by part of the area of a normal curve (see next chapter) is very much less than that involved in Markoff's approximation. Nevertheless, the hypergeometric approximation needs fewer mathematical assumptions than does the normal approximation and for this reason we shall deal with it in some detail.

It is required to evaluate $P\{k > k_1\}$, that is, it is required to evaluate

$$\sum_{k=k_1+1}^{n} P_{n.k} = \frac{n!}{(k_1+1)!\,(n-k_1-1)!} p^{k_1+1} q^{n-k_1-1} + \frac{n!}{(k_1+2)!\,(n-k_1-2)!} p^{k_1+2} q^{n-k_1-2} + \ldots + p^n.$$

The assumption is made that $k_1 \geqslant np$. This assumption is not, however, restricting, for if $k_1 < np$ then $\sum_{k=0}^{k_1} P_{n.k}$ could be evaluated by this same method and subtracted from unity to give the required probability. The first term of the required sum may be put outside a bracket as follows:

$$\sum_{k=k_1+1}^{n} P_{n.k} = \frac{n!}{(k_1+1)!\,(n-k_1-1)!} p^{k_1+1} q^{n-k_1-1} \times \left[1 + \frac{n-k_1-1}{k_1+2}\frac{p}{q} + \frac{(n-k_1-1)(n-k_1-2)}{(k_1+2)(k_1+3)}\frac{p^2}{q^2} + \ldots\right]$$

and this outside term may be evaluated quickly with the aid of tables of log-factorials and logarithms. It remains therefore to sum the finite series within the brackets which will be recognized as a particular case of the hypergeometric series

$$F(\alpha, \beta, \gamma, z) = 1 + \frac{\alpha.\beta}{\gamma}\frac{z}{1!} + \frac{\alpha(\alpha+1)\beta(\beta+1)}{\gamma(\gamma+1)}\frac{z^2}{2!} + \ldots,$$

where

$$\alpha = -n + k_1 + 1, \quad \beta = 1, \quad \gamma = k_1 + 2, \quad z = -p/q.$$

If we define

$$X_{2n} = F(\alpha+n, \beta+n, \gamma+2n, z),$$
$$X_{2n+1} = F(\alpha+n, \beta+n+1, \gamma+2n+1, z),$$

easy algebra will verify that

$$X_{2n} = F(\alpha+n-1, \beta+n, \gamma+2n-1, z)$$
$$+z\frac{(\beta+n)(\gamma+2n-1-\alpha-n+1)}{(\gamma+2n-1)(\gamma+2n)}F(\alpha+n, \beta+n+1, \gamma+2n+1, z)$$

or
$$X_{2n} = X_{2n-1}+z\frac{(\beta+n)(\gamma-\alpha+n)}{(\gamma+2n-1)(\gamma+2n)}X_{2n+1}.$$

Similarly
$$X_{2n+1} = X_{2n}+z\frac{(\alpha+n)(\gamma-\beta+n)}{(\gamma+2n)(\gamma+2n+1)}X_{2n+2}.$$

Writing a_{2n} and a_{2n+1} for the coefficients of X_{2n+1} and X_{2n+2}, respectively, we shall have generally

$$X_{v-1} = X_v - a_v X_{v+1} z.$$

Hence, if we give v values 1, 2, ... successively, we obtain a series of relationships between the X's the first of which will be

$$X_0 = X_1 - a_1 X_2 z \quad \text{or} \quad \frac{X_1}{X_0} = 1\Big/\left[1-a_1 z\frac{X_2}{X_1}\right].$$

By utilizing the successive relationships we may write this down as a continued fraction

$$\frac{X_1}{X_0} = \cfrac{1}{1-\cfrac{a_1 z}{1-\cfrac{a_2 z}{1-\ddots\,-\cfrac{a_v z}{\cfrac{X_v}{X_{v+1}}}}}}.$$

Let $X_0 = 1$ and

$$X_1 = F(\alpha, \beta+1, \gamma+1, z) = F(-n+k_1+1, 1, k_1+2, -p/q)$$

so that

$$\alpha = -n+k_1+1, \quad \beta = 0, \quad \gamma = k_1+1, \quad z = -p/q.$$

Writing
$$d_\omega = -za_{2\omega}, \quad c_\omega = za_{2\omega-1},$$

it may be shown by substitution that

$$d_\omega = \frac{\omega(n+\omega)}{(k_1+2\omega+1)(k_1+2\omega)}\frac{p}{q}, \quad c_\omega = \frac{(n-\omega-k_1)(k_1+\omega)}{(k_1+2\omega-1)(k_1+2\omega)}\frac{p}{q}.$$

Hence if S is the sum of the hypergeometric series which we are proceeding to evaluate, i.e. if

$$S = 1 + \frac{n-k_1-1}{k_1+2}\frac{p}{q} + \frac{(n-k_1-1)(n-k_1-2)}{(k_1+2)(k_1+3)}\frac{p^2}{q^2} + \ldots$$

then
$$S = \frac{X_1}{X_0} = \cfrac{1}{1-\cfrac{c_1}{1+\cfrac{d_1}{1-\cfrac{c_2}{1+\cfrac{d_2}{1-\ddots - \cfrac{c_{n-k_1-1}}{1+\cfrac{d_{n-k_1-1}}{1}}}}}}}.$$

Referring back to our definition of c_ω it will be seen that if c_1 is positive and less than unity so will be all the other c's. c_1 was defined as

$$c_1 = \frac{n-k_1-1}{k_1+2}\frac{p}{q}.$$

k_1 is essentially positive and less than n, p and q are positive fractions and n is a positive integer. It follows therefore that c_1 is positive or at least zero. For c_1 to be a fraction it is necessary that

$$(k_1+2)q > (n-k_1-1)p \quad \text{or} \quad k_1+2 > np+p.$$

It was assumed that $k_1 \geqslant np$, so the inequality holds good.

d_ω is essentially positive and, accordingly, if we consider the sum of the continued fraction

$$s_i = \cfrac{c_i}{1+\cfrac{d_i}{1-c_{i+1}\ddots}}$$

it is clear that $c_i > s_i > 0$

and that
$$s_i = \cfrac{c_i}{1+\cfrac{d_i}{1-s_{i+1}}}.$$

This last expression will give us all that is required for evaluating S. The necessary steps in the calculation will be:

(1) Choose i to be any number desired. Obviously the greater i the more accurate will be the approximation. Uspensky

suggests taking $i = 5$, but this is only something which may be learnt by experience and it may be possible to take i smaller than 5 for certain values of n. It will not pay to make i too large because the approximation will not then save a great deal of calculation.

(2) Having chosen i, calculate c_{i+1}, thus obtaining the upper limit to the inequality

$$c_{i+1} > s_{i+1} > 0.$$

(3) Calculate d_i and c_i and, using the limits for s_{i+1}, obtain limits for s_i from the relation

$$s_i = \frac{c_i}{1 + \dfrac{d_i}{1 - s_{i+1}}}.$$

(4) The calculations (3) are repeated to obtain successively limits for $s_{i-1}, s_{i-2}, \ldots, s_1, S$.

(5) The binomial term outside the bracket is evaluated either directly by logarithms or by means of an inequality discussed later.

(6) The multiplication of the results of calculations (4) and (5) give an approximation to the required sum of binomial probabilities.

Example. Find $\sum_{k=k_1+1}^{n} P_{n.k}$ when $n = 80$, $k_1 = 40$, $p = 0{\cdot}4$. This may be evaluated directly and exactly from the incomplete B-function ratio tables. We have

$$P\{k > k_1\} = I_{0{\cdot}4}(41, 40) = 0{\cdot}0271{,}236.$$

In practice we should not contemplate using the hypergeometric approximation if the exact value could be obtained from tables. However, for the purposes of illustration let us apply the theory which we have just outlined. First it is necessary to test whether the only assumption holds. Is $k_1 \geqslant np$? Here $k_1 = 40$ and $np = 32$ and we may therefore proceed.

(1) Following Uspensky let us choose $i = 5$.

(2) $0 < s_6 < 0{\cdot}39316$.

(3) and (4) give

ω	c_ω	d_ω	
5	0·42857	0·11111	$0{\cdot}36224 < s_5 < 0{\cdot}38571$
4	0·46809	0·09524	$0{\cdot}40526 < s_4 < 0{\cdot}40727$
3	0·51240	0·07678	$0{\cdot}45364 < s_3 < 0{\cdot}45381$
2	0·56237	0·05522	$0{\cdot}51073 < s_2 < 0{\cdot}51075$
1	0·61905	0·02990	$0{\cdot}58340 = s_1 = 0{\cdot}58340$

$S = 2{\cdot}40038$

(5) The binomial term

$$\frac{n!}{(k_1+1)!\,(n-k_1-1)!}p^{k_1+1}q^{n-k_1-1}$$

$$=\frac{80!}{41!\,39!}(0{\cdot}4)^{41}(0{\cdot}6)^{39}=0{\cdot}011300.$$

(6) The required sum of the binomial probabilities is 0·027124.

It will be seen that this agrees well with the value as calculated from the incomplete B-function ratio tables but the calculation involved is rather heavy. Even so the calculations are few in number compared with those which would be necessary in order to evaluate the binomial series term by term. The hypergeometric approximation involves no restricting assumptions and may be made as accurate as desired by increasing the size of i which is at choice. It should therefore be used for sums of binomial probabilities outside the range of the B-function ratio tables for which a definite accuracy is required. We shall discuss later other approximate methods which give the sums of binomial probabilities quickly, but these approximations (the normal curve and Poisson's limit, treated in succeeding chapters) rest on certain mathematical restrictive assumptions and it is not always possible to judge the accuracy of the results obtained from their use. While therefore they may be adequate for the rough determination of a probability level they are certainly not satisfactory for calculating a sum of terms when other calculations are to be based on the result.

Exercise. Given that the constant probability that an event will happen in a single trial is 1/3, find the probability that in 100 trials 45 or more events will happen.

Exercise. A halfpenny is tossed 200 times. The head fell uppermost on 153 occasions and the tail on 47. If it is assumed that the constant probability of head or tail in a single trial is 1/2, would you consider such an experimental result to be exceptional?

Exercise. Consider Weldon's dice experiment of the previous chapter. Evaluate the probability of obtaining 9 or more dice with 5 or 6 uppermost when 12 dice are thrown.

3. Approximate Evaluation of a Single Binomial Probability

When the n of the binomial series is small it is an easy matter to evaluate any given binomial term. When n is large it will be necessary to refer to tables of log-factorials and it is possible that occasions may arise when these are not readily available or, as is sometimes the case, the binomial index may lie outside the range of existing tables. It is rare however, that the computer has not access to tables of logarithms, and it is useful therefore to give inequalities for a single binomial probability when the binomial index is large. These inequalities will occur again in a slightly different form in the proof of Laplace's theorem (next chapter) and the approximation will accordingly serve the further purpose of making the reader familiar with them. It is required to evaluate

$$P_{n.k} = \frac{n!}{k!(n-k)!} p^k q^{n-k},$$

when n, k, and $n-k$ are all large numbers.

It is known by Stirling's theorem that

$$m! = (2\pi m)^{\frac{1}{2}} m^m e^{-m+\theta(m)},$$

where

$$\frac{1}{12m+6} < \theta(m) < \frac{1}{12m}.$$

Expanding the factorials in the expression for $P_{n.k}$ we have

$$\frac{P_{n.k}}{\left(\frac{n}{2\pi k(n-k)}\right)^{\frac{1}{2}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k}} = \exp\left[\theta(n) - \theta(k) - \theta(n-k)\right].$$

From the inequalities for $\theta(m)$ given above, since n is greater than k or $n-k$ it will follow that

$$\frac{1}{12k} + \frac{1}{12(n-k)} - \frac{1}{12n} > \theta(k) + \theta(n-k) - \theta(n) >$$

$$\frac{1}{12k+6} + \frac{1}{12(n-k)+6} - \frac{1}{12n+6}$$

and therefore that

$$\exp\left[\frac{1}{12n}-\frac{1}{12k}-\frac{1}{12(n-k)}\right]$$
$$<\exp\left[\theta(n)-\theta(k)-\theta(n-k)\right]<$$
$$\exp\left[\frac{1}{12n+6}-\frac{1}{12k+6}-\frac{1}{12(n-k)+6}\right].$$

Hence

$$\exp\left[\frac{1}{12n}-\frac{1}{12k}-\frac{1}{12(n-k)}\right]$$
$$<\frac{P_{n.k}}{\left(\frac{n}{2\pi k(n-k)}\right)^{\frac{1}{2}}\left(\frac{np}{k}\right)^{k}\left(\frac{nq}{n-k}\right)^{n-k}}<$$
$$\exp\left[\frac{1}{12n+6}-\frac{1}{12k+6}-\frac{1}{12(n-k)+6}\right],$$

all the terms of which can be evaluated by logarithms.

Example. Test this approximation for the binomial term

$$\frac{80!}{41!\,39!}(0{\cdot}4)^{41}(0{\cdot}6)^{39},$$

the exact value of which is 0·011330.

We begin by evaluating the divisor of $P_{n.k}$ by logarithms.

$$\left(\frac{n}{2\pi k(n-k)}\right)^{\frac{1}{2}}\left(\frac{np}{k}\right)^{k}\left(\frac{nq}{n-k}\right)^{n-k}=0{\cdot}011335.$$

The left- and right-hand sides of the inequality may be determined either from tables of the negative exponential function, or by logarithms, or in this case perhaps more easily from the first three terms of the negative exponential series.

Thus

$$0{\cdot}99688<\frac{P_{n.k}}{0{\cdot}011335}<0{\cdot}99692,$$

which gives $0{\cdot}011330=P_{n.k}=0{\cdot}011330.$

The approximation in this case agrees with the exact value to six decimal places.

It will be recognized that, provided tables of the log-factorials are available, there is no advantage in using this approximation in preference to calculating the exact value; for the arithmetic

involved in the use of the approximation is, if anything, a little heavier than for the exact value. If, however, no tables but ordinary logarithms are available then the approximation must perforce be used.

Exercise. Calculate the exact values of the binomial probabilities for

(i) $n = 20, \quad k = 11, \quad p = 0{\cdot}4;$

(ii) $n = 200, \quad k = 110, \quad p = 0{\cdot}4$

and compare with the approximate method.

REFERENCES AND READING

Further illustrations of the hypergeometric approximation may be found in J. V. Uspensky, *Introduction to Mathematical Probability*, whose proof of the conr exion between the hypergeometric approximation and the binomial series has been closely followed here. The connexion between the binomial series and the incomplete B-function ratio is well known. It was given by Karl Pearson who believed it to be new (Karl Pearson, *Biometrika*, XVI, p. 202, 'Note on the Relationship of the Incomplete B-function to the sum of the first p terms of the binomial $(a+b)^n$'), but it was almost certainly known before this century.

The hypergeometric approximation has been discussed by J. H. Müller (*Biometrika*, XXII, p. 284, 'Application of continued functions to the evaluation of certain integrals with special reference to the Incomplete B-function'). Müller believed his method to be original. It does not appear, however, to differ greatly from that outlined by Uspensky and attributed by him to Markoff.

CHAPTER V

REPLACEMENT OF THE BINOMIAL SERIES BY THE NORMAL CURVE

The fact that it is possible under certain conditions to replace a binomial series by a normal curve has been known for many years. The first derivation of the curve was given by Abram de Moivre (1718) in *The Doctrine of Chances* and he derived it in order to be able to express a certain sum of probabilities. The formula was not, however, stated explicitly as a theorem and it was not until the advent of Laplace (1812) with his *Théorie des Probabilités* that the relationship between the binomial series and the normal curve was given clear mathematical expression. Since Laplace it has become increasingly frequent for those seeking to give numerical form to the sum of a given number of binomial terms to refer to the normal approximation, and the tabled areas of the normal curve are used for such evaluations on occasions when the restrictions and assumptions of the approximation can hardly hope to be justified. We shall try to show at a later stage the limitations within which the application of the theorem may appear to be legitimate. The theorem may be stated in many forms. We shall state it in the following way:

Laplace's Theorem

If n is the number of absolutely independent trials, such that in each trial the probability of a certain event E is always equal to p, whatever the result of the preceding trials, and if k denotes the number of trials in which an event E occurs, then whatever the numbers z_1 and z_2, where $z_1 < z_2$, the probability

$$P\left\{z_1 \leqslant \frac{k-np}{\sqrt{(npq)}} \leqslant z_2\right\} \to \frac{1}{\sqrt{(2\pi)}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz \quad \text{as} \quad n \to \infty.$$

Before proceeding with the direct proof of the theorem it will be convenient to begin by stating and proving a lemma which will be necessary at different stages of the proof. This lemma appears to be due to Duhamel.

LEMMA. Consider two sequences of positive sums

$$S_1, \quad S_2, \quad \ldots, \quad S_n, \quad \ldots,$$
$$\Sigma_1, \quad \Sigma_2, \quad \ldots, \quad \Sigma_n, \quad \ldots,$$

such that

$$S_n = \sum_{i=1}^{N_n} P_{n.i}, \quad \Sigma_n = \sum_{i=1}^{N_n} P'_{n.i}, \quad \text{and} \quad \lim_{n\to\infty} N_n = \infty.$$

If, whatever $\epsilon > 0$, it is possible to find n_ϵ, such that

$$\left| \frac{P'_{n.i}}{P_{n.i}} - 1 \right| < \epsilon \quad \text{for} \quad i = 1, 2, \ldots, N_n \quad \text{and} \quad n > n_\epsilon$$

then the existence of a finite limit

$$\lim_{n\to\infty} S_n = S$$

implies that of Σ_n, namely

$$\lim_{n\to\infty} \Sigma_n = \lim_{n\to\infty} S_n = S$$

and conversely.

Proof of lemma. If S_n tends to a finite limit S then there exist fixed numbers M and n_M such that for any $n > n_M$, $S_n < M$. Write

$$P'_{n.k} - P_{n.k} = \epsilon_{n.k} P_{n.k}.$$

Sum and take absolute values. We have then

$$\left| \sum_{k-1}^{N_n} P'_{n.k} - \sum_{k=1}^{N_n} P_{n.k} \right| \leqslant \sum_{k=1}^{N_n} P_{n.k} |\epsilon_{n.k}|.$$

Consider any η as small as desired and write

$$\epsilon = \eta/M.$$

If n is greater than both n_M and n_ϵ then we shall have both

$$S_n < M \quad \text{and} \quad |\epsilon_{n.k}| < \epsilon.$$

Hence

$$\left| \sum_{k=1}^{N_n} P'_{n.k} - \sum_{k=1}^{N_n} P_{n.k} \right| \leqslant \sum_{k=1}^{N_n} P_{n.k} |\epsilon_{n.k}| < \epsilon \sum_{k=1}^{N_n} P_{n.k} < \epsilon M = \eta.$$

Now η may be chosen as small as desired. It follows therefore that if S_n tends to a finite limit S then Σ tends to the same limit and the lemma is proved.

Proof of theorem. We begin by rewriting the probability

$$P\left\{z_1 \leqslant \frac{k-np}{\sqrt{(npq)}} \leqslant z_1\right\} \quad \text{as} \quad P\{np + z_1\sqrt{(npq)} \leqslant k \leqslant np + z_2\sqrt{(npq)}\}.$$

Denote by k_1 the smallest integer such that

$$np + z_1\sqrt{(npq)} \leqslant k_1,$$

let k_2 be the largest integer such that

$$np + z_2\sqrt{(npq)} \geqslant k_2,$$

and substitute in the probability

$$P\{np + z_1\sqrt{(npq)} \leqslant k \leqslant np + z_2\sqrt{(npq)}\} = P\{k_1 \leqslant k \leqslant k_2\}$$
$$= P\{(k = k_1) + (k = k_1 + 1) + \ldots + (k = k_2)\}.$$

The probabilities that $k = k_1$ and the succeeding terms are binomial probabilities as given in the statement of the theorem. Hence

$$P\{(k = k_1) + (k = k_1 + 1) + \ldots + (k = k_2)\}$$
$$= \sum_{k=k_1}^{k_2} P_{n.k} = \sum_{k=k_1}^{k_2} \frac{n!}{k!(n-k)!} p^k q^{n-k}.$$

We want to find an approximation to this sum and we therefore look for an expression $P'_{n.k}$. If an expression $P'_{n.k}$ can be found such that

$$\frac{P'_{n.k}}{P_{n.k}} - 1 = \epsilon_{n.k}$$

and if for any number $\epsilon > 0$, where ϵ is as small as desired, we may find a number n_ϵ, such that for $n > n_\epsilon$ and $k_1 \leqslant k \leqslant k_2$

$$|\epsilon_{n.k}| < \epsilon$$

then, by Duhamel's lemma if there is a finite limit to $\sum_{k=k_1}^{k_2} P'_{n.k}$ there is a limit to $\sum_{k=k_1}^{k_2} P_{n.k}$ and these two limits will be equal.

Stirling's expansion for $n!$ may be written in the following form for n large

$$n! = n^n \exp\left[-\left(n - \frac{\theta_1}{12n}\right)\right]\sqrt{(2\pi n)},$$

where $0 < \theta_1 < 1$. Expanding the factorials in $P_{n.k}$ by means of this expression, after some arrangement we obtain

$$P_{n.k} = \frac{n!}{k!(n-k)!} p^k q^{n-k} = \left(\frac{np}{k}\right)^{k+\frac{1}{2}} \left(\frac{nq}{n-k}\right)^{n-k+\frac{1}{2}}$$
$$\times \frac{1}{\sqrt{(2\pi npq)}} \exp\left[\frac{\theta_1}{12n} - \frac{\theta_2}{12k} - \frac{\theta_3}{12(n-k)}\right],$$

where $0<\theta_1, \theta_2, \theta_3, <1$. Write $P'_{n.k}$ equal to the terms on the right-hand side which do not include the exponential and it follows that

$$\frac{P'_{n.k}}{P_{n.k}} = \exp\left[-\frac{\theta_1}{12n}+\frac{\theta_2}{12k}+\frac{\theta_3}{12(n-k)}\right].$$

Since n, k and $n-k$ are positive numbers and $0<\theta_1, \theta_2, \theta_3, <1$ it is clear that

$$-\frac{1}{12n}<-\frac{\theta_1}{12n}+\frac{\theta_2}{12k}+\frac{\theta_3}{12(n-k)}<\frac{1}{12k}+\frac{1}{12(n-k)},$$

but the right-hand side is dependent on k as well as n. Now by definition k_1 is the smallest integer such that

$$k_1 \geqslant np+z_1\sqrt{(npq)}$$

and it will follow therefore that

$$\frac{1}{12k_1} \leqslant \frac{1}{12(np+z_1\sqrt{(npq)})}.$$

Similarly from the definition of k_2 it will follow that

$$\frac{1}{12(n-k_2)} \leqslant \frac{1}{12(nq-z_2\sqrt{(npq)})}.$$

These inequalities will hold good for any k, such that $k_1 \leqslant k \leqslant k_2$, if we write k instead of k_1 and k_2. The fundamental inequality containing the θ's accordingly becomes

$$-\frac{1}{12n}<-\frac{\theta_1}{12n}+\frac{\theta_2}{12k}+\frac{\theta_3}{12(n-k)}<$$

$$\frac{1}{12(np+z_1\sqrt{(npq)})}+\frac{1}{12(nq-z_2\sqrt{(npq)})}.$$

We may now choose n as large as required so that the right- and left-hand sides of the inequality will differ from zero by as little as desired. It follows that

$$\exp\left[-\frac{\theta_1}{12n}+\frac{\theta_2}{12k}+\frac{\theta_3}{12(n-k)}\right]$$

will differ from unity by as little as desired and that $P'_{n.k}$ will therefore satisfy the conditions of the lemma. We may proceed,

then, to look for a limit to $\sum_{k=k_1}^{k_2} P'_{n.k}$ knowing that if it exists then it will be the desired limit of $\sum_{k=k_1}^{k_2} P_{n.k}$. $P'_{n.k}$ was defined as

$$P'_{n.k} = \frac{1}{\sqrt{(2\pi npq)}} \left(\frac{np}{k}\right)^{k+\frac{1}{2}} \left(\frac{nq}{n-k}\right)^{n-k+\frac{1}{2}}.$$

Write $$k = np + z\sqrt{(npq)},$$

where $$z_1 \leqslant z \leqslant z_2.$$

If we denote by Δz the difference between successive values of z, it follows, since k may take only integer values, that

$$k+1 = np + (z+\Delta z)\sqrt{(npq)},$$

from which it is obvious that

$$\frac{\Delta z}{\sqrt{(2\pi)}} = \frac{1}{\sqrt{(2\pi npq)}}$$

and that Δz tends to zero as n increases without limit. Rearrange $P'_{n.k}$ and take logarithms.

$$\log P'_{n.k} = \log \frac{\Delta z}{\sqrt{(2\pi)}} - (k+\tfrac{1}{2}) \log\left(1 + z\sqrt{\frac{q}{np}}\right) - (n-k+\tfrac{1}{2}) \log\left(1 - z\sqrt{\frac{p}{nq}}\right).$$

The last two terms on the right-hand side may be expanded as a series in the form

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3(1+\phi x)^3},$$

where $0 < |\phi| < 1$, from which, by writing $R_{n.z}$ for the collected terms of z^3, we shall have

$$\log P'_{n.k} = \log \frac{\Delta z}{\sqrt{(2\pi)}} - \frac{z}{2}\left(\sqrt{\frac{q}{np}} - \sqrt{\frac{p}{nq}}\right) - R_{n.z} z^3 - z^2\left[\frac{1}{2} - \frac{1}{4}\left(\frac{q^2+p^2}{npq}\right)\right].$$

The lemma of Duhamel may now be applied again. Let

$$\log P''_{n.k} = \log \frac{\Delta z}{\sqrt{(2\pi)}} - \tfrac{1}{2}z^2.$$

It will be seen that $P''_{n.k}/P'_{n.k}$ will differ from unity by as little as desired for some value of n greater than a given number. From the lemma, therefore, if there is a limit to $\sum_{k=k_1}^{k_2} P''_{n.k}$ it will also be the limit to $\sum_{k=k_1}^{k_2} P'_{n.k}$ and therefore to $\sum_{k=k_1}^{k_2} P_{n.k}$. Using the definition that a definite integral is the limit of a sum, we have

$$\lim_{n\to\infty} \sum_{k=k_1}^{k_2} P''_{n.k} = \lim_{n\to\infty} \sum_{k=k_1}^{k_2} \left(\frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}z^2}\Delta z\right) = \frac{1}{\sqrt{(2\pi)}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz$$

and therefore

$$\lim_{n\to\infty} \sum_{k=k_1}^{k_2} P_{n.k} = \frac{1}{\sqrt{(2\pi)}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz$$

and the theorem is proved.

The β_1 and β_2 of the binomial series were shown to be

$$\beta_1 = \frac{1}{npq} - \frac{4}{n}; \quad \beta_2 = 3 + \frac{1-6pq}{npq}.$$

Provided therefore that p remains finite, it can be seen that as n increases without limit β_1 will tend to 0 and β_2 to 3, that is to the β_1 and β_2 of the normal curve.

From the definitions given in statistical theory of the moments of a distribution there would appear to be no reason why the β_1 and β_2 of the distribution of a variate which may only take discrete values should not be calculated. Yet in making such a calculation the student should ever bear in mind exactly what it is he is doing. β_1 and β_2 are two measures devised by Karl Pearson to express skewness and flatness of frequency curves. Now the distribution of a binomial variate can never be a frequency curve. It consists of a discrete set of points and can never be the distribution of a continuous variate.

The fact that this is so, however, does not prevent us from seeking to express the sum of a number of binomial probabilities in terms of a continuous function. That we may do this was seen in the last chapter when it was shown that the sum of a number of binomial probabilities may be expressed exactly in terms of the incomplete B-function ratio. There is no reason why we should not express the sum of a number of binomial probabilities in like manner by means of an area of the normal curve. If it is

desired to find $P\{k \geqslant k_1\}$, where k is a binomial variate, then, with the customary notation, we may reduce this variable by its mean and standard deviation, i.e.

$$\frac{k-np}{\sqrt{(npq)}}$$

and refer this to the extensive existing tables of areas of the normal curve. Such a procedure is quite a legitimate one for it implies the conversion of one function so that entry may be made in the known tables of another function.

It may be shown simply, using the Euler-Maclaurin theorem, that a better approximation to the sum of a number of binomial terms may be made by entering the normal tables with

$$\frac{k-np \pm \frac{1}{2}}{\sqrt{(npq)}} \quad \text{instead of} \quad \frac{k-np}{\sqrt{(npq)}}.*$$

Thus we may give the rules that

$$P\{k \geqslant k_1\} \simeq \frac{1}{\sqrt{(2\pi)}} \int_{\frac{k_1-np-\frac{1}{2}}{\sqrt{(npq)}}}^{+\infty} e^{-\frac{1}{2}t^2} dt$$

and

$$P\{k \leqslant k_1\} \simeq \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\frac{k_1-np+\frac{1}{2}}{\sqrt{(npq)}}} e^{-\frac{1}{2}t^2} dt.$$

These rules may be memorized easily by imagining that the binomial probabilities may be represented by a frequency distribution in which $P_{n.k}$ is constant over the range $(k-\frac{1}{2})$ to $(k+\frac{1}{2})$. Such an assumption is, of course, wholly fallacious, since the binomial probabilities are a discrete set of points, but it is a useful aid to memory if not pursued too tenaciously. Table A below gives illustration of the effect of this corrective term.

The cases were chosen arbitrarily and it will be noted that in each case the effect of the corrective term is to bring the normal approximation more closely in line with the exact probability enumerated from the B-function ratio tables. Even for $n = 10$ the evaluation of the probability sum, as given by the corrected normal deviate, is not very different from the exact value. This

* The reader may compare this corrective factor of $\frac{1}{2\sqrt{(npq)}}$ with the 'continuity' correction for χ^2 in the case of a 2×2 table.

close correspondence for n comparatively small is of interest because so far we have not discussed the appropriate number below which n should not be taken for the normal approximation to have validity; in Laplace's theorem it will be remembered that the area of the normal curve only tends to represent the binomial sum as n increases without limit.

TABLE A. *Sum of binomial probabilities estimated by an area of the normal curve*

				$P\{k \geqslant k_1\}$		
n	p	q	k_1	Exact	Normal $\left(\frac{k_1-np}{\sqrt{(npq)}}\right)$	Normal $\left(\frac{k_1-np-\frac{1}{2}}{\sqrt{(npq)}}\right)$
10	0·3	0·7	5	0·1503	0·0838	0·1503
10	0·5	0·5	5	0·6230	0·5000	0·6241
10	0·7	0·3	5	0·9526	0·9162	0·9577
20	0·3	0·7	10	0·0480	0·0256	0·0436
20	0·5	0·5	10	0·5881	0·5000	0·5882
20	0·7	0·3	10	0·9829	0·9744	0·9858
30	0·3	0·7	15	0·0169	0·0084	0·0143
30	0·5	0·5	15	0·5722	0·5000	0·5721
30	0·7	0·3	15	0·9936	0·9916	0·9952

To obtain some idea of the degree of approximation involved is difficult in that there are various ways in which it may be desired to use the normal areas. It is conceivable that an estimate of each of the binomial probabilities in a given series may be required. If p is not very different from $\frac{1}{2}$, then for n as little as 5 the areas of the normal curve will agree with each of the binomial probabilities to within 1%. For $n = 10$ the correspondence is good. However, the main use to which we may expect to put the normal approximation is for tests of significance.

The statistician is concerned with the setting up of arbitrary probability levels whereby hypotheses may be tested. Because these levels are arbitrarily chosen, the acceptance or rejection of a hypothesis cannot be insisted on too rigorously if the calculated probability falls near the significance level. For example, if the 5 % level of significance is decided on *a priori*, and the calculated probability was found to be 0·057, the statistician would feel no more certain about the acceptance of the hypothesis under test than he would about its rejection for a calculated probability of

0·043. In the absence of further evidence he would consider that the issue was in doubt. It follows therefore that for approximate tests of significance which may be carried out first of all on raw material in order to make preliminary judgments, the reduced binomial variate

$$\frac{k-np-\frac{1}{2}}{\sqrt{(npq)}},$$

which it is recognized may not be exact, may be used in conjunction with the normal probability scales. This is advantageous in that the normal probability levels are easily remembered and a quick rough test of significance may therefore be made.

If the rejection level is 10 %, then n may be as small as 5 and p may vary from 0·1 to 0·9 if a variation of 4 % is allowed in the level. For instance, the error involved in assuming normality for $n = 5$ and $p = 0{\cdot}1$ is 2 % at the 10 % level. A smaller rejection level will not have much meaning for $n = 5$. For $n = 10$ and upwards with p varying from 0·3 to 0·7 the error involved is of the order of 1 % at the 5 % level. The lower the significance level chosen the more likely is the normal approximation to be in error and at the 0·005 level only the value $p = q = \frac{1}{2}$ will be found to lie within 0·004 to 0·006 for n as large as 50. However, provided it is remembered that the test is approximate then little error will result from its use. As a general rule it may be remembered that for $p < 0{\cdot}5$, whatever n, the normal test will tend to overemphasize the significance at the upper significance level, while for $p > 0{\cdot}5$ it will underestimate it.

Example. The proportion of male births within the whole population is of the order of 0·51. A family is observed composed of 10 males and no females. Assuming that order of birth does not affect the probability of being born male, is it considered that a family so constituted is exceptional? The probabilities of obtaining 0, 1, 2, ..., 10 males in 10 offspring will be given by the generating function $(0{\cdot}49+0{\cdot}51)^{10}$.

Hence the probability that out of 10 children all will be males will be given approximately by

$$\frac{10-10(0{\cdot}51)-0{\cdot}5}{(10\times 0{\cdot}51\times 0{\cdot}49)^{\frac{1}{2}}} = 2{\cdot}785$$

referred to tables of the normal probability integral. The

numerical figure for the probability is 0·0027. That is to say it might be expected that a family so constituted would occur less than 3 times in 1000 families composed of 10 offspring.

Example. From previous experience it has been found that when a river overflows on a road an average of six in every ten cars manage to get through the flood. Fifty cars attempt the crossing. What is the probability that thirty-five or more will get through?

Here the generating function is $(0{\cdot}4+0{\cdot}6)^{50}$ and

$$P\{k \geqslant 35\}$$

is desired. Using the normal approximation the deviate is

$$\frac{35-30-0{\cdot}5}{(30\times 0{\cdot}6\times 0{\cdot}4)^{\frac{1}{2}}} = 0{\cdot}89,$$

giving a normal probability of 0·19. That is to say, the chance is only approximately 4 to 1 that 35 or more cars will get through.

Example. At a given distance, x feet, from the explosion of a bomb, it is known that the probability of a pane of glass being smashed is 0·3. What is the probability that out of 100 panes of glass situated at x feet from the explosion 40 or more will be smashed?

The normal deviate is

$$\frac{40-30-0{\cdot}5}{(100\times 0{\cdot}3\times 0{\cdot}7)^{\frac{1}{2}}} = 2{\cdot}07$$

and the equivalent normal probability is 0·02. We should say therefore that it is doubtful whether so many panes of glass will be smashed at a distance x feet.

Example. If the chance of a house being hit by an incendiary bomb is 0·1 and if 25 bombs fall randomly within the area in which the house is situated, what is the probability that the house receives more than one bomb?

The normal deviate is

$$\frac{2-2{\cdot}5-0{\cdot}5}{(25\times 0{\cdot}1\times 0{\cdot}9)^{\frac{1}{2}}} = -0{\cdot}67$$

and the chance that the house receives more than one bomb is

therefore 0·75. The exact probability calculated from the B-function tables is

$$I_{0\cdot1}(2, 24) = 0\cdot74.$$

On three-quarters of the occasions on which 25 bombs fall in an equivalent area, the house will receive two or more bombs.

REFERENCES AND READING

A detailed development of the replacement of the sum of a number of binomial probabilities by the normal probability integral will be found in recent papers by S. Bernstein, where the error term is included.

A simplified version of Bernstein's theorem has been put forward by J. V. Uspensky, *Introduction to Mathematical Probability.*

A simplified version of the proof given in this chapter will be found in J. L. Coolidge, *Introduction to Mathematical Probability.*

Most statistical text-books use the normal approximation, with or without the corrective term, with a certain amount of indiscrimination. The student will find many examples to work through in such text-books.

CHAPTER VI

POISSON'S LIMIT FOR BINOMIAL PROBABILITIES

When p is very small but finite, it is necessary for n to be very large indeed before the normal integral will approximate to a sum of binomial probabilities. The incomplete B-function ratio tables do not extend below $p = 0{\cdot}01$ and it is necessary therefore to find another method of approximating to the required sum. This we do in Poisson's limit for binomial probabilities.

During a war the ideas of the individual regarding the fundamental sets on which probabilities are calculated are liable to vary according to whether he is exposed to immediate risk or not. For example, to the individual under fire the subjective probability set will be composed of two alternatives, survival or non-survival. To the statistician, however, calculating a probability on the total number of persons exposed to risk, the chance of any one individual being hurt was found to be very small. In fact, many chances for what might be expressed as war risks were found to be small and examples for which Poisson's limit to the binomial was valid were numerous.

Poisson's limit to the binomial, like the normal approximation, has been considerably misused in statistical practice. It is sometimes referred to as Poisson's law of Small Numbers; a bad nomenclature in that the limit loses its binomial parentage and may lead to misunderstanding of the fact that the 'law' is nothing more than a limit, under certain conditions, for binomial probabilities.

Poisson's Limit for Binomial Probabilities

Theorem. If n is the number of absolutely independent trials, and if in each trial the probability of an event E is constant and equal to p, where p is small but finite, and if k denote the number

of trials in which an event E occurs, then provided m and k remain finite

$$P_{n.k} \to \frac{m^k e^{-m}}{k!}$$

as n increases without limit, where $P_{n.k}$ has its usual meaning and $m = np$.

It is convenient to begin with a simple inequality.* If

$$1 \leqslant \alpha \leqslant \beta - 1,$$

then
$$\left(1-\frac{\beta}{n}\right) < \left(1-\frac{\alpha}{n}\right)\left(1-\frac{(\beta-\alpha)}{n}\right) \leqslant \left(1-\frac{\beta}{2n}\right)^2,$$

where n is a positive integer. This inequality is self-evident. From this inequality it is clear that if

$$T^2 = \prod_{\alpha=1}^{\beta-1}\left(1-\frac{\alpha}{n}\right)\left(1-\frac{(\beta-\alpha)}{n}\right),$$

then
$$\left(1-\frac{\beta}{n}\right)^{\beta-1} < T^2 < \left(1-\frac{\beta}{2n}\right)^{2\beta-2}$$

and
$$\left(1-\frac{\beta}{n}\right)^{\frac{1}{2}(\beta-1)} < T < \left(1-\frac{\beta}{2n}\right)^{\beta-1}.$$

We shall find it necessary to use the inequality in this form.

$P_{n.k}$ is defined as

$$P_{n.k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}.$$

Writing $m = np$ and rearranging

$$P_{n.k} = \frac{1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{(k-1)}{n}\right)}{\left(1-\frac{m}{n}\right)\left(1-\frac{m}{n}\right)\cdots\left(1-\frac{m}{n}\right)} \frac{m^k}{k!}\left(1-\frac{m}{n}\right)^n$$

$$= \left(1-\frac{m}{n}\right)^{n-k} \frac{m^k}{k!} \prod_{t=0}^{k-1}\left(1-\frac{t}{n}\right).$$

Applying the fundamental inequality we may obtain an inequality for $P_{n.k}$.

$$\left(1-\frac{m}{n}\right)^{n-k} \frac{m^k}{k!}\left(1-\frac{k}{n}\right)^{\frac{1}{2}(k-1)} < P_{n.k} < \left(1-\frac{m}{n}\right)^{n-k} \frac{m^k}{k!}\left(1-\frac{k}{2n}\right)^{k-1}.$$

* See J. V. Uspensky, *Introduction to Mathematical Probability*, pp. 135–7. The proof given here follows his outline.

Since k, m and n are positive it may be shown, by means of a transformation of the type

$$\left(1-\frac{m}{n}\right)^n = \exp\left[n\log_e\left(1-\frac{m}{n}\right)\right],$$

that $$\left(1-\frac{m}{n}\right)^{n-k} < \exp\left[-m+\frac{k}{n}m-\frac{n-k}{2n^2}m^2\right]$$

and $$\left(1-\frac{k}{2n}\right)^{k-1} < \exp\left[-\frac{k(k-1)}{2n}\right].$$

The right-hand side of the inequality for $P_{n.k}$ may therefore be rewritten as follows:

$$P_{n.k} < \frac{m^k}{k!}e^{-m}\exp\left[\frac{k}{n}m-\frac{n-k}{2n^2}m^2-\frac{k(k-1)}{2n}\right].$$

Consider now the left-hand side of the inequality for $P_{n.k}$ and use the same device as before.

$$\left(1-\frac{m}{n}\right)^{n-k} = \left(1+\frac{m}{n-m}\right)^{-(n-k)}$$

$$> \exp\left[-m+\frac{m(k-m)}{n-m}+\frac{n-k}{2}\frac{m^2}{(n-m)^2}-\frac{(n-k)}{3}\frac{m^3}{(n-m)^3}\right]$$

and $$\left(1-\frac{k}{n}\right)^{\frac{1}{2}(k-1)} = \left(1+\frac{k}{n-k}\right)^{-\frac{1}{2}(k-1)} > \exp\left[-\frac{k(k-1)}{2(n-k)}\right],$$

from which it follows that

$$P_{n.k} > \frac{m^k}{k!}e^{-m}\exp\left[\frac{km}{n}-\frac{n-k}{2}\frac{m^2}{n^2}-\frac{k(k-1)}{2n}\right]$$

$$\times\left[1-\theta\left(\frac{n-k}{3}\frac{m^3}{(n-m)^3}+\frac{k^3}{2n(n-k)}\right)\right],$$

where $0<\theta<1$. If therefore we write

$$\psi = \exp\left[\frac{km}{n}-\frac{n-k}{2}\frac{m^2}{n^2}-\frac{k(k-1)}{2n}\right]$$

and $$\phi = 1-\theta\left[\frac{n-k}{3}\frac{m^3}{(n-m)^3}+\frac{k^3}{2n(n-k)}\right],$$

the inequality for $P_{n.k}$ becomes

$$\psi < \frac{P_{n.k}}{\frac{m^k}{k!}e^{-m}} < \psi.\phi.$$

Now provided k and m remain finite as n increases, ψ and θ will differ from unity by as little as desired, and therefore both sides of the inequality will differ from unity by as little as desired.

It follows that

$$P_{n.k} \to \frac{m^k}{k!} e^{-m} \quad \text{as} \quad n \to \infty$$

provided both m and k are finite.

Under the assumptions through which the Poisson limit is reached it would be possible intuitively to write down its moments. They are, however, quickly reached by the elementary method previously used for binomial probabilities. If

$$P_m = \frac{m^k}{k!} e^{-m} = \lim_{n \to \infty} P_{n.k} \quad (k \text{ and } m \text{ finite}),$$

then

$$\mu'_r = \sum_{k=0}^{\infty} \frac{m^k}{k!} e^{-m} k^r,$$

from which it is seen that

$$\mu'_1 = m, \quad \mu_2 = m, \quad \mu_3 = m, \quad \mu_4 = 3m^2 + m$$

and

$$\beta_1 = \frac{1}{m}, \quad \beta_2 = 3 + \frac{1}{m}.$$

One point should be noticed here. The summation for moments was taken over the range $k = 0$ to $k = +\infty$ and not, as is strictly correct, over the range $k = 0$ to $k = n$. It is necessary to do this because

$$\sum_{k=0}^{\infty} \frac{m^k}{k!} e^{-m} = 1, \quad \text{while} \quad \sum_{k=0}^{n} \frac{m^k}{k!} e^{-m} \neq 1.$$

The approximation involved by this alteration of the limit of the summation sign is, however, of negligible proportions as may easily be shown by an actual calculation of the terms involved.

The Poisson limit has several advantages over the true binomial probabilities provided the conditions laid down for its use are justified. The incomplete B-function ratio tables are actually tabulated for arguments of 0·01 of the constant probability p, but where p is less than 0·01 interpolation into the tables becomes difficult indeed. The Poisson Limit is extensively tabled and the extraction of probabilities for p small is thus a simple procedure. In mathematical form it is more tractable to handle than the true binomial probability, while for arithmetical

purposes the fact that the mean and variance are equal means that only one set of calculations need be carried out.

Example. Compare the true binomial probabilities and those obtained by assuming Poisson's limit for $n = 10$ and $p = 0{\cdot}1$.

k		0	1	2	3
$I_p(k, n-k+1) - I_p(k+1, n-k)$	Binomial	0·34868	0·44631	0·20501	0·05889
$m^k e^{-m}/k!$	Poisson	0·36788	0·36788	0·18394	0·06131
k		4	5	6	7
$I_p(k, n-k+1) - I_p(k+1, n-k)$	Binomial	0·01130	0·00150	0·00014	0·00001
$m^k e^{-m}/k!$	Poisson	0·01533	0·00307	0·00051	0·00007
k		8	9	10	Total
$I_p(k, n-k+1) - I_p(k+1, n-k)$	Binomial	0·00000	0·00000	0·00000	1·00000
$m^k e^{-m}/k!$	Poisson	0·00001	0·00000	0·00000	1·00000

At first sight the agreement between the two series does not seem too good. This is partly because of the large number of decimal places taken; for considering that n is only equal to 10 the agreement when the first two decimal places only are taken is as close as could be expected. For p smaller than 0·1 the agreement between the exact calculated values and the Poisson limit should be closer, as also for the same p of 0·1 and larger n. It is well to note, however, that such divergences do occur. Poisson's limit should not be applied blindly and without due regard for the conditions of the theorem.

Example. A caterpillar $2x$ inches long starts to cross at right angles a one-way cycle-track T yards wide at a speed of f feet per second. If cycles are passing this particular spot at random intervals but at an average rate of n per second, what is the probability that the caterpillar reaches the other side safely? It may be assumed that the impress of a cycle tyre on the ground is equal to t inches and that the caterpillar may only be touched to be considered hurt.

If the caterpillar is $2x$ inches long and the impress of the tyre on the road is t inches then the best estimate we can make of the probability that a caterpillar will be run over by a single cycle is

$$p = \frac{2x+t}{36T}.$$

If the speed of the caterpillar is f feet per second and the track is T yards wide the caterpillar will take $3T/f$ seconds to cross the track. Moreover, if cycles are passing with an average frequency of n per second during these $3T/f$ seconds an average of $3Tn/f$ cycles will pass the given spot. We require the chance that none of these cycles do more than graze the caterpillar, that is we require

$$\left(1-\frac{2x+t}{36T}\right)^{3Tn/f}.$$

Generally the probability of one cycle running over the caterpillar will be very small so that the Poisson limit will give us the approximate evaluation of this chance as

$$\left(1-\frac{2x+t}{36T}\right)^{3Tn/f} \doteq \exp\left[-\frac{n}{12f}(2x+t)\right].$$

It will be remembered that this answer is correct only if the original estimate concerning the single constant probability is correct. We have no means of judging from the question whether this is so.

Example. Over a generation of students have been amused by Bortkewitsch's data, now classical, of deaths from kicks by a horse in 10 Prussian Army Corps during 20 years. The material he gave was as follows:

Actual deaths per corps	Frequency observed	Frequency Poisson's limit
0	109	109
1	65	66
2	22	20
3	3	4
4	1	1
5	—	—
Total	200	200

The mean of the observed frequency is 0·61 from which we have

$$m = np = 0{\cdot}61; \quad p = 0{\cdot}003.$$

The chance of being killed by a kick from a horse in any one year in any one Army corps is therefore extremely small. The frequency using Poisson's limit may be found directly from tables (Molina) entering with m. It may well have been that at the time at which

Bortkewitsch wrote conditions were sufficiently stable to allow him to consider that p could be constant over a period of 20 years. This state of affairs is, however, hardly likely to obtain to-day and it should be remembered in the grouping of such data that the fundamental assumption is that there is a small but constant probability that the event will happen in any one trial.

Exercise. The emission of α-particles from a radioactive substance was measured by Rutherford and Geiger. If t is the number of particles observed in units of time of $\frac{1}{8}$ minute, and if n_t is the number of intervals in which t particles were observed then the experimental results may be expressed by the following table:

t	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
n_t	57	203	383	525	532	408	273	139	49	27	10	4	—	1	1	2612

Calculate the probability that a single particle will be observed in a single time unit and fit Poisson's limit to the observed frequencies if it is considered justifiable.

Example. A man is standing some distance from blasting operations in a quarry. Five small pieces of debris fall randomly within a space of 10 sq.yd. If the plan area of a man is 2 sq.ft., what is the probability that if he were standing within the 10 sq.yd. he would not have been hit by any of the pieces of debris?

It is stated that the debris fall randomly within 10 sq.yd. This implies that of the 90 sq.ft. any one part is as likely to be hit as any other, and that we may say therefore that the probability of any 2 sq.ft. receiving a single piece of debris is

$$p = \tfrac{2}{90}.$$

Five pieces of debris are received in 10 sq.yd. Hence the probability that none of these is received by any particular 2 sq.ft. is

$$(1-\tfrac{2}{90})^5 \simeq e^{-\frac{1}{9}}.$$

Conversely the probability that he will be hit by at least one piece is

$$1-(1-\tfrac{2}{90})^5 \simeq 1-e^{-\frac{1}{9}}.$$

We have noted previously that the fundamental assumption of the binomial theorem on probabilities and therefore of Poisson's

limit is that the probability is constant from trial to trial. This does not mean that if an observed series of frequencies is found to be graduated closely by a Poisson series the underlying probability must necessarily be constant, although given full information it might be a reasonable hypothesis to make. If the probability of a single event varies considerably from trial to trial so that the material collected is not homogeneous, Poisson's limit may be a good fit to the collected data, but it is possible that a series known as the negative binomial may be more appropriate. It is easy to see how such a case might arise. We have seen that for the binomial

$$np = \mu_1', \quad npq = \mu_2,$$

whence $q = \mu_2/\mu_1'$. If μ_2 is greater than μ_1' then q is greater than unity and $p = 1-q$ is negative. Since μ_1' is positive this implies that n must be negative also and that a series of the form

$$(q+(-p))^{-n}$$

would be the appropriate one to fit. In certain cases a negative binomial may be expected from hypothesis, as in the theory given by Yule for the proportion of a population dying after the nth exposure to a disease. Unless, however, some such theoretical background can be constructed for a problem the advantage of fitting such a series may be questioned. The reader must remember that p is the probability that an event will occur in a single trial and is *a priori* postulated as lying between 0 and 1. n is the number of trials and must be a positive integer. It would appear wrong therefore to carry out calculations in which p and n are given negative values. Further, the object of fitting a series to obtain any graduations of observed frequency is to enable conclusions to be drawn from such a fitting, and it is difficult to see what conclusions could be drawn from the fitting of a negative binomial unless it is expected on theoretical grounds.

It was first pointed out by 'Student' that series in which the variance is greater than the mean arise from the probability p not remaining constant from trial to trial. These situations are not uncommon in bacteriological work. 'Student' found that the distribution of cells in a haemacytometer did not follow a Poisson series although it might reasonably be expected to do so. The hypothesis put forward to 'explain' this was that the presence

of one cell in a square of the haemacytometer altered the probability that there would be another, owing to the first exerting some attraction on the second, and so on. Karl Pearson writes, 'if two or more Poisson series be combined term by term *from the first*, then the compound will always* be a negative binomial' and remarks that this theorem was suggested to him by 'Student'. It is, however, not altogether certain that the converse of the theorem will hold good.

Pearson actually considered only the case of graduating the negative binomial by two separate Poisson series, which may account for the fact that this method is not always satisfactory in practice. His object was to obtain some means whereby a satisfactory interpretation could be given to the observational data. Actually, of course, unless there is an *a priori* reason to expect such a dichotomy, the splitting of the data into two Poisson series may not help very much.

Pearson's Theorem for the Graduation of the Negative Binomial

A series of N observations, for each of which the fundamental probability p may vary, may be described by the two Poisson series

$$\sum_{k=0}^{\infty}\left[v_1\left(\frac{m_1^k}{k!}e^{-m_1}\right)+v_2\left(\frac{m_2^k}{k!}e^{-m_2}\right)\right],$$

where m_1 and m_2 are the roots of the equation

$$m^2(a_2-a_1^2)-m(a_3-a_1a_2)+a_3a_1-a_2^2=0,$$

and a_1, a_2, a_3 are obtained from the moments of the series of N observations by means of the relationships

$$a_1=\mu_1', \quad a_2=\mu_2'-\mu_1', \quad a_3=\mu_3'-3\mu_2'+2\mu_1'$$

and

$$\frac{v_2}{N}=\frac{\mu_1'-m_2}{m_1-m_2}; \quad \frac{v_1}{N}=\frac{\mu_1'-m_1}{m_2-m_1}.$$

The proof of the theorem is straightforward and may be left to the student as an exercise.

* This is not strictly true.

Example of use of the theorem. 'Student' gave the count of yeast cells in 400 squares of a haemacytometer in the following table:

Number of yeast cells	0	1	2	3	4	5	Total
Frequency	213	128	37	18	3	1	400

Here $\mu_1' = 0{\cdot}6825, \quad \mu_2 = 0{\cdot}8117, \quad \mu_3 = 1{\cdot}0876,$

giving $q = 1{\cdot}19, \quad p = -0{\cdot}19, \quad n = -3{\cdot}59,$

so that the negative binomial would be

$$400(1{\cdot}19 - 0{\cdot}19)^{-3{\cdot}59}.$$

Solving the equations given in the theorem, we obtain

$$v_1 = 237, \quad m_1 = 0{\cdot}385,$$
$$v_2 = 163, \quad m_2 = 1{\cdot}116,$$

whence by calculating out these two Poisson series a good fit is obtained.

No. of yeast cells	0	1	2	3	4	5	6	7
1st Poisson series	161·44	62·11	11·95	1·53	0·15	0·01	0·00	0·00
2nd Poisson series	53·32	59·52	32·22	12·36	3·45	0·77	0·14	0·02
Total	215	122	44	14	4	1		
Observed frequency	213	128	37	18	3	1	—	—

Many other negative binomials may be graduated in this way by the addition of two Poisson series. However, as we have noted, the dichotomy is not always satisfactory and possibly this is because more than two Poisson series may be required adequately to describe the observations.

Neyman has discussed what he terms a new class of 'contagious distributions' which, it seems, will be applicable to many types of heterogeneous data and which will moreover give at least as good a fit and probably a better fit than many of the existing series. The moments of the distribution may be derived by the reader at a later stage since they will follow most naturally from the application of the theory of characteristic functions. However, the practical use of the theorem is pertinent at this point and we shall therefore state it without proof.

Neyman's Theorem describing data in which $\mu_2 > \mu_1'$

A series of N observations, for each of which the fundamental probability, p, may vary, may be described by the series

$$P\{X = k+1\} = \frac{m_1 m_2 e^{-m_2}}{k+1} \sum_{t=0}^{k} \frac{m_2^t}{t!} P\{X = k-t\},$$

where $$P\{X = 0\} = \exp(-m_1(1-e^{-m_2}))$$

X being the number of "successes", 0, 1, 2, ...,

and $$m_2 = (\mu_2 - \mu_1')/\mu_1', \quad m_1 = \mu_1'/m_2.$$

μ_2 and μ_1' are calculated from the observations; m_1 and m_2 are essentially positive.

There appears to be no reason why this series should not fit adequately all binomial type series for which μ_2 is greater than μ_1'. Since the positive binomial will probably be sufficient when μ_2 is less than μ_1', and the simple Poisson for μ_2 approximately equal to μ_1', it will be seen that Neyman's series extends the range of theoretical distributions necessary for describing frequency distributions for which $\mu_2 > \mu_1'$.

We must remark, however, as for the negative binomial, that the fitting of the series will only be of practical use provided the estimated parameters are capable of physical interpretation.

Example. Greenwood and Yule give a table of frequency of accidents in 5 weeks to 647 women working on H.E. shells. A simple Poisson distribution fitted to these figures does not graduate the observed frequencies very well, the reason being, it is supposed, that accident proneness is different for different

Number of accidents	Observed frequency	Poisson distribution	Negative binomial distribution	Neyman's series
0	447	406	442	448
1	132	189	140	128
2	42	45	45	49
3	21	7	14	16
4	3	1	5	5
5	2	0·1	2	1
Total	647	648	648	647

women. Greenwood and Yule fit a type of negative binomial to the material and obtain a good fit. The writer has fitted Neyman's series to the same material and it will be seen that this series gives a slightly better fit than the negative binomial. It is, however, difficult to see what the parameters of either distribution mean. The only drawback to the use of Neyman's series would appear to be the relatively heavy computation which is involved if the k of the series is over 10 (say). However, this is a slight failing, because it is rare that series of this type are found with k very large and in any case it should not be difficult to devise a suitable computational scheme.

Exercise. The number of defective teeth in alien Jewish children (boys) aged 12 years is given in the table below:

t = No. of teeth affected	0	1	2	3	4	5	6	7	8	9	10	11	12	Total
n_t = No. of boys with t teeth affected	73	56	37	52	31	18	22	6	9	3	2	2	2	313

Fit (*a*) Poisson's series, (*b*) the negative binomial series, (*c*) Neyman's contagious series to this material. Can you suggest any reason why the variance is larger than might have been expected?

REFERENCES AND READING

The derivation of Poisson's limit to the binomial probability will be found in any statistical text-book. Possibly the most satisfying is that given by J. V. Uspensky, *Introduction to Mathematical Probability*, but there are various methods at choice. Poisson's limit is extensively tabled in E. C. Molina, *Poisson's Exponential Binomial Limit.*

The negative binomial is discussed by 'Student', *Collected Papers*, no. 9, and may be found in

G. U. Yule and M. G. Kendall, *Introduction to the Theory of Statistics*,

M. G. Kendall, *The Advanced Theory of Statistics*, Vol. I. Kendall discusses Yule's survivor theory on pp. 125–6 of Vol. I.

Karl Pearson's paper on the graduation of the negative binomial by two Poisson series may be found in *Biometrika*, XI, p. 139.

Neyman's derivation of his contagious distribution may be found in *Annals of Mathematical Statistics*, X, p. 35.

The description of heterogeneous material is a statistical problem of great interest and the student should read the papers mentioned above and others to which these will lead him.

CHAPTER VII

PROBABILITIES *A POSTERIORI* CONFIDENCE LIMITS

In the first chapter we defined the population probability of a characteristic as being the proportion of units possessing that characteristic within the population. It is clear therefore that if the population probability is known then it is possible to specify all the various compositions which a sample drawn from that population may have and the probability that each of these compositions will arise in the random drawing of a single sample. For example, when considering a pack of 52 cards, if the probability of drawing any one card is 1/52 and five cards are drawn randomly from the pack, all variations of the 5 cards can be enumerated and the probability of drawing any one particular set may be calculated. Thus from a knowledge of the population we are able to specify the probability of the sample and the most probable composition of the sample. Such probabilities are often referred to as probabilities *a priori* in that prior knowledge of the population probability is necessary for their evaluation.

We now turn to what are termed probabilities *a posteriori* and we find that the position is the reverse of the *a priori* probabilities. Now all that is known is the composition of the sample and it is required from this knowledge to estimate the most probable composition of the population from which it has been drawn. Obviously if repeated sampling could be carried out then the composition of the parent population, i.e. the proportion of individuals possessing the characteristic A, can be estimated very nearly. However, it is not always possible for this repeated sampling to be carried out and we shall therefore discuss methods of estimating a population probability which have been put forward in the past (and mostly rejected), and the present-day method of confidence intervals.

For many years the centre of statistical controversy was centred around the theorem on probabilities *a posteriori*, loosely spoken of as Bayes' theorem. Thomas Bayes himself may have

had doubts about the validity of the application of his theorem. At any rate he withheld its publication and it was only after his death that it was found among his papers and communicated to the Royal Society by his friend Richard Price in 1763. Laplace incorporated it in his *Théorie Analytique des Probabilités* and used the theorem in a way which we cannot at this present time consider justifiable. However, since the theorem was given the weight of Laplace's authority, the validity of its application was assumed by many writers of the nineteenth century; we find famous statisticians such as Karl Pearson and Edgeworth defending it, and it still holds a prominent place in such elementary algebra text-books as have chapters on probability. Yet the modern point of view is that, strictly speaking, the application of the theorem in statistical method is wholly fallacious except under very restricted conditions.

Bayes' Theorem

An event E may happen only if one of the set $E_1, E_2, \ldots, E_k$, of mutually exclusive and only possible events occurs. The probability of the event E_t, given that E has occurred, is given by

$$P\{E_t \mid E\} = \frac{P\{E_t\}\, P\{E \mid E_t\}}{\sum_{j=1}^{k} P\{E_j\}\, P\{E \mid E_j\}}.$$

$P\{E_t \mid E\}$ is spoken of as the *a posteriori* probability of the event E_t.

Proof of Theorem. The proof of the theorem is simple.

$$P\{E.E_t\} = P\{E\}\, P\{E_t \mid E\} = P\{E_t\}\, P\{E \mid E_t\},$$

whence
$$P\{E_t \mid E\} = P\{E_t\}\, P\{E \mid E_t\}/P\{E\}.$$

It is stated that E may only occur if one of the set $E_1, E_2, \ldots, E_k$, occurs, and since these mutually exclusive events are also the only possible ones,

$$P\{E\} = P\{E.E_1\} + P\{E.E_2\} + \ldots + P\{E.E_k\}.$$

Each of the probabilities on the right-hand side may be expanded as before, for example,

$$P\{E.E_1\} = P\{E_1\}\, P\{E \mid E_1\}$$

and the proof of the theorem follows.

In the statement of the theorem we have written of a set $E_1, E_2, \ldots, E_k$ of mutually exclusive and only possible events. It

is, however, possible to speak of them as a set of hypotheses, and for this reason Bayes' theorem is sometimes referred to as a formula for the probability of hypotheses and sometimes as a theorem on the probability of causes.

However, no matter what the title of the theorem, it is clear that unless $P\{E_i\}$, i.e. the prior probability of the event E_i, is known, the application of the formula cannot be valid. Thus, if we regard the event E_i as the hypothesis that the sample E has been drawn from one of a given set of populations, it is clear that the probability of this hypothesis will rarely be known. If the composition of the super population generating the set of populations is known, or, in the language of the theorem, if the prior probability of the event E_i is known, then the validity of the application of the theorem is not in question; but we must then consider what occasion would arise in statistical practice in which it is necessary to calculate a further probability.

If $P\{E_i\}$ is not known, it follows that some assumption regarding its value must be made before $P\{E_i \mid E\}$ can be calculated. There is no reason why this assumption should not be made, but the fact which is often overlooked is that, given such an assumption, $P\{E_i \mid E\}$ will only be correct under this assumption and will vary according to the nature of the assumption. It has been customary to assume that all compositions of the populations $E_1, E_2, \ldots, E_k$ are equally probable, and from this to draw inferences regarding the most probable composition of the population from which the sample has been drawn. It is legitimate to make this assumption, but if it is made then it should be stated that under the assumption that all population compositions (all hypotheses, all causes) are equally likely, the probability that E is associated with E_i is a certain value.

Possibly it is unnecessary to labour this point further for at the present time there are few adherents of Bayes' theorem. We shall consider some examples for which the application of Bayes' theorem is valid, and some for which we shall show that the probability will vary according to the original hypothesis regarding the populations.

Example. Assume that there are three urns each containing a certain number of balls. The first urn contains 1 white, 2 red and 3 black balls; the second 2 white, 3 red and 1 black; and the

third 3 white, 1 red and 2 black. The balls are indistinguishable one from another, except for colour, and it may be assumed that the probability of drawing one given ball from any urn is 1/6. An urn is chosen at random and from it two balls are chosen at random. These two balls are one red and one white. What is the probability that they came from the second urn?

If the urn is selected at random then all three urns are equally probable and we have

$$P\{E_1\} = P\{\text{Urn } 1\} = \tfrac{1}{3}, \quad P\{E_2\} = P\{\text{Urn } 2\} = \tfrac{1}{3},$$
$$P\{E_3\} = P\{\text{Urn } 3\} = \tfrac{1}{3}.$$

The event E consists of the drawing of two balls, 1 white and 1 red, from either E_1 or E_2 or E_3.

$$P\{E \mid E_1\} = \frac{\text{number of ways in which 2 balls drawn from } E_1 \text{ can be 1 white and 1 red}}{\text{total number of ways in which 2 balls can be drawn from } E_1}$$
$$= 2\,\frac{2!\,4!}{6!} = \frac{2}{15}.$$

Similarly

$$P\{E \mid E_2\} = 6\,\frac{2!\,4!}{6!} = \frac{2}{5}, \quad P\{E \mid E_3\} = 3\,\frac{2!\,4!}{6!} = \frac{1}{5}.$$

By Bayes' theorem

$$P\{E_t \mid E\} = \frac{P\{E_t\}\,P\{E \mid E_t\}}{\sum_{j=1}^{3} P\{E_j\}\,P\{E \mid E_j\}}$$

and we have

$$P\{E_1 \mid E\} = \tfrac{2}{11}, \quad P\{E_2 \mid E\} = \tfrac{6}{11}, \quad P\{E_3 \mid E\} = \tfrac{3}{11}.$$

It is a little uncertain how such probabilities may be interpreted.

Example. A box contains a very large number of identical balls one-half of which are coloured white and the rest black. From this population ten balls are chosen randomly and put in another box and the result of drawing 5 balls randomly with replacement from these ten is that 4 showed black and 1 white. What is the most probable composition of the ten balls?

The probability that there are k white balls in the ten is, from the description 'very large' population,

$$\frac{10!}{k!(10-k)!}\,\frac{1}{2^{10}} = P\{E_k\}.$$

If there are k white balls in the population then the probability that 5 balls drawn from the ten with replacement will show 4 black and 1 white will be

$$P\{E \mid E_k\} = \frac{5!}{4!\,1!}\left(\frac{k}{10}\right)^1\left(1-\frac{k}{10}\right)^4.$$

Applying Bayes' theorem and reducing, we have

$$P\{E_k \mid E\} = \frac{10!}{k!(10-k)!}\left(\frac{k}{10}\right)^1\left(1-\frac{k}{10}\right)^4 \Big/ \sum_{k=1}^{9} \frac{10!}{k!(10-k)!}\left(\frac{k}{10}\right)^1\left(1-\frac{k}{10}\right)^4.$$

Letting k take in turn values 1, 2, 3, ..., 7, 8, 9 (we exclude zero because it is known that one white ball is among the ten, and 10 because at least one black ball is known to be present), we may draw up the following table:

k	1	2	3	4	5	6	7	8	9
$P\{E_k \mid E\}$	0·02	0·11	0·24	0·30	0·22	0·09	0·02	0·00	0·00

The most probable composition of the ten balls is therefore four white and six black.

Coolidge gives an interesting illustration of the effect of two different hypotheses in problems of this type where the composition of the original population is not known. He propounds the following problem:

'An urn contains N identical balls, black and white, in unknown proportion. A ball is drawn out and replaced n times, the balls being mixed after each drawing, with the result that just r white balls are seen. What is the probability that the urn contains exactly R white balls?' As in the previous problem

$$P\{E \mid E_R\} = \frac{n!}{r!(n-r)!}\left(\frac{R}{N}\right)^r\left(1-\frac{R}{N}\right)^{n-r},$$

but now the probabilities of the compositions of the population are not known. We cannot therefore apply Bayes' theorem unless we make some hypothesis about these probabilities. Coolidge suggests

Hypothesis I. All compositions of the population are equally likely.

Hypothesis II. The population has been formed by drawing balls at random from a super-population in which black and white are of equal proportions.

The student may invent other hypotheses for himself.

For hypothesis I we are given

$$P\{E_R\} = 1/N - 1,$$

for we rule out the cases that all are white and that all are black. We have therefore

$$P\{E_R \mid E\} = \binom{R}{N}^r \left(1 - \frac{R}{N}\right)^{n-r} \Big/ \sum_{R=1}^{N-1} \binom{R}{N}^r \left(1 - \frac{R}{N}\right)^{n-r}$$

and the most probable composition of the population, given hypothesis I, is

$$\frac{R}{N} = \frac{r}{n}.$$

In other words, the proportion observed in the sample is the most probable proportion in the population.

For hypothesis II we are given

$$P\{E_R\} = \frac{N!}{R!(N-R)!} \left(\frac{1}{2}\right)^N,$$

which gives

$$P\{E_R \mid E\} = \frac{\frac{N!}{R!(N-R)!} \binom{R}{N}^r \left(1 - \frac{R}{N}\right)^{n-r}}{\sum_{R=1}^{N-1} \frac{N!}{R!(N-R)!} \binom{R}{N}^r \left(1 - \frac{R}{N}\right)^{n-r}},$$

whence by finding the value of R/N which maximizes this expression we deduce that the most probable composition of the N balls from which the n balls were drawn with replacement, is

$$\frac{R}{N} = \frac{1}{2} \frac{N+2r}{N+n}.$$

Thus by making two different assumptions regarding the composition of the probabilities of the compositions of the population we are led to two different conclusions. If we take the population as $N = 10$, the sample as $n = 5$ and the number of white in the sample as $r = 1$ we shall have for the most probable composition of the population

Hypothesis I $\frac{R}{N} = \frac{2}{10}$. *Hypothesis II* $\frac{R}{N} = \frac{4}{10}$.

Both these results are correct if the original hypotheses are accepted but neither is correct if we limit ourselves strictly to the conditions of the problem which states that the composition of the population is not known.

This example illustrates clearly the fallaciousness of Bayes' theorem as it is generally applied. We shall take the point of view that in general Bayes' theorem will not be applicable in statistical work. An exception to this, which will be discussed in a later chapter, is the application to Mendelian hypotheses.

The statistical problem associated with Bayes' theorem is one which touches statisticians very nearly. Much of the working career of a present-day mathematical statistician is spent in attempting to draw valid inferences about a population when the only information available to him is that obtainable from a sample (or samples) drawn from that population. This need to draw inferences about the whole from the part has possibly always been felt by workers in probability, though not so markedly as to-day, and this may be the reason why the use of Bayes' theorem persisted for so many years; its inadequacy must have been recognized many times but it was used because no one could think of anything better.

The objective of the method of confidence intervals, a statistical concept which was devised to overcome the impasse created by the too liberal use of Bayes' theorem, is the estimation of limits within which we may be reasonably sure that a given population parameter will lie; these limits are estimated from the information provided by the sample. For example, since we have discussed the binomial theorem in some detail let us suppose that it is desired to estimate p, the proportion of individuals possessing a certain character in a given population, the only information at our disposal being the number f who possess that character in a sample of size n which has been randomly and independently drawn from the population. We should, possibly, for lack of anything better, take the proportion in the sample as an estimate of the population probability, but it is unnecessary to point out that this estimate will vary both according to the size of the sample and the number of samples available.

If p, the population probability, is known, then by the binomial theorem the probabilities of obtaining $0, 1, 2, \ldots, n$ units

possessing the given characteristic in a sample of size n can be enumerated. If any positive fraction ϵ is arbitrarily chosen, where $0 < \epsilon < 1$, two points f_1/n and f_2/n can be found such that

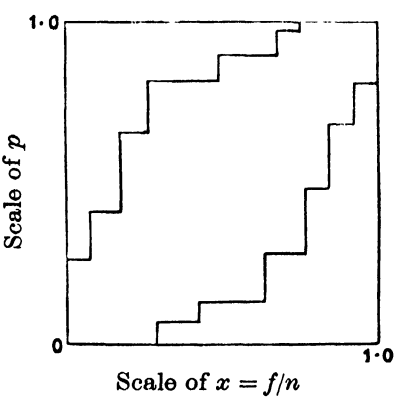
$$P\{x < f_1/n\} \leqslant \tfrac{1}{2}\epsilon \quad \text{and} \quad P\{x < f_2/n\} \leqslant \tfrac{1}{2}\epsilon,$$

whence

$$P\{f_1/n \leqslant x \leqslant f_2/n\} \geqslant 1 - \epsilon,$$

for given values of p and n. If $P\{x \mid p\}$ had been a continuous function then it would have been possible to choose f_1 and f_2 so that each of the above probabilities was exactly equal to $\frac{1}{2}\epsilon$. Since, however, for the binomial probabilities $P\{x \mid p\}$ is discontinuous, it becomes necessary to choose f as the nearest integer satisfying the inequality.

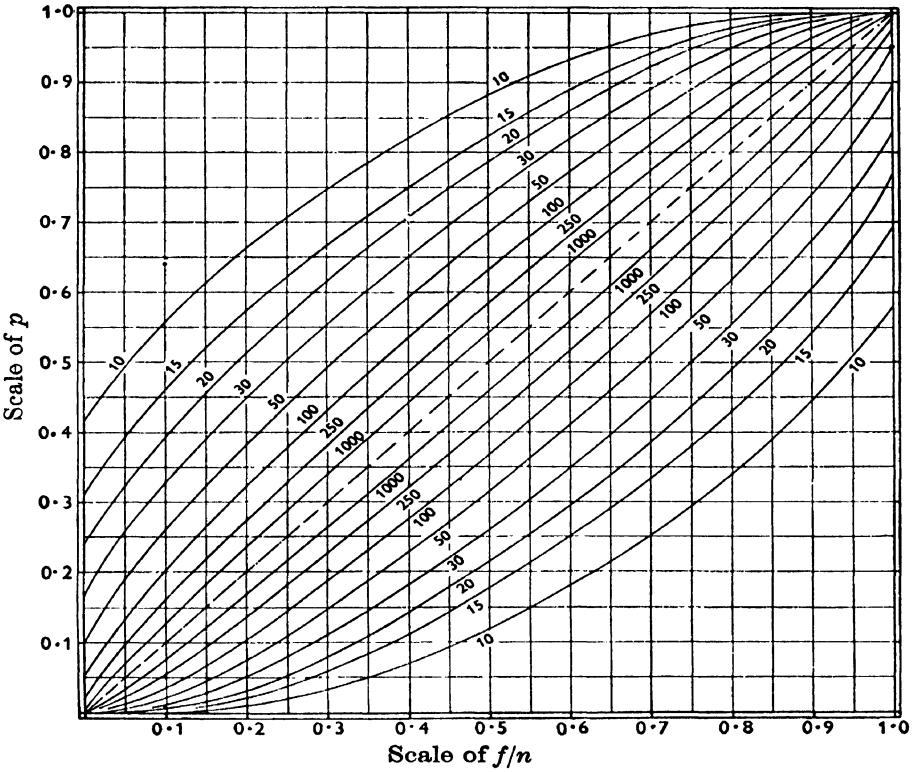
If n is kept constant but different values are taken for p, $0 < p < 1$, it will be possible to find f_1 and f_2 to satisfy the above inequalities for each value of p chosen. It will therefore be possible, for one given value of n and one given value of ϵ, to draw a diagram something like this:



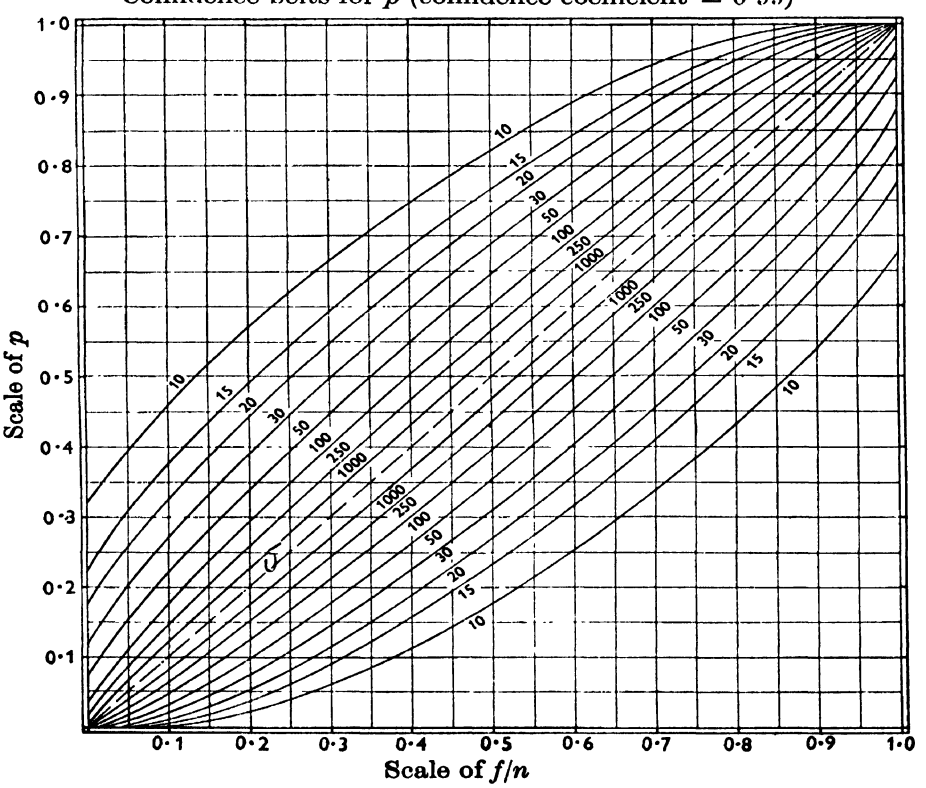
ϵ is usually arbitrarily chosen to be equal to 0·05 or 0·01 and it follows from the construction of the diagram that knowing p, the population probability, and n, the size of the sample, we may read off the two fractions f_1/n and f_2/n and be confident that only once in twenty times ($\epsilon = 0{\cdot}05$) or once in one hundred times ($\epsilon = 0{\cdot}01$) would we expect the probability as estimated from the sample to fall by chance outside these limits. The curves will be different for different values of n and ϵ but they will all follow the same kind of pattern.

Diagrams somewhat similar to these were first drawn by E. S. Pearson and C. J. Clopper. Their curves are reproduced

Confidence belts for p (confidence coefficient = 0·95)



Confidence belts for p (confidence coefficient = 0·99)



here by permission of the Editor of *Biometrika*. Although strictly they should have been drawn as a series of steps, as in the illustration above, the authors smoothed them by joining the outer points of the steps by a smooth curve.

Before we pass on to consider the estimation of an interval for an unknown population parameter from these curves, we may perhaps mention one use of the curves which is sometimes overlooked. Suppose it is desired to test the hypothesis that a sample, in which the observed proportion of units possessing a given character is f/n, had come from a population in which the proportion was p. This type of problem may arise in the testing of Mendelian hypotheses when it is possible to calculate *a priori* what the proportion in the population should be.

Example. From the mating of two individuals of genetical compositions **AA** and **Aa** the offspring must have the only possible genetical compositions **AA** and **Aa** by the Mendelian hypothesis. Among the 20 offspring from several matings of this type 14 were observed to be **AA** and 6 **Aa**. Is this consistent with the Mendelian hypothesis? By the Mendelian hypothesis (see next chapter),

$$P\{\mathbf{AA}\} = \tfrac{1}{2} = P\{\mathbf{Aa}\}.$$

The observed proportion of **AA** in a sample of 20 was $14/20 = 0{\cdot}7$. Reference to the confidence belt for $n = 20$, $\epsilon = 0{\cdot}05$ at the point $p = \frac{1}{2}$, shows that once in twenty times, owing to random errors, the sample proportion will lie outside the limits 0·25 and 0·75. In our case the observed proportion lies within these limits and we may say that there is nothing in the data to contradict the Mendelian hypothesis.

The procedure carried through in the example does not differ from that which we have previously discussed in other examples on the binomial theorem, except that in this case the limits are already calculated. If the sample proportion had happened to fall outside the sample limits as given by the chart, then the argument would be that only once in twenty times would this be expected to happen through chance, that twenty to one are rather long odds and that the original Mendelian hypothesis may not be tenable. At least the statistician would be justified in asking for a check of the original assumptions.

In using the 0·05 or 0·01 level for the acceptance or rejection

of hypotheses it must be remembered that on an average of once in twenty times (or once in one hundred times) the hypothesis tested will be correct but that the arbitrarily chosen limits will reject it as untenable. Obviously it will not be possible for the statistician to predict when such an occasion will arise and it will be necessary for him therefore to balance the sensitivity of the limits against the risk of making a false decision. In a lifetime of statistical work he will run the risk of averaging one in twenty wrong decisions if he always chooses the 0·05 level as his criterion.

This use of the confidence curves for the testing of hypotheses is not all-important, however, because the population probability is rarely known; in fact where Bayes' theorem is not directly applicable for purposes of estimation then no direct hypothesis can be tested. The use to which the confidence curves are most often put is to estimate an interval which will include the unknown population parameter 95 (or 99) times in 100. The confidence belts were constructed by considering all values of p between 0 and 1 and all possible values of x $(= f/n)$ for each value of p. Consider a function $\phi(p)$ which will have the property that the value of $\phi(p)$ at the point $p = \alpha$ will be the probability that $p = \alpha$. This function $\phi(p)$ we may call the *a priori* elementary probability law of p and we note that it is unknown.

The probability that any given point (x, p) will lie within the confidence belt, for a sample of size n, is

$$P\{(x,p) \mid n\} = \text{Probability that } (x,p) \text{ lies within the confidence belt for } n$$

$$= \sum_{\text{All } p} \phi(p) \sum_{x>f_1/n}^{x<f_2/n} P\{x \mid p\}.$$

Owing to the way in which the confidence belts were constructed we shall have

$$P\{(x,p) \mid n\} \geqslant \sum_{\text{All } p} \phi(p)\,(1-\epsilon) = 1-\epsilon.$$

Hence the probability that any pair of values (x, p) will lie within the confidence belt is greater than or equal to $1-\epsilon$, ϵ being the small positive fraction at choice. It follows that if we make the statement that the pair of values (x, p) will always lie within the confidence belt we shall be wrong in making this statement on a proportion ϵ of occasions.

Suppose now that we have an observed proportion $x = f/n$ and it is desired to estimate confidence limits for the unknown population probability, p. This may be done directly from the confidence belt. The abscissa is x and if an ordinate is drawn through x cutting the confidence belt at p_1 and p_2, then

$$P\{p_1 \leqslant p \leqslant p_2\} \geqslant 1 - \epsilon$$

and p_1 and p_2 are called the confidence limits for p. Because p is a population parameter it cannot vary, but the limits p_1 and p_2, dependent as they are on n and f, will vary from sample to sample; but however p_1 and p_2 may vary, in stating that the interval p_1 to p_2 will cover the true population value we shall be right in making this statement on a proportion $(1-\epsilon)$ of occasions. The value ϵ is at choice and is usually taken as 0·05 or 0·01. The statistician must balance the smaller interval for p if ϵ is large against the increased chance that the interval will not cover the true value.

We may note that because of the method of construction of the confidence belt the *a priori* distribution of p does not matter. Thus we have taken a step away from the restrictions of Bayes' theorem.

REFERENCES AND READING

Bayes' theorem is discussed by most writers on probability and the objections raised here are generally brought forward. For those who would like to read a defence of the theorem there is H. Jeffreys, *The Theory of Probability*, but it should be added that few statisticians accept Jeffreys' arguments. J. L. Coolidge, *An Introduction to Mathematical Probability* has a stimulating discussion of Bayes' theorem as has also J. V. Uspensky, *Introduction to Mathematical Probability*. R. A. Fisher, 'Uncertain Inference', *Proc. American Academy of Arts and Sciences*, LXXI, no. 4, gives interesting criticisms of Bayes' theorem and develops his own theory of fiducial inference published some years previously.

I have not touched on Fisher's theory in this chapter. We may note that he develops a theory which differs to a certain extent from that of Neyman and that both theories have their protagonists. Neyman first put forward (in English) his theory of confidence intervals in J. Neyman, 'On two different aspects of the representative method', *J. R. Statist. Soc.* 1934, and extended it later in 'Outline of a theory of statistical estimation based on the classical theory of probability', *Phil. Trans.* A, CCXXXVI, p. 333.

It is for the student to read both Fisher and Neyman and to make up his own mind which theory he prefers to adopt.

CHAPTER VIII

SIMPLE GENETICAL APPLICATIONS

It is perhaps surprising that the field of genetics has not made a more universal appeal to writers on probability. The hypotheses governing the simpler aspects of inheritance appear to be clear-cut and it is intellectually more satisfying to apply the fundamentals of probability to a subject which is of practical importance rather than to follow the orthodox procedure and discuss the hazards of the gaming tables. There are many text-books on genetics in which probability applications are set out, and this present chapter does not pretend to instruct the serious student of genetics; what is attempted is to give the student of probability a small idea of how the elementary theorems of probability may be of use.

The simple Mendelian laws of inheritance postulate the hypothesis that there are 'atoms' of heredity known as genes. These genes are associated in pairs and an offspring from the mating of two individuals receives one gene from the pair from each parent. Thus if we write **AA** for a pair of dominant genes, and **aa** for a pair of recessive genes, the genetical composition of the offspring of the mating of **AA** × **aa** can only be **Aa**. Such a genetical composition will be spoken of as a hybrid. From such simple assumptions it is possible to specify the probabilities of any type of genetical composition arising in the offspring of any given mating. If we write X_1 and X_2 for the genetical composition of the parents and Y for that of their offspring, we shall have, considering one pair of genes only, the following alternatives.

(i) $X_1 = \mathbf{AA}, X_2 = \mathbf{AA}$. $P\{Y = \mathbf{AA} \mid X_1 = \mathbf{AA}, X_2 = \mathbf{AA}\} = 1,$
$P\{Y = \mathbf{Aa} \mid X_1 = \mathbf{AA}, X_2 = \mathbf{AA}\} = 0,$
$P\{Y = \mathbf{aa} \mid X_1 = \mathbf{AA}, X_2 = \mathbf{AA}\} = 0.$

(ii) $X_1 = \mathbf{Aa}, X_2 = \mathbf{AA}$. $P\{Y = \mathbf{AA} \mid X_1 = \mathbf{Aa}, X_2 = \mathbf{AA}\} = \frac{1}{2},$
$P\{Y = \mathbf{Aa} \mid X_1 = \mathbf{Aa}, X_2 = \mathbf{AA}\} = \frac{1}{2},$
$P\{Y = \mathbf{aa} \mid X_1 = \mathbf{Aa}, X_2 = \mathbf{AA}\} = 0.$

Similarly for $X_1 = \mathbf{AA}$ and $X_2 = \mathbf{Aa}$.

(iii) $X_1 = \mathbf{aa}, X_2 = \mathbf{AA}$. $P\{Y = \mathbf{AA} \mid X_1 = \mathbf{aa}, X_2 = \mathbf{AA}\} = 0,$
$P\{Y = \mathbf{Aa} \mid X_1 = \mathbf{aa}, X_2 = \mathbf{AA}\} = 1,$
$P\{Y = \mathbf{aa} \mid X_1 = \mathbf{aa}, X_2 = \mathbf{AA}\} = 0.$

Similarly for $X_1 = \mathbf{AA}$ and $X_2 = \mathbf{aa}$.

(iv) $X_1 = \mathbf{Aa}, X_2 = \mathbf{Aa}$. $P\{Y = \mathbf{AA} \mid X_1 = \mathbf{Aa}, X_2 = \mathbf{Aa}\} = \frac{1}{4},$
$P\{Y = \mathbf{Aa} \mid X_1 = \mathbf{Aa}, X_2 = \mathbf{Aa}\} = \frac{1}{2},$
$P\{Y = \mathbf{aa} \mid X_1 = \mathbf{Aa}, X_2 = \mathbf{Aa}\} = \frac{1}{4}.$

(v) $X_1 = \mathbf{aa}, X_2 = \mathbf{Aa}$. $P\{Y = \mathbf{AA} \mid X_1 = \mathbf{aa}, X_2 = \mathbf{Aa}\} = 0,$
$P\{Y = \mathbf{Aa} \mid X_1 = \mathbf{aa}, X_2 = \mathbf{Aa}\} = \frac{1}{2},$
$P\{Y = \mathbf{aa} \mid X_1 = \mathbf{aa}, X_2 = \mathbf{Aa}\} = \frac{1}{2}.$

Similarly for $X_1 = \mathbf{Aa}$ and $X_2 = \mathbf{aa}$.

(vi) $X_1 = \mathbf{aa}, X_2 = \mathbf{aa}$. $P\{Y = \mathbf{AA} \mid X_1 = \mathbf{aa}, X_2 = \mathbf{aa}\} = 0,$
$P\{Y = \mathbf{Aa} \mid X_1 = \mathbf{aa}, X_2 = \mathbf{aa}\} = 0,$
$P\{Y = \mathbf{aa} \mid X_1 = \mathbf{aa}, X_2 = \mathbf{aa}\} = 1.$

These results follow directly from the application of elementary probability theorems. The probabilities thus obtained are sometimes spoken of as the Mendelian ratios.

The study of the inheritance of a particular pair of genes in a population is often rendered difficult by the fact that there is a selective factor in mating of which it is necessary to take account. Karl Pearson discussed this 'coefficient of assortative mating' for human populations and there is no doubt that it obtains for many animal populations also. In fact, it is difficult to think of any population in which it is reasonably certain that the mating is at random and is not affected by the genetical composition of the parents. The process of random mating is styled *Panmixia*, and we shall discuss a simplified form of this process. It may be questioned, since random mating is rarely met with in practice, whether it is worth while discussing. From the point of view of applying the theory of probability to genetical material it possibly is not, but from the point of view of understanding the application of probability theory to genetical theory the study of *Panmixia* will not be without value.

Assume that in a given population the proportions of males who are dominants (**AA**), hybrids (**Aa**) and recessives (**aa**) are p_1, q_1 and r_1, where $p_1+q_1+r_1 = 1$, and that the corresponding proportions for females are p_2, q_2 and r_2, where $p_2+q_2+r_2 = 1$. If it is further assumed that *Panmixia* operates, we may proceed to calculate the proportions of dominants, hybrids and recessives in the first, second and third filial generations. As before, write X_1 and X_2 to represent the genetical composition of the parents and Y for the genetical composition of their offspring. It follows then that the proportion of dominants in the first filial generation is given by

$$\begin{aligned} P\{Y = \mathbf{AA}\} = {} & P\{(X_1 = \mathbf{AA})(X_2 = \mathbf{AA})(Y = \mathbf{AA})\} \\ & + P\{(X_1 = \mathbf{AA})(X_2 = \mathbf{Aa})(Y = \mathbf{AA})\} \\ & + P\{(X_1 = \mathbf{AA})(X_2 = \mathbf{aa})(Y = \mathbf{AA})\} \\ & + P\{(X_1 = \mathbf{Aa})(X_2 = \mathbf{AA})(Y = \mathbf{AA})\} \\ & + P\{(X_1 = \mathbf{Aa})(X_2 = \mathbf{Aa})(Y = \mathbf{AA})\} \\ & + P\{(X_1 = \mathbf{Aa})(X_2 = \mathbf{aa})(Y = \mathbf{AA})\} \\ & + P\{(X_1 = \mathbf{aa})(X_2 = \mathbf{AA})(Y = \mathbf{AA})\} \\ & + P\{(X_1 = \mathbf{aa})(X_2 = \mathbf{Aa})(Y = \mathbf{AA})\} \\ & + P\{(X_1 = \mathbf{aa})(X_2 = \mathbf{aa})(Y = \mathbf{AA})\}. \end{aligned}$$

Each of these probabilities may be evaluated from first principles. For example

$$\begin{aligned} & P\{(X_1 = \mathbf{AA})(X_2 = \mathbf{AA})(Y = \mathbf{AA})\} \\ & \quad = P\{(X_1 = \mathbf{AA})\}\, P\{(X_2 = \mathbf{AA}) \mid (X_1 = \mathbf{AA})\} \\ & \qquad \times P\{(Y = \mathbf{AA}) \mid (X_1 = \mathbf{AA})(X_2 = \mathbf{AA})\}, \end{aligned}$$

or, since random mating was assumed,

$$\begin{aligned} & P\{(X_1 = \mathbf{AA})(X_2 = \mathbf{AA})(Y = \mathbf{AA})\} \\ & \quad = P\{(X_1 = \mathbf{AA})\}\, P\{(X_2 = \mathbf{AA})\} \\ & \qquad \times P\{(Y = \mathbf{AA}) \mid (X_1 = \mathbf{AA})(X_2 = \mathbf{AA})\}, \end{aligned}$$

whence, on substitution, we have

$$P\{(X_1 = \mathbf{AA})(X_2 = \mathbf{AA})(Y = \mathbf{AA})\} = p_1 p_2.$$

If P_1 be the total probability that Y is a dominant, then, by similar calculations for each individual term and substitution in the formula, it is found that

$$P\{Y = \mathbf{AA}\} = P_1 = (p_1 + \tfrac{1}{2}q_1)(p_2 + \tfrac{1}{2}q_2).$$

Similarly it may be shown that

$$P\{Y = \mathbf{Aa}\} = Q_1 = (p_1 + \tfrac{1}{2}q_1)(r_2 + \tfrac{1}{2}q_2) + (p_2 + \tfrac{1}{2}q_2)(r_1 + \tfrac{1}{2}q_1)$$

and $$P\{Y = \mathbf{aa}\} = R_1 = (r_1 + \tfrac{1}{2}q_1)(r_2 + \tfrac{1}{2}q_2).$$

It is easy to verify that $P_1 + Q_1 + R_1 = 1$. The proportions of dominants, hybrids and recessives in the first filial generation are therefore P_1, Q_1, R_1.

We now assume that random mating again occurs, and calculate the proportions of dominants, hybrids and recessives in the second filial generation Z. By a process identical with that for the first filial generation it may be shown that

$$\begin{aligned} P_2 = P\{Z = \mathbf{AA}\} &= (P_1 + \tfrac{1}{2}Q_1)^2, \\ Q_2 = P\{Z = \mathbf{Aa}\} &= 2(P_1 + \tfrac{1}{2}Q_1)(R_1 + \tfrac{1}{2}Q_1), \\ R_2 = P\{Z = \mathbf{aa}\} &= (R_1 + \tfrac{1}{2}Q_1)^2. \end{aligned}$$

Again if W is the third filial generation then

$$\begin{aligned} P_3 = P\{W = \mathbf{AA}\} &= (P_2 + \tfrac{1}{2}Q_2)^2 = (P_1 + \tfrac{1}{2}Q_1)^2, \\ Q_3 = P\{W = \mathbf{Aa}\} &= 2(P_2 + \tfrac{1}{2}Q_2)(R_2 + \tfrac{1}{2}Q_2) \\ &= 2(P_1 + \tfrac{1}{2}Q_1)(R_1 + \tfrac{1}{2}Q_1), \\ R_3 = P\{W = \mathbf{aa}\} &= (R_2 + \tfrac{1}{2}Q_2)^2 = (R_1 + \tfrac{1}{2}Q_1)^2, \end{aligned}$$

remembering that $P_1 + Q_1 + R_1 = 1$. It follows, then, that provided the mating is always at random and no extraneous factors intervene, the genetical compositions of the population do not change in proportion after the first filial generation. The population after the first filial generation may be regarded therefore as stable genetically.

Example. A breeder wishes to produce seeds of red flowering plants (**AA** or **Aa**). For this purpose he repeatedly performs a mass selection, consisting in the early removal from his fields of all plants with white flowers (**aa**) before they open. Thus he removes the possibility of any plants being fertilized by the pollen of pure recessives. Assume that the process of reproduction of plants left untouched in the field satisfies the definition of *Panmixia* and that in a particular year the percentage of plants removed because they would have flowered white was $r = 4/25$. Calculate the proportion of white flowers to be expected from the seeds of the plants left growing on the field.

Since repeated selection has been performed it may be assumed, if p, q and r are the proportions of dominants, hybrids and recessives respectively, that

$$p = (p+\tfrac{1}{2}q)^2, \quad q = 2(p+\tfrac{1}{2}q)(r+\tfrac{1}{2}q), \quad r = (r+\tfrac{1}{2}q)^2$$

which give, on solution of the equations,

$$p = (1-\sqrt{r})^2, \quad q = 2\sqrt{r}(1-\sqrt{r}), \quad r = r,$$

for the proportion of dominants, hybrids and recessives in a stable population. The recessives are artificially removed, i.e. $r = 0$, and the proportions accordingly become

$$p' = \frac{1-\sqrt{r}}{1+\sqrt{r}}, \quad q' = \frac{2\sqrt{r}}{1+\sqrt{r}}, \quad r' = 0.$$

If this population now mates according to *Panmixia* the proportion of recessives, r_1, in the next generation from a population so composed will be

$$r_1 = (r' + \tfrac{1}{2}q')^2 = \frac{r}{(1+\sqrt{r})^2}.$$

r was given equal to 4/25, and therefore the proportion of white flowers (recessives), given by substituting in the expression for r_1, is $r_1 = 4/49$, that is, the proportion of white flowers has been almost halved by a single selective process.

The procedure set out in this example of artificially destroying a certain proportion of the population raises some interesting queries as to what will happen if the selection is carried out a number of times, and what the number of repetitions will need to be if the proportion of recessives is to fall below a given number. The whole problem of random mating and artificial selection can be represented geometrically.

We have seen that in a genetically stable population

$$p = (1-\sqrt{r})^2, \quad q = 2\sqrt{r}(1-\sqrt{r}), \quad r = r,$$

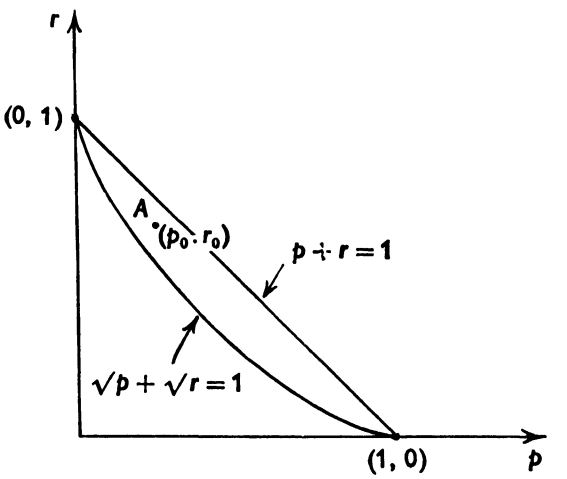
from which

$$\sqrt{p}+\sqrt{r} = 1,$$

which is a parabola. Also since $p+q+r = 1, p+r \leqslant 1$. The composition of a genetically stable population is therefore represented by that part of the parabola lying between the points $(1, 0)$ and $(0, 1)$.

Consider a point A with co-ordinates (p_0, r_0). If p_0 and r_0 represent the proportion of dominants and recessives in a population, then for *Panmixia*, as already proved, the proportions p_1 and r_1 of dominants and recessives in the first filial generation will be

$$p_1 = (p_0 + \tfrac{1}{2}q_0)^2 = \tfrac{1}{4}(1 + (p_0 - r_0))^2,$$
$$r_1 = (r_0 + \tfrac{1}{2}q_0)^2 = \tfrac{1}{4}(1 - (p_0 - r_0))^2.$$



(Not to scale)

If p_0 and r_0 are constant, then the point with co-ordinates (p_1, r_1) will be given by the intersection with the parabola of the perpendicular to the line

$$p + r = 1$$

through A. If (p_0, r_0) is a point on the parabola then (p_1, r_1) is the same point. Hence, if the mating in a population is supposed to be according to *Panmixia* the composition of the next generation may easily be found by geometrical drawing.

We may now suppose that we have a population the composition of which is genetically stable, and the co-ordinates of which on the parabola are (p, r). If all the recessives in this population are destroyed the proportion of dominants will be $p/(1-r)$ and the composition of the population will be represented by a point A_0' with co-ordinates $(p/(1-r), 0)$. It will be noted that the point A_0' is also the point of intersection with the abscissa of the line joining the points $(0, 1)$ and $A_0, (p, r)$.

If the population represented by A_0' is allowed to mate randomly, then, following the previous analysis, the composition

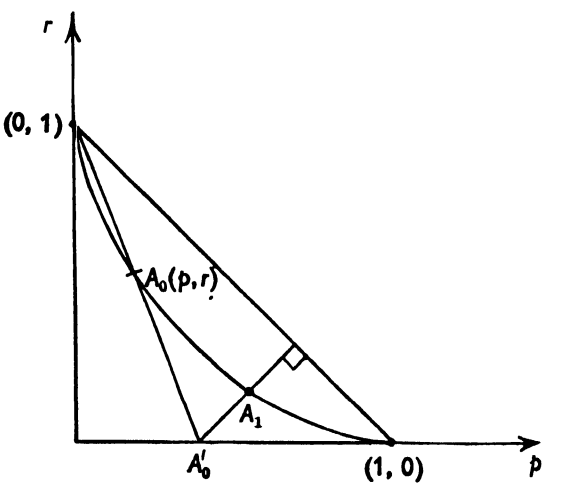
of the next generation will be given by the co-ordinates of the point of intersection with the parabola of a line at right angles to

$$p+r=1$$

and passing through A_0'. The co-ordinates of this point A_1 are

$$\left[\left(\frac{1}{(1+\sqrt{r})^2}\right), \quad \left(\frac{r}{(1+\sqrt{r})^2}\right)\right].$$

The effect of applying selection to the population represented by A_1, and then allowing it to mate randomly, can be seen by



(Not to scale)

following the same process; A_1 is joined to (0, 1) and the point of intersection of this line with the abscissa is A_1'. The line through A_1' at right angles to the line joining (1, 0) and (0, 1) gives the composition of the next generation in its point of intersection with the parabola. Suppose this process is carried out n times. It is easily shown that the co-ordinates of A_n and A_n' are

$$A_n: \quad \left[\left(\frac{1+(n-1)\sqrt{r}}{1+n\sqrt{r}}\right)^2, \left(\frac{\sqrt{r}}{1+n\sqrt{r}}\right)^2\right],$$

$$A_n': \quad \left[\frac{1+(n-1)\sqrt{r}}{1+(n+1)\sqrt{r}}, 0\right].$$

The proportion of hybrids may always be found from the relation

$$p+q+r=1.$$

If the selection process is carried out n times, then it is seen from A_n that the proportion of recessives in the nth filial generation will be

$$\frac{r}{(1+n\sqrt{r})^2}.$$

Example. If the proportion of recessives was 4/25 in the original population, how many times will the selection process need to be carried out in order that the proportion of recessives in the last generation should be less than 0·01?

Here $r = 4/25$, and it is required to determine n such that

$$\frac{4/25}{(1+n.2/5)^2} < 0{\cdot}01.$$

n must at least equal 8, that is at least 8 selections must be made.

Correlation between the Composition of Siblings

It will be assumed that there is a population which is genetically stable and in which the mating is random. The proportions of dominants, hybrids and recessives are p, q and r for both male and female. By an application of the fundamental probability laws, as at the beginning of this chapter, the probability of a pair of offspring having given genetical compositions may be calculated, and hence the correlation between the genetical compositions of two offspring.

If the population is genetically stable, then, as before,

$$p = (1-\sqrt{r})^2, \quad q = 2\sqrt{r}(1-\sqrt{r}), \quad r = r.$$

If X_1 and X_2 are the two parents and Y_1 and Y_2 the two offspring, then

$$\begin{aligned}
&P\{(Y_1 = \mathbf{AA})(Y_2 = \mathbf{AA})\} \\
&\quad = P\{(X_1 = \mathbf{AA})(X_2 = \mathbf{AA})(Y_1 = \mathbf{AA})(Y_2 = \mathbf{AA})\} \\
&\qquad + P\{(X_1 = \mathbf{AA})(X_2 = \mathbf{Aa})(Y_1 = \mathbf{AA})(Y_2 = \mathbf{AA})\} \\
&\qquad + P\{(X_1 = \mathbf{Aa})(X_2 = \mathbf{AA})(Y_1 = \mathbf{AA})(Y_2 = \mathbf{AA})\} \\
&\qquad + P\{(X_1 = \mathbf{Aa})(X_2 = \mathbf{Aa})(Y_1 = \mathbf{AA})(Y_2 = \mathbf{AA})\}.
\end{aligned}$$

The other possible matings can be neglected because they could

not produce offspring the genetical composition of which was **AA**. After expansion of the probabilities, for example,

$$P\{(X_1 = \mathbf{Aa})(X_2 = \mathbf{AA})(Y_1 = \mathbf{AA})(Y_2 = \mathbf{AA})\} = P\{(X_1 = \mathbf{Aa})\} \times P\{(X_2 = \mathbf{AA})\} P\{(Y_1 = \mathbf{AA}) \mid (X_1 = \mathbf{Aa})(X_2 = \mathbf{AA})\} \times P\{(Y_2 = \mathbf{AA}) \mid (X_1 = \mathbf{Aa})(X_2 = \mathbf{AA})\},$$

it is found that

$$P\{(Y_1 = \mathbf{AA})(Y_2 = \mathbf{AA})\} = \tfrac{1}{4}(1-\sqrt{r})^2(2-\sqrt{r})^2.$$

Similarly

$$P\{(Y_1 = \mathbf{AA})(Y_2 = \mathbf{Aa})\} = \frac{\sqrt{r}}{2}(1-\sqrt{r})^2(2-\sqrt{r}),$$

$$P\{(Y_1 = \mathbf{AA})(Y_2 = \mathbf{aa})\} = \frac{r}{4}(1-\sqrt{r})^2,$$

$$P\{(Y_1 = \mathbf{Aa})(Y_2 = \mathbf{Aa})\} = \sqrt{r}(1-\sqrt{r})(1+\sqrt{r}-r),$$

$$P\{(Y_1 = \mathbf{Aa})(Y_2 = \mathbf{aa})\} = \frac{r}{2}(1-\sqrt{r})(1+\sqrt{r}),$$

$$P\{(Y_1 = \mathbf{aa})(Y_2 = \mathbf{aa})\} = \frac{r}{4}(1+\sqrt{r})^2.$$

Accordingly a correlation table (given on p. 91) may be drawn up, each cell of which will be the joint probability of Y_1 and Y_2 having two given genetical compositions.

If **AA**, **Aa** and **aa** are arbitrarily assigned values, 1, 0 and -1, the table of probabilities can be treated as a correlation table and the correlation coefficient between the genetical compositions of Y_1 and Y_2 worked out. The total 'frequency' in the table is unity. Hence

$$\rho = \frac{\Sigma Y_1 Y_2 - \Sigma Y_1 \Sigma Y_2}{\sigma_{Y_1}\sigma_{Y_2}}$$

and easy algebra will give that ρ, the correlation coefficient between the genetical composition of the offspring, is 0·5. The correlation between the genetical composition of the offspring will thus appear to be independent of the original proportions in the population.

Exercise. Assume that there is a population which is genetically stable and in which the mating is random. Let the proportions of dominants, hybrids and recessives be p, q and r for both male and female. Show that the correlation between parent and offspring is 0·5.

Y_2 \ Y_1	aa	Aa	AA	Totals	Scale
AA	$\frac{r}{4}(1-\sqrt{r})^2$	$\frac{\sqrt{r}}{2}(1-\sqrt{r})^2(2-\sqrt{r})$	$\frac{1}{4}(1-\sqrt{r})^2(2-\sqrt{r})^2$	$(1-\sqrt{r})^2$	1
.....	.....	.....	.....	.....	.....
Aa	$\frac{r}{2}(1-\sqrt{r})(1+\sqrt{r})$	$\sqrt{r}(1-\sqrt{r})(1+\sqrt{r}-r)$	$\frac{\sqrt{r}}{2}(1-\sqrt{r})^2(2-\sqrt{r})$	$2\sqrt{r}(1-\sqrt{r})$	0
.....	.....	.....	.....	.....	.....
aa	$\frac{r}{4}(1+\sqrt{r})^2$	$\frac{r}{2}(1-\sqrt{r})(1+\sqrt{r})$	$\frac{r}{4}(1-\sqrt{r})^2$	r	-1
Totals	r	$2\sqrt{r}(1-\sqrt{r})$	$(1-\sqrt{r})^2$	1	
Scale	-1	0	1		

Elimination of Races by Selective Breeding

In *Panmixia* we have discussed the problem of the elimination of recessives in a population by destroying the recessives of each successive generation. We shall now study another aspect of the problem whereby genes carrying undesirable characteristics are eliminated by purposively mating individuals with others having a different genetical composition. This type of race improvement is an everyday practical problem, particularly in cattle breeding, where, by careful choice of bulls to serve the herd, a farmer may convert a herd of parents with an indifferent milk yield into a herd of descendants with a good milk yield. Let

$$\mathbf{r}_1\mathbf{r}_1, \quad \mathbf{r}_2\mathbf{r}_2, \quad \mathbf{r}_3\mathbf{r}_3, \quad \ldots, \quad \mathbf{r}_n\mathbf{r}_n$$

denote n pairs of genes which it is desired to eliminate from a race, r_e. No assumption is made whether these genes are dominant, hybrid or recessive. Further, let

$$\mathbf{R}_1\mathbf{R}_1, \quad \mathbf{R}_2\mathbf{R}_2, \quad \mathbf{R}_3\mathbf{R}_3, \quad \ldots, \quad \mathbf{R}_n\mathbf{R}_n$$

denote n pairs of genes belonging to an individual of race R_i. It is desired to introduce these n pairs of genes into r_e. Individuals of R_i and r_e are mated. The genetical composition of the first generation must be

$$F_1 = R_i \times r_e = \mathbf{R}_1\mathbf{r}_1, \mathbf{R}_2\mathbf{r}_2, \ldots, \mathbf{R}_n\mathbf{r}_n,$$

since the offspring will receive one gene of each type from each parent. Now mate the F_1 generation with individuals of the R parent race. Individuals of the F_2 (second) generation will receive one of a pair of genes from R_i and one from F_1. The gene from F_1 may be **R** or **r**. Thus

$$F_2 = R_i \times F_1 = \mathbf{R}_1X_1, \mathbf{R}_2X_2, \ldots, \mathbf{R}_nX_n,$$

where X may be **R** or **r**. Suppose that this backcrossing is carried out $(s+1)$ times so that

$$F_{s+1} = R_i \times F_s.$$

Let $P_{n.k}^{s+1}$ denote the probability that an individual of the $(s+1)$st generation will possess exactly k genes of the n genes of type **r** that it is desired to eliminate. Further, let $p_i(s+1)$ be the probability that an individual of the F_{s+1} generation will possess

a gene of type $\mathbf{r}_t$. From the law of inheritance of genes it follows that

$$p_t(s+1) = \frac{1}{2}p_t(s) = \frac{1}{2^2}p_t(s-1) = \ldots = \frac{1}{2^s}p_t(1).$$

$p_t(1)$ is by definition the probability that an individual of the first generation will possess a gene of type $\mathbf{r}_t$. We have seen that it is certain that this will be so and therefore

$$p_t(s+1) = \frac{1}{2^s}.$$

This result is independent of t and will hold for any pair of genes.

The probability that an individual of the F_{s+1} generation will possess exactly k genes of type $\mathbf{r}$ will be, from the binomial theorem,

$$P_{n.k}^{s+1} = \frac{n!}{k!(n-k)!}\left(\frac{1}{2^s}\right)^k\left(1-\frac{1}{2^s}\right)^{n-k}.$$

It is desired, if possible, to eliminate the genes of r_e completely and we are therefore concerned with the case of k equal zero, that is, with the probability that an individual of the F_{s+1} generation will possess no genes of the type to be eliminated. This probability will be

$$P_{n.0}^{s+1} = \left(1-\frac{1}{2^s}\right)^n.$$

As s increases without limit

$$P_{n.0}^{s+1} \to 1$$

irrespective of the number of genes n.

Example. If $n = 12$, what is the smallest value of s in order that the most probable composition of an individual of F_{s+1} will be the composition of an individual of R_i?

It has been proved for the binomial that if the greatest term of the expansion $(q+p)^n$ is at the integer k_0, then

$$(n+1)p - 1 \leqslant k_0 < (n+1)p.$$

In this present example it is required that k_0 should be zero and hence it will be necessary that

$$(n+1)p < 1,$$

i.e. that

$$13\left(\frac{1}{2^s}\right) < 1,$$

from which it follows that s must equal 4.

Example. If $n = 6$, show how the distribution of $P^{s+1}_{n.k}$ changes as s increases. From theory

$$P^{s+1}_{6.k} = \frac{6!}{k!(6-k)!}\left(\frac{1}{2^s}\right)^k\left(1-\frac{1}{2^s}\right)^{6-k}$$

and the problem reduces to that of calculating a number of binomial probabilities.

Table of $P^{s+1}_{n.k}$ *for* $n = 6$

s \ k	0	1	2	3	4	5	6
1	0·015,625	0·093,750	0·234,375	0·312,500	0·234,375	0·093,750	0·015,625
2	0·177,978	0·355,957	0·296,631	0·131,836	0·032,959	0·004,395	0·000,244
3	0·448,795	0·384,681	0·137,386	0·026,169	0·002,804	0·000,160	0·000,004
4	0·678,934	0·271,574	0·045,262	0·004,023	0·000,201	0·000,005	—
5	0·826,553	0·159,978	0·012,901	0·000,555	0·000,013	—	—
6	0·909,836	0·086,651	0·003,439	0·000,073	0·000,001	—	—

Example. What is the smallest number of backcrossings necessary in order that the probability that an individual of the $(s+1)$st generation possessing no genes of type **r** shall be at least equal to 0·99? Assume $n = 12$.

It is required that $$P^{s+1}_{12.0} \geqslant 0{\cdot}99,$$

i.e. that $$\left(1-\frac{1}{2^s}\right)^{12} \geqslant 0{\cdot}99.$$

In order to satisfy this inequality s must be at least equal to 11.

Bayes' Theorem and Mendelian Hypotheses

In a previous chapter the conditions under which the application of Bayes' theorem is thought to be legitimate have been set out, and it was stated that these conditions were nearly always fulfilled for Mendelian hypotheses. It is proposed now to illustrate by means of examples the application of the theorem in this case.

Example. From the mating of two dominant-looking hybrids, **Aa** × **Aa**, a dominant-looking offspring is obtained of composition **A***x*, x being unknown. This individual is mated with another hybrid and as a result of this mating n individuals are obtained, all of which look like dominants. What is the *a posteriori* probability, that x = **A**?

From the mating of the parent hybrids we may obtain an individual the genetical composition of which is **aa**, **Aa**, or **AA**. The first alternative is ruled out because we are told that the individual is dominant-looking. Let h_1 be the hypothesis that $x = \mathbf{A}$ and h_2 the hypothesis that $x = \mathbf{a}$. It is required to find the probability that the hypothesis h_1 is true. Consider now the mating of Ax with another hybrid. If $x = \mathbf{A}$ we have, for a single offspring, Y,

$$P\{Y = \mathbf{AA} \text{ or } \mathbf{Aa}\} = P\{Y = \mathbf{AA} \text{ or } \mathbf{Aa} \mid X_1 = \mathbf{AA}, X_2 = \mathbf{Aa}\} = 1.$$

If $x = \mathbf{a}$ then

$$P\{Y = \mathbf{AA} \text{ or } \mathbf{Aa}\} = P\{Y = \mathbf{AA} \text{ or } \mathbf{Aa} \mid X_1 = \mathbf{Aa}, X_2 = \mathbf{Aa}\} = \tfrac{3}{4}.$$

Hence the probability of obtaining n dominant-looking offspring under hypotheses h_1 and h_2 will be

$$P\{n(\mathbf{AA} \text{ or } \mathbf{Aa}) \mid h_1\} = 1$$
$$P\{n(\mathbf{AA} \text{ or } \mathbf{Aa}) \mid h_2\} = (\tfrac{3}{4})^n.$$

The *a priori* probabilities of the hypotheses h_1 and h_2 will be

$$P\{h_1\} = \tfrac{1}{3}, \quad P\{h_2\} = \tfrac{2}{3},$$

for the possible offspring from the mating of two hybrids are **AA**, **Aa**, **aA** and **aa** and the last alternative is ruled out because the individual is dominant-looking. All the probabilities necessary for the calculation of probabilities by Bayes' theorem have been enumerated. Accordingly

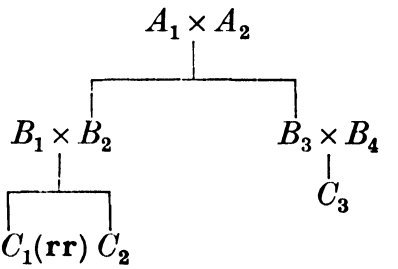
$$P\{h_1 \mid n(\mathbf{AA} \text{ or } \mathbf{Aa})\} = \frac{P\{h_1\} P\{n(\mathbf{AA} \text{ or } \mathbf{Aa}) \mid h_1\}}{P\{h_1\} P\{n(\mathbf{AA} \text{ or } \mathbf{Aa}) \mid h_1\} + P\{h_2\} P\{n(\mathbf{AA} \text{ or } \mathbf{Aa}) \mid h_2\}}$$
$$= \frac{\frac{1}{3} \cdot 1}{\frac{1}{3} \cdot 1 + \frac{2}{3}(\frac{3}{4})^n} = \frac{1}{1 + 2(\frac{3}{4})^n}$$

and
$$P\{h_2 \mid n(\mathbf{AA} \text{ or } \mathbf{Aa})\} = \frac{2}{2 + (\frac{4}{3})^n}.$$

Let $n = 4$,
$$P\{h_1 \mid 4(\mathbf{AA} \text{ or } \mathbf{Aa})\} = 0{\cdot}61,$$
$$P\{h_2 \mid 4(\mathbf{AA} \text{ or } \mathbf{Aa})\} = 0{\cdot}39,$$

and we should not be certain of either hypothesis unless more offspring from a further mating of x with a hybrid were obtained.

Example. A certain pair of genes has the property that pure recessive individuals with the composition **rr** possess a certain defect X, while the dominants **RR** and the hybrids **Rr** are normal. Consider the following pedigree in which one individual, C_1, possesses the inherited defect, X, all others appearing normal.



C_2 and C_3 intend to marry and it is required to calculate the probability that a single one of their offspring would possess the defect X. Assume that A_2 and B_4 were selected at random from a population of apparently normal individuals and that the probability that any one of these has a hidden gene **r** is $p(\mathbf{r}) = 0{\cdot}001$. (B.Sc. London, 1937.)

It is stated that C_1 is **rr** and that both B_1 and B_2 are normal individuals. It follows that both B_1 and B_2 must have the composition **Rr**. Further, since A_1 and A_2 are normal, then either A_1 or A_2 must have the composition **Rr**. Let A_1 be the parent who passed the **r** gene to B_2.

The offspring from the mating of B_1 and B_2 can have the compositions **RR**, **Rr** or **rr**. C_2 is normal and cannot be **rr** and we have therefore

$$P\{C_2 = \mathbf{RR}\} = \tfrac{1}{3}, \quad P\{C_2 = \mathbf{Rr}\} = \tfrac{2}{3}.$$

This completes the left-hand branch of the pedigree.

Let us now consider the right-hand branch.

$$\begin{aligned} P\{B_3 = \mathbf{RR}\} &= P\{(A_1 = \mathbf{Rr})\ (A_2 = \mathbf{RR})\ (B_3 = \mathbf{RR})\} \\ &\quad + P\{(A_1 = \mathbf{Rr})\ (A_2 = \mathbf{Rr})\ (B_3 = \mathbf{RR})\} \\ &= P\{A_1 = \mathbf{Rr}\}\ P\{A_2 = \mathbf{RR}\} \\ &\quad \times P\{(B_3 = \mathbf{RR}) \mid (A_1 = \mathbf{Rr})\ (A_2 = \mathbf{RR})\} \\ &\quad + P\{A_1 = \mathbf{Rr}\}\ P\{A_2 = \mathbf{Rr}\} \\ &\quad \times P\{(B_3 = \mathbf{RR}) \mid (A_1 = \mathbf{Rr})\ (A_2 = \mathbf{Rr})\}. \end{aligned}$$

It is known that A_1 must be **Rr** and therefore

$$P\{A_1 = \mathbf{Rr}\} = 1,$$

but what is the probability that A_2 is a dominant or a hybrid? We are told that the probability of an individual possessing a hidden gene of type **r** is 0·001. It follows therefore that if an individual such as A_2 or B_4 is chosen at random from the population,

$$P\{A_2 = \mathbf{Rr}\} = 0{\cdot}001 \quad \text{and} \quad P\{A_2 = \mathbf{RR}\} = 0{\cdot}999.$$

Hence

$$P\{B_3 = \mathbf{RR}\} = 0{\cdot}999 \times \tfrac{1}{2} + 0{\cdot}001 \times \tfrac{1}{4} = 0{\cdot}49975.$$

Similarly $P\{B_3 = \mathbf{Rr}\} = 0{\cdot}50000$

and therefore $P\{B_3 = \mathbf{rr}\} = 0{\cdot}00025.$

B_3 is, however, reported as normal and accordingly the *a posteriori* probabilities will be

$$P\{B_3 = \mathbf{RR}\} = 0{\cdot}499875,$$
$$P\{B_3 = \mathbf{Rr}\} = 0{\cdot}500125.$$

The *a priori* probabilities that C_3 is a dominant or a hybrid may be calculated in the same way as for B_3.

$$\begin{aligned} P\{C_3 = \mathbf{RR}\} = {} & P\{(B_3 = \mathbf{RR})(B_4 = \mathbf{RR})(C_3 = \mathbf{RR})\} \\ & + P\{(B_3 = \mathbf{RR})(B_4 = \mathbf{Rr})(C_3 = \mathbf{RR})\} \\ & + P\{(B_3 = \mathbf{Rr})(B_4 = \mathbf{RR})(C_3 = \mathbf{RR})\} \\ & + P\{(B_3 = \mathbf{Rr})(B_4 = \mathbf{Rr})(C_3 = \mathbf{RR})\}. \end{aligned}$$

Expanding and substituting numerical values we have

$$P\{C_3 = \mathbf{RR}\} = 0{\cdot}74956$$

and by a similar process

$$P\{C_3 = \mathbf{Rr}\} = 0{\cdot}25031, \quad P\{C_3 = \mathbf{rr}\} = 0{\cdot}00013,$$

whence the probabilities *a posteriori* for C_3 can be deduced to be

$$P\{C_3 = \mathbf{RR}\} = 0{\cdot}74966, \quad P\{C_3 = \mathbf{Rr}\} = 0{\cdot}25034.$$

Let Y be the offspring if C_2 and C_3 marry. It is required to find the probability that Y will possess the defect, i.e. that $Y = \mathbf{rr}$.

$$P\{Y = \mathbf{rr}\} = P\{(C_2 = \mathbf{Rr})(C_3 = \mathbf{Rr})(Y = \mathbf{rr})\} = 0{\cdot}042.$$

Exercise. Consider the genes **R** and **r** as given in the preceding example. Let F_1 and F_2 be two consecutive generations of a population. Denote by q_1 the probability that an individual, chosen at random from the apparently normal individuals of F_1, will be a hybrid (**Rr**). Assume that the matings of apparently normal individuals are at random.

(i) What is the probability, q_2, that an apparently normal individual, Y, of F_2 whose parents are externally normal, will be a hybrid?

(ii) What is the probability, $P\{(Z_1 = \mathbf{Rr})(Z_2 = \mathbf{Rr}) \mid n\}$, that the apparently normal individuals Z_1 and Z_2 will have the compositions $Z_1 = \mathbf{Rr}$ and $Z_2 = \mathbf{Rr}$ when it is known that their offspring, n in number, are all externally normal?

(iii) What is the probability that an apparently normal individual, W, of F_2 will have the composition **Rr**, given that his parents and $(n-1)$ brothers are known to be externally normal?

(iv) If the probability in (iii) is $q_2(n)$, find the limit of $q_2(n)$ as n tends to infinity. Put $q_1 = 0{\cdot}001$ and $n = 1, 2, 3$, in turn and see how the knowledge that the parents and siblings of W are externally normal influences the probability that W has a hidden gene, **r**. (B.Sc. London, 1937.)

REFERENCES AND READING

Any elementary text-book on genetics will give the reader more genetical terminology than has been assumed as known here. Applications of probability to genetical problems are spread widely through genetical literature. We may mention two books by K. Mather, *Statistical Analysis in Biology* and *The Measurement of Linkage in Heredity*, in which the student will find a number of biological problems treated statistically. Chapter IX of R. A. Fisher, *Statistical Methods for Research Workers*, may also be read with profit.

The main ideas of the present chapter were obtained from lectures by Karl Pearson and J. Neyman. The interpretation of these ideas is the writer's own.

CHAPTER IX

MULTINOMIAL THEOREM AND SIMPLE COMBINATORIAL ANALYSIS

Thus far in probability we have been concerned chiefly with fundamental probability sets the elements of which possess two alternative characteristics only; an event may happen or not happen, a ball may be black or white, and so on. No discussion of discrete probabilities would, however, be complete without some investigation of the case where an individual of the fundamental probability set may possess one of *several* different characteristics. The binomial theorem gives a method for the calculation of probabilities when there are two alternatives; we now turn to the multinomial theorem which applies to cases in which more than two alternatives need to be considered.

In stating that an element of the fundamental probability set may possess any one of k mutually exclusive properties, these properties being the only possible, we are formulating a general proposition a particular case of which might be that an event may happen in k different ways and so on. If the fundamental probability set is composed of N elements, N_1 of which possess the property A_1, N_2 of which possess the property $A_2, \ldots, N_k$ of which possess the property A_k, where the N elements may be actual recorded happenings or a mathematical model, then p_i, the probability that an element of the fundamental probability set possesses the property A_i, will be

$$p_i = \frac{N_i}{N} \quad (i = 1, 2, \ldots, k)$$

by definition.

Suppose now that n independent trials are made. The probability that a single trial will result in an element being found to possess a given characteristic is defined, and we proceed, as in the case of the binomial, to ask, what is the probability that as a result of these n trials r_1 elements will be found to possess the character A_1, r_2 the character $A_2, \ldots, r_k$ the character A_k?

MULTINOMIAL THEOREM. An event may happen in k mutually exclusive ways which are also the only possible. The probability

that in n trials the event will happen r_1 times in the first way, r_2 in the second, ..., r_k times in the kth way is

$$\frac{n!}{r_1!\,r_2!\ldots r_k!}\,p_1^{r_1}p_2^{r_2}\ldots p_k^{r_k},$$

where p_i is the probability that the event will happen in the ith way in a single trial and $\sum_{i=1}^{k} p_i = 1$.

Since the k ways are mutually exclusive and are also the only possible, the event must happen in one of the given ways. If it were required to find the probability that it would happen in the first way for the first r_1 trials, in the second way for the next r_2 trials, and so on, the probability would be simply

$$p_1^{r_1}p_2^{r_2}\ldots p_k^{r_k}.$$

No order, however, is specified and it is necessary therefore to enumerate the number of ways in which $r_1, r_2, \ldots, r_k$ trials can be arranged subject to the restriction that

$$r_1+r_2+\ldots+r_k = n.$$

The expression $(p_1+p_2+\ldots+p_k)^n$ is the product of

$$(p_1+p_2+\ldots+p_k)$$

by itself $n-1$ times, i.e.

$$(p_1+p_2+\ldots+p_k)^n$$
$$= (p_1+p_2+\ldots+p_k)(p_1+p_2+\ldots+p_k)\ldots(p_1+p_2+\ldots+p_k).$$

Every term in the expansion of the left-hand side is formed by taking one symbol out of each of the n brackets of the right-hand side. Hence the number of ways in which any term $p_1^{r_1}p_2^{r_2}\ldots p_k^{r_k}$ will appear in the final expansion will be the number of ways of arranging n symbols when r_1 must be p_1, r_2 must be $p_2, \ldots, r_k$ must be p_k.

This is the same requirement as for the arrangement of probabilities. It follows that the probability of obtaining r_1 trials of the first kind, r_2 of the second and so on will be given by the complete term $p_1^{r_1}p_2^{r_2}\ldots p_k^{r_k}$ in the expansion of

$$(p_1+p_2+\ldots+p_k)^n$$

and that this probability is

$$\frac{n!}{r_1!\,r_2!\,\dots\,r_k!}\,p_1^{r_1}p_2^{r_2}\dots p_k^{r_k}.$$

The expression $(p_1+p_2+\dots+p_k)^n$ may be spoken of as the generating function of the probabilities.

When only two alternatives are possible, such as when an event may or may not happen, then

$$r_1 = k, \quad r_2 = n-k, \quad p_1 = p, \quad p_2 = 1-p = q,$$

$$p_3 = p_4 = \dots = p_k = 0$$

and the probability that an event will happen exactly k times in n trials is

$$\frac{n!}{k!(n-k)!}\,p^k q^{n-k},$$

as found in Chapter III.

Example. A bag contains 5 white, 7 green, 12 red and 14 black balls. The balls are indistinguishable from each other except by colour. A ball is drawn and replaced, after its colour had been noted, on ten occasions. If any ball is as likely to be drawn as any other, what is the probability that of the ten balls seen 3 will be white, 3 green, 2 red and 2 black?

Answer: $$\frac{10!}{3!\,3!\,2!\,2!}\left(\frac{5}{38}\right)^3\left(\frac{7}{38}\right)^3\left(\frac{12}{38}\right)^2\left(\frac{14}{38}\right)^2.$$

The binomial and multinomial theorems are, if equal probability of all elements in the fundamental probability set is assumed or established, simple propositions which fit into a general mathematical scheme of arrangements generally known as combinatorial analysis. The ideas and theorems used in combinatorial analysis, as far as probability is concerned, are not new—many of them were known to Laplace—but they do not appear to be as well known as they should. We may discuss here certain simple aspects of this analysis both in relation to the theory of probability and in its application to statistical method, but it will be necessary first of all to define certain quantities and to state some of their properties.

Differences of Zero

If $x_1, x_2, \ldots, x_n$ are a series of numbers, or the values of a given function at successive entries of the tabled argument, then it is conventional to write

$$
\begin{array}{lllllll}
x_1 & & & & & & \\
 & \Delta x_1 & & & & & \\
x_2 & & \Delta^2 x_1 & & & & \\
 & \Delta x_2 & & \Delta^3 x_1 & & & \\
x_3 & & \Delta^2 x_2 & & \Delta^4 x_1 & & \\
 & \Delta x_3 & & \Delta^3 x_2 & & \ddots & \\
x_4 & & \Delta^2 x_3 & & \Delta^4 x_2 & & \ddots \\
 & \Delta x_4 & & \Delta^3 x_3 & & \ddots & \\
\vdots & & \ddots & & \ddots & & \ddots \\
 & & & \ddots & & \ddots & \\
 & & & & \ddots & &
\end{array}
$$

where $\Delta x_1 = x_2 - x_1$, $\Delta^2 x_1 = \Delta x_2 - \Delta x_1$, $\Delta^3 x_1 = \Delta^2 x_2 - \Delta^2 x_1$, and so on. The differences associated with x_1 are often spoken of as the *leading* differences.

If $x_1 = x^s$, $x_2 = (x+1)^s, \ldots, x_n = (x+\overline{n-1})^s$, then we have

$$
\begin{array}{lllllll}
x^s & & & & & & \\
 & \Delta x^s & & & & & \\
(x+1)^s & & \Delta^2 x^s & & & & \\
 & \Delta(x+1)^s & & \Delta^3 x^s & & & \\
(x+2)^s & & \Delta^2(x+1)^s & & \Delta^4 x^s & & \\
 & \Delta(x+2)^s & & \Delta^3(x+1)^s & & \ddots & \\
(x+3)^s & & \Delta^2(x+2)^s & & \Delta^4(x+1)^s & & \ddots \\
\vdots & & & \ddots & & \ddots & \\
\vdots & & & & \ddots & & \ddots \\
 & & & & & \ddots &
\end{array}
$$

and further, if x is put equal to zero,

$$
\begin{array}{lllllll}
0^s & & & & & & \\
 & \Delta(0)^s & & & & & \\
1^s & & \Delta^2(0)^s & & & & \\
 & \Delta(1)^s & & \Delta^3(0)^s & & & \\
2^s & & \Delta^2(1)^s & & \Delta^4(0)^s & & \\
 & \Delta(2)^s & & \Delta^3(1)^s & & \ddots & \\
3^s & & \Delta^2(2)^s & & \Delta^4(1)^s & & \ddots \\
 & \Delta(3)^s & & \ddots & & \ddots & \\
4^s & & \ddots & & \ddots & & \ddots \\
\vdots & & & \ddots & & \ddots & \\
\vdots & & & & \ddots & & \ddots
\end{array}
$$

The leading differences $\Delta(0)^s, \Delta^2(0)^s, \ldots, \Delta^r(0)^s$ are named difference quotients of zero, or more often simply differences of zero.

It is easily proved that

$$\Delta^r(0)^s = r! \quad \text{if } r = s,$$

and

$$\Delta^r(0)^s = 0 \quad \text{if } r > s.$$

It is possible to show (see for example Milne-Thompson, p. 36) that the following recurrence relationship holds:

$$\Delta^r(0)^s = r\Delta^r(0)^{s-1} + r\Delta^{r-1}(0)^{s-1},$$

or, if each side is divided by $r!$, then

$$\frac{\Delta^r(0)^s}{r!} = r\,\frac{\Delta^r(0)^{s-1}}{r!} + \frac{\Delta^{r-1}(0)^{s-1}}{(r-1)!}.$$

It is curious, since these differences have been used by probabilists for over a century, that no table of them appeared before 1925, when Karl Pearson and E. M. Elderton tabled the expression $\Delta^r(0)^{r+s}/(r+s)!$ for values of r and s ranging by integer values from 1 to 20. W. L. Stevens, who was not aware of this earlier table, calculated $\Delta^r(0)^s/r!$ for r and s ranging by integer values from 1 to 25, in 1937. A straightforward relationship exists between the rth differences of zero and the Bernoulli numbers of order r, namely

$$\frac{\Delta^r(0)^s}{r!} = \frac{s!}{r!(s-r)!}\,B_{s-r}^{(-r)} \quad \text{for all } r \leqslant s.$$

Generally

$$B_s^{(r)} = -\frac{r}{s}\sum_{t=1}^{s}(-1)^t\,\frac{s!}{t!(s-t)!}\,B_t^{(1)}B_{s-r}^{(r)},$$

the first ten Bernoulli numbers obtained from this relationship being, after $B_0^{(r)}$,

$$B_0^{(r)} = 1, \quad B_1^{(r)} = -\frac{r}{2}, \quad B_2^{(r)} = \frac{r}{12}(3r-1), \quad B_3^{(r)} = -\frac{r^2}{8}(r-1),$$

$$B_4^{(r)} = \frac{r}{240}(15r^3 - 30r^2 + 5r + 2), \quad B_5^{(r)} = -\frac{r^2}{96}(r-1)(3r^2 - 7r - 2)$$

$$B_6^{(r)} = \frac{r}{4032}(63r^5 - 315r^4 + 315r^3 + 91r^2 - 42r - 16),$$

$$B_7^{(r)} = -\frac{r^2}{1152}(9r^5 - 63r^4 + 105r^3 + 7r^2 - 42r - 16),$$

$$B_8^{(r)} = \frac{r}{34560}(135r^7 - 1260r^6 + 3150r^5 - 840r^4 - 2345r^3 - 540r^2 + 404r + 144),$$

$$B_9^{(r)} = -\frac{r^2}{7680}(15r^7 - 180r^6 + 630r^5 - 448r^4 - 665r^3 + 100r^2 + 404r + 144),$$

$$B_{10}^{(r)} = \frac{r}{101,376}(99r^9 - 1485r^8 + 6930r^7 - 8778r^6 - 8085r^5 + 8195r^4 + 11,792r^3 + 2068r^2 - 2288r - 768).$$

When $r = 1$ these numbers reduce to the quantities usually designated Bernoulli numbers. It is possible, by the use of Bernoulli numbers of order r, to solve any probability problem where $s - r$ is small without recourse to tables.

A simple System of Arrangements

Assume that there is a box B which is divided into N equal and identical compartments. k identical balls, where $k \leqslant N$, are dropped into the box at random and no restriction is placed on the number of balls which may fall into any one compartment. The problem will consist in enumerating the number of different patterns into which the k balls may arrange themselves among the N compartments.

If it is equally likely that any one ball will fall in any one compartment, then the first ball will have a choice of N equally likely alternatives; so will the second, and the third, and so on, so that the total number of patterns in which the k balls may arrange themselves is N^k. This total number of patterns will be the sum of the sets of different patterns in which the k balls may fall. One set of patterns will be when the k balls fall into k different compartments and k only. They cannot fall into more than k compartments because there are only k balls. Another set of patterns will be when k balls fill exactly $k - 1$ compartments, which will imply that $k - 2$ compartments contain one ball each

and one compartment has 2 balls, and so on. The different patterns may be enumerated in the following way:

k balls in	k compartments in	$c_k . N(N-1)\dots(N-k+1)$	ways.
k „	$k-1$ „	$c_{k-1} . N(N-1)\dots(N-k+2)$	„
⋮	⋮	⋮	
k	$k-t$ „	$c_{k-t} . N(N-1)\dots(N-k+t+1)$	„
⋮	⋮	⋮	
k	1 „	$c_1 N$	„

$c_1, c_2, \dots, c_k$ are constants for a given k and depend only on k. That is to say, once the number of compartments to be filled is fixed (say $k-t$), the distribution of the $k-t$ compartments among the N has been given and it is only left to enumerate the possible distribution of the k balls in the $k-t$ compartments. All sets of patterns have been enumerated above and hence

$$N^k = c_1 \frac{N!}{(N-1)!} + c_2 \frac{N!}{(N-2)!} + \dots + c_t \frac{N!}{(N-t)!} + \dots + c_k \frac{N!}{(N-k)!}.$$

It remains to evaluate the c's. This follows immediately by differencing* both sides of the equation t times for $t = 1, 2, \dots, k$ and then putting N equal zero, whence

$$c = \frac{\Delta^t (0)^k}{t!}$$

and we have

$$N^k = \frac{N!}{(N-1)!} \frac{\Delta^1 (0)^k}{1!} + \frac{N!}{(N-2)!} \frac{\Delta^2 (0)^k}{2!} + \dots + \frac{N!}{(N-t)!} \frac{\Delta^t (0)^k}{t!} + \dots + \frac{N!}{(N-k)!} \frac{\Delta^k (0)^k}{k!}.$$

The enumeration is, however, not yet complete. We have shown that the total number of ways in which k balls may arrange

* This is in effect an application of Gregory's theorem, viz.

$$u(x+n\omega) = E^{n\omega} u(x) = (1+\omega \underset{\omega}{\Delta})^n u(x)$$

$$= u(x) + \frac{n!}{1!(n-1)!} \omega \underset{\omega}{\Delta^1} u(x) + \frac{n!}{2!(n-2)!} \omega^2 \Delta^2 u(x) + \dots + \omega^n \underset{\omega}{\Delta^n} u(x).$$

themselves in t compartments is c_t, as given above, but it is of interest to enumerate the different distributions of k *within* these t compartments. Thus we may wish to know in how many compartments there will be just one ball, and so on. This process is sometimes called enumerating the different partitions of k by t. An easy extension of the multinomial theorem, which gives the number of ways in which a_1 compartments contain just one ball, a_2 compartments just two balls, ..., a_j compartments just j balls, is

$$\frac{k!}{(1!)^{a_1}a_1!(2!)^{a_2}a_2!\dots(j!)^{a_j}a_j!},$$

where the restrictions are that

$$\sum_{i=1}^{j} a_i = t \quad \text{and} \quad \sum_{i=1}^{j} i.a_i = k$$

and some, but not all, of the a's may be equal to zero. The sum of this expression, taken over all possible partitions of k by t, will be equal to c_t.

Example. Five balls are dropped at random in 10 compartments. Given all conditions are equally probable, in how many ways can they arrange themselves in order to occupy 3 compartments and 3 compartments only?

Here $N = 10, k = 5, t = 3$.

The number of ways will therefore be

$$\frac{\Delta^3(0)^5}{3!}.10.9.8.$$

It has been shown that

$$\frac{\Delta^3(0)^5}{3!} = \frac{5!}{2!3!}B_2^{(-3)} = \frac{5}{2}\frac{5.4}{2} = 25.$$

The total number of ways will be therefore 18,000.

To enumerate these ways in detail we must discuss the different partitions of 5 by 3. These will be

$$\begin{matrix} 3 & 1 & 1 \\ 2 & 2 & 1 \end{matrix}$$

and we therefore have

$$\frac{5!}{(1!)^2 2!(3!)^1 1!} = 10, \quad \frac{5!}{(2!)^2 2!(1!)^1 1!} = 15,$$

which we note add together to make 25. Hence the number of ways in which 5 balls may occupy 3 out of 10 compartments, with 3 balls in one and 1 in each of the two others, is 7200 ways, and with 1 ball in one and 2 in each of the two others is 10,800.

Finally, since the balls are dropped at random and each ball is equally likely to drop in any given compartment we may calculate the probability for each partition. The total number of ways in which the 5 balls may distribute themselves in the 10 compartments is 10^5. Hence

$$P\{3, 1, 1 \mid 3, 5, 10\} = \frac{7200}{100{,}000} = 0{\cdot}072,$$

$$P\{2, 2, 1 \mid 3, 5, 10\} = \frac{10{,}800}{100{,}000} = 0{\cdot}108,$$

and the total chance that if the 5 balls are dropped at random they will occupy 5 compartments only is 0·180.

Example. What is the most likely distribution of balls if 5 are dropped at random in 10 compartments?

5 balls	in	1	compartment	in	10 ways
5 „		2	compartments	in	1,350 „
5 „		3	„	„	18,000 „
5 „		4	„	„	50,400 „
5 „		5	„	„	30,240 „
			Total		100,000 „

Hence the most likely distribution is 5 balls in 4 compartments.

The problem considered by Laplace was a little more complicated than the foregoing. He considered a list of n different numbers all of which had the same probability of being drawn. r of these numbers were randomly chosen. They were noted and returned to the population of n. He then discussed the probability that after i sets of drawings of r, q or more different numbers would have been seen.

The number of distributions possible in a single set of r drawings is

$$\frac{n!}{r!(n-r)!},$$

since all alternatives are given as equally probable. The number of distributions possible in i sets will be

$$\left(\frac{n!}{r!(n-r)!}\right)^i.$$

The number of cases in which the number 1 will not be drawn will be given by excluding that number from the list of n. This will be

$$\left(\frac{(n-1)!}{r!(n-r-1)!}\right)^i$$

and therefore the total number of ways in which the number 1 can be drawn is

$$\left(\frac{n!}{r!(n-r)!}\right)^i - \left(\frac{(n-1)!}{r!(n-r-1)!}\right)^i = \left(\frac{1}{r!}\right)^i \Delta^1((n-1)(n-2)\dots(n-r))^i.$$

Following the same argument it can be shown that the number of ways in which 1 and 2 may be drawn is

$$\left(\frac{1}{r!}\right)^i \Delta^2((n-2)(n-3)\dots(n-r))^i,$$

from which, by further extension of the argument, the number of ways in which q different numbers may be drawn is

$$\left(\frac{1}{r!}\right)^i \Delta^q((n-q)(n-q-1)\dots(n-r))^i = C(n,r,i,q) \text{ (say)}.$$

The probability that q different numbers will be seen in i sets of drawings will therefore be

$$C(n,r,i,q)\left(\frac{r!(n-r)!}{n!}\right)^i$$

$$= \frac{\Delta^q((n-q)(n-q-1)\dots(n-r))^i}{(n(n-1)\dots(n-r+1))^i} = P(n,r,i,q) \text{ (say)}.$$

COROLLARY. If the probability is required that after i sets of drawings all n of the numbers will have been seen, then, writing t for the dummy variable to be put equal to zero after differencing,

$$P(n,r,i,n) = \frac{\Delta^n(t(t-1)\dots(t-r+1))^i}{(n(n-1)\dots(n-r+1))^i}.$$

COROLLARY. If the number drawn on each occasion is 1, i.e. if $r = 1$, then the probability that after i drawings of one number all n of the numbers will have been seen is

$$P\{n,1,i,n\} = \frac{\Delta^n(0)^i}{n^i}.$$

Example. A court of discipline is drawn from 4 members, 2 of which are chosen at random for any one sitting. What is the chance that in six sittings all four will have served?

$$(n(n-1)(n-2)\dots(n-r+1))^i = (4!/2!)^6 = 12^6,$$

$$\Delta^4(t(t-1))^6 = \Delta^4(0)^{12} - 6\Delta^4(0)^{11} + 15\Delta^4(0)^{10} - 20\Delta^4(0)^9 + 15\Delta^4(0)^8 - 6\Delta^4(0)^7 + \Delta^4(0)^6,$$

which may be evaluated from either Stevens' tables or those of Pearson and Elderton. The required probability is 0·94.

Example. The committee of a learned society is 12 in number. One member retires each month and is replaced by a new person. If the retiring member is chosen randomly, what is the probability that after 12 months have passed, none of the members will then have served 12 months?

We require

$$P\{12, 1, 12, 12\} = \frac{\Delta^{12}(0)^{12}}{12!} = \frac{12!}{12^{12}}.$$

The evaluation of a probability such as $\Delta^r(0)^s$ is not easy when r and s are both large. The tables of the difference quotients of zero extend, as we have already pointed out, to r and $s = 25$ only. After these limits have been reached it becomes necessary to use an approximation to these differences if the probability is to be evaluated.

Karl Pearson discusses two such approximations, one due to Laplace and one due to De Moivre, and remarks that De Moivre's approximation may be preferred over that of Laplace in that fewer approximations are involved and the formula is on the whole easier of application. The problem would appear to be the replacement of the series

$$\frac{\Delta^r(0)^s}{r^s} = 1 - r\left(1-\frac{1}{r}\right)^s + \frac{r(r-1)}{2!}\left(1-\frac{2}{r}\right)^s - \frac{r(r-1)(r-2)}{3!}\left(1-\frac{3}{r}\right)^s + \dots$$

by one which is easily summable. If we write

$$\left(1-\frac{1}{r}\right)^{st} \quad \text{for} \quad \left(1-\frac{t}{r}\right)^s,$$

as did De Moivre, then

$$\frac{\Delta^r(0)^s}{r^s} \simeq \left(1 - \left(\frac{r-1}{r}\right)^s\right)^r.$$

For r large this approximation is adequate and enables the required probability to be calculated if the values of r and s fall outside the existing tables. For r small the Bernoulli polynomials of order r may be used.

REFERENCES AND READING

Discussion of the multinomial theorem is often very restricted in probability text-books. There is usually much more in text-books of statistical theory and the student should read the derivation of the χ^2 distribution from the multinomial distribution.

Enough has been given here for the student to understand what is meant by a difference quotient of zero. For those who wish to take the subject a little further there is L. M. Milne-Thompson, *Calculus of Finite Differences*. A knowledge of this calculus is useful for many problems in combinatorial analysis.

Applications of difference quotients of zero to probability problems will be found, among many other places, in P. S. Laplace, *Théorie des Probabilités* (1812), Livre II, chap. II; K. Pearson, Introduction to *Tables for Statisticians and Biometricians*, Part II; W. L. Stevens, *Ann. Eugen.* VIII, p. 57, 'Significance of Grouping'.

For further reading in the theory of combinatorial analysis the student might begin with P. A. MacMahon, *Elements of Combinatorial Analysis.*

CHAPTER X

RANDOM VARIABLES. ELEMENTARY LAWS AND THEOREMS

During the preceding chapters attention has been confined to discontinuous or discrete probabilities. This restriction of the field is purposeful in that in outlining a new subject it is simpler for the reader to understand if the sets of points which are discussed are denumerable. All fundamental theorems, however, relating to the addition and multiplication of probabilities do not state explicitly that the discontinuous case is being considered (except in the proof) and these theorems will be found to apply for the case where there is continuity or perhaps where there is a compound of continuity and discontinuity.

The distinction between discontinuity and continuity will be preserved in discussing random variables. It is comparatively easy to prove all theorems relating to random variables when the variable is discontinuous. When the variable is continuous the same theorems may be shown to be true using the theory of sets. For the person interested primarily in statistical applications, however, it is often sufficient to prove the theorem for the discontinuous case and to see intuitively that for the continuous variable the substitution of an integral for a summation sign will generalize the theorem.

DEFINITION. x is a random variable if, whatever the number a, there exists a probability that x is less than or equal to a, i.e. if $P\{x \leqslant a\}$ exists.

This is quite a general definition. Consider the case of a binomial probability.

$$P\{k \leqslant k_1\} = \sum_{k=0}^{[k_1]} \frac{n!}{k!(n-k)!} p^k q^{n-k}.*$$

The probability that $k \leqslant k_1$ exists and k is therefore a random variable. If x is normally distributed, i.e. if

$$P\{x \leqslant a\} = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \int_{-\infty}^{a} \exp\left[-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2\right] dx,$$

then x is a random variable, normally distributed.

* $[k_1]$ = the largest integer not greater than k_1.

DEFINITION. The elementary probability law of the discontinuous random variable x is a function the value of which corresponding to any given value, say $x = x'$, is the probability that x takes the value x', i.e.

$$p_x(x') = P\{x = x'\}.$$

When the probability law of x is discontinuous x will be referred to as a discontinuous random variable and when continuous as a continuous random variable.

DEFINITION. The integral probability law of a continuous random variable x is a function $F(x)$ having the property that

$$F(x) = P\{\alpha < x < \beta\} = \int_{\alpha}^{\beta} p(x)\,dx,$$

where α and β are any two numbers. $p(x)$ is sometimes called the elementary probability law of the continuous variable x and sometimes its frequency function.

Example. $$P_{n.k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

is the elementary probability law of a discontinuous random binomial variable k.

Example. $$p(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2\right]$$

may be spoken of as the elementary probability law of a continuous random normal variable x. Its integral probability law will be

$$P\{\alpha < x < \beta\} = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \int_{\alpha}^{\beta} \exp\left[-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2\right] dx.$$

DEFINITION. Assume that x is a discontinuous random variable which may take the mutually exclusive and only possible values $u_1, u_2, \ldots, u_m$. Let the elementary probability law of x be written $p_x(u_i)$ for $i = 1, 2, \ldots, m$. Then the mean value of x in repeated sampling or, in other words, the *expectation* of x is defined as

$$\mathscr{E}(x) = u_1 . p_x(u_1) + u_2 . p_x(u_2) + \ldots + u_m . p_x(u_m) = \sum_{i=1}^{m} u_i . p_x(u_i).*$$

* I do not like the growing English practice of writing the expectation of x as $E(x)$. E has passed into common use as a linear difference operator in the calculus of finite differences and there exists the possibility of confusion if the same symbol is used for expectation. I have followed the continental practice in using $\mathscr{E}$.

ILLUSTRATION. k is a discontinuous random variable which may take values $0, 1, 2, \ldots, n$ with corresponding probabilities $P_{n.0}, P_{n.1}, \ldots, P_{n.r}, \ldots, P_{n.n}$. What is the expectation of k?

$$\mathscr{E}(k) = 0.P_{n.0} + 1.P_{n.1} + 2.P_{n.2} + \ldots + r.P_{n.r} + \ldots + n.P_{n.n} = \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^k q^{n-k} = np.$$

COROLLARY. If k is a discontinuous random variable, so is k^2 or any power of k, and the expectation of k^2 is

$$\mathscr{E}(k^2) = 0^2.P_{n.0} + 1^2.P_{n.1} + \ldots + r^2.P_{n.r} + \ldots + n^2.P_{n.n} = \sum_{k=0}^{n} k^2 \frac{n!}{k!(n-k)!} p^k q^{n-k} = n(n-1)p^2 + np.$$

DEFINITION. The expectation of a continuous random variable x is defined as

$$\mathscr{E}(x) = \int_{-\infty}^{+\infty} x.p(x)\,dx,$$

where $p(x)$ is the elementary probability law of x as defined.

COROLLARY. If $p(x)$ is the elementary probability law of a continuous random variable x, then the expectation of a continuous function of x, say $f(x)$, will be

$$\mathscr{E}(f(x)) = \int_{-\infty}^{+\infty} f(x)\,p(x)\,dx.$$

It is clear that the expectation of a function cannot always be evaluated. For example, consider the simple probability law for x

$$p(x) = 1 \quad \text{for } 0 < x < 1$$
$$= 0 \quad \text{for } x \text{ outside these limits}$$

and find the expectation of $1/x$.

$$\mathscr{E}\left(\frac{1}{x}\right) = \int_0^1 \frac{1}{x}.1.dx = \log x \Big|_0^1$$

and this is infinite at the lower limit. Or again, suppose that the elementary probability law of the discontinuous random variable k is

$$p(k) = \frac{e^{-1}}{k!} \quad \text{for } k = 0, 1, 2, \ldots, +\infty$$

and find $\mathscr{E}(k!)$.

$$\mathscr{E}(k!) = \sum_{k=0}^{\infty} k! \frac{e^{-1}}{k!} = e^{-1} \sum_{k=0}^{\infty} 1,$$

which cannot be evaluated.

THEOREM. If x is a discontinuous random variable which may take values in ascending order of magnitude $u_1, u_2, \ldots, u_m$ (these values being mutually exclusive and the only possible ones), with corresponding probabilities $p_1, p_2, \ldots, p_m$, then

$$u_1 \leqslant \mathscr{E}(x) \leqslant u_m.$$

It is given that $u_1 \leqslant u_2 \leqslant \ldots \leqslant u_m.$

By definition

$$\mathscr{E}(x) = \sum_{i=1}^{m} u_i p_i \leqslant \sum_{i=1}^{m} u_m p_i = u_m,$$

$$\mathscr{E}(x) = \sum_{i=1}^{m} u_i p_i \geqslant \sum_{i=1}^{m} u_1 p_i = u_1.$$

Hence $$u_1 \leqslant \mathscr{E}(x) \leqslant u_m$$

and it follows that the expectation of x must lie between its greatest and its least values.

THEOREM. If x is a continuous random variable whose elementary probability law is $p(x)$, then the expectation of any bounded function, $f(x)$, of x exists, and is contained between the upper and lower bounds of the function.

A function $f(x)$ is said to be bounded if there exist two numbers m and M such that $$m \leqslant f(x) \leqslant M.$$

By definition

$$\mathscr{E}(f(x)) = \int_{-\infty}^{+\infty} f(x)\, p(x)\, dx \leqslant M \int_{-\infty}^{+\infty} p(x)\, dx = M.$$

$$\geqslant m \int_{-\infty}^{+\infty} p(x)\, dx = m.$$

Hence $$m \leqslant \mathscr{E}(f(x)) \leqslant M.$$

It has been tacitly assumed that $f(x)$ is real. If $f(x)$ is a complex function then it may be said to be bounded if there is a number M such that the modulus of this function does not exceed M. The expectation of the real and imaginary parts of the function may each be demonstrated to exist.

Example. The integral probability law of a random variable x is

$$F(x) = P\{\alpha < x < \beta\} = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \int_{\alpha}^{\beta} \exp\left[-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2\right] dx.$$

Find

(i) $\mathscr{E}(x)$, (ii) $\mathscr{E}(x^2)$.

$$\mathscr{E}(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\int_{-\infty}^{+\infty} x \exp\left[-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2\right] dx = \xi,$$

$$\mathscr{E}(x^2) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\int_{-\infty}^{+\infty} x^2 \exp\left[-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2\right] dx = \xi^2+\sigma^2.$$

The mean value in repeated sampling of a random variable x which follows the normal probability law is thus seen to be the mean of the normal curve and the expectation of its square $\sigma^2+\xi^2$.

Example. The integral probability law of a random variable x is

$$F(x) = P\{\alpha < x < \beta\} = \frac{1}{\Gamma(f)}\int_{\alpha}^{\beta} x^{f-1}e^{-x}\,dx \quad \text{for } 0 \leqslant \alpha \leqslant \beta < +\infty,$$

where f is an integer. Find the expectation of x^k.

$$\mathscr{E}(x^k) = \frac{1}{\Gamma(f)}\int_0^{+\infty} x^{k+f-1}e^{-x}\,dx = \frac{\Gamma(f+k)}{\Gamma(f)} = \frac{(f+k-1)!}{(f-1)!},$$

provided k is an integer, remembering the relationship

$$\Gamma(f) = (f-1)\,\Gamma(f-1) \quad \text{and} \quad \Gamma(0) = 1.$$

DEFINITION. The relative probability law of a discontinuous random variable x_1, relative to other random variables

$$x_2, \quad x_3, \quad \ldots, \quad x_k,$$

is a function the value of which for $x_1 = u$, given

$$x_2 = v, \quad x_3 = \xi, \quad \ldots, \quad x_k = w,$$

will be the relative probability that $x_1 = u$, given

$$x_2 = v, \quad x_3 = \xi, \quad \ldots, \quad x_k = w,$$

i.e.

$$p_{x_1 \mid x_2, x_3, \ldots, x_k}(u \mid v, \xi, \ldots, w) = P\{(x_1 = u) \mid (x_2 = v), (x_3 = \xi), \ldots, (x_k = w)\}.$$

For two variables this reduces to

$$p_{x_1 \mid x_2}(u \mid v) = P\{(x_1 = u) \mid (x_2 = v)\}.$$

This definition may simply (and obviously) be extended to meet the case of the continuous random variable.

DEFINITION. If the random variables x_1 and x_2 are independent, then

$$p_{x_1 | x_2}(u \mid v) = P\{(x_1 = u) \mid (x_2 = v)\} = P\{(x_1 = u)\} = p_{x_1}(u).$$

THEOREM. The expectation of the sum of two discontinuous random variables is the sum of their expectations, whether the variables are independent or not.

Consider two random variables x and y. x may take values $u_1, u_2, \ldots, u_n$, which are the only possible, with probabilities $p_1, p_2, \ldots, p_n$. y may take values $v_1, v_2, \ldots, v_m$, which are the only possible, with probabilities $p'_1, p'_2, \ldots, p'_m$. Let P_{ij} be the joint probability that x takes the value u_i while y takes the value v_j, i.e. let

$$P_{ij} = P\{x = u_i)\,(y = v_j)\},$$

then

$$\mathscr{E}(x+y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij}(u_i + v_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij} u_i + \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij} v_j.$$

The order of summation is quite arbitrary and we may write therefore

$$\mathscr{E}(x+y) = \sum_{i=1}^{n} u_i \sum_{j=1}^{m} P_{ij} + \sum_{j=1}^{m} v_j \sum_{i=1}^{n} P_{ij}.$$

Now since P_{ij} is the probability that x takes the value u_i while y takes the value v_j, $\sum_{j=1}^{m} P_{ij}$ will be the probability that x takes the value u_i while y takes any of the values $v_1, v_2, \ldots, v_m$. Hence

$$\sum_{j=1}^{m} P_{ij} = p_i, \quad \sum_{i=1}^{n} P_{ij} = p'_j.$$

It follows that

$$\mathscr{E}(x+y) = \sum_{i=1}^{n} u_i p_i + \sum_{j=1}^{m} v_j p'_j = \mathscr{E}(x) + \mathscr{E}(y)$$

and the theorem is proved.

THEOREM. The expectation of the sum of k discontinuous random variables is equal to the sum of their expectations.

Let the k discontinuous random variables be $x_1, x_2, \ldots, x_k$. By the preceding theorem

$$\mathscr{E}\left(\sum_{i=1}^{k} x_i\right) = \mathscr{E}(x_1 + x_2 + \ldots + x_k) = \mathscr{E}(x_1) + \mathscr{E}(x_2 + x_3 + \ldots + x_k)$$

and therefore by continued application of the theorem

$$\mathscr{E}\left(\sum_{i=1}^{k} x_i\right) = \sum_{i=1}^{k} \mathscr{E}(x_i).$$

Example. Assume that there are k random variables

$$x_1, \quad x_2, \quad \ldots, \quad x_k$$

and that

$$p_{x_i}(t) = \frac{m_i^t}{t!} e^{-m_i} \quad \text{for } t = 0, 1, 2, \ldots + \infty, \quad \text{and} \quad i = 1, 2, \ldots, k.$$

Find $$\mathscr{E}\left(\sum_{i=1}^{k} x_i \right).$$

For any x_i we have

$$\mathscr{E}(x_i) = \sum_{t=0}^{\infty} t \frac{m_i}{t!} e^{-m_i} = m_i.$$

By the preceding theorem

$$\mathscr{E}\left(\sum_{i=1}^{k} x_i \right) = \sum_{i=1}^{k} \mathscr{E}(x_i) = \sum_{i=1}^{k} m_i.$$

The theorems regarding the sum of k random variables can be proved to hold for either continuous or discontinuous variables. The proof of the theorem will be assumed for the former case.

THEOREM. x and y are two discontinuous random variables. If x is independent of y then y is independent of x.

Let the joint probability law of xy be p_{xy},

$$p_{xy}(uv) = P\{(x = u)(y = v)\} = P\{(x = u)\} P\{(y = v) \mid (x = u)\}$$
$$= P\{(y = v)\} P\{(x = u) \mid (y = v)\},$$

x is given independent of y. It follows by definition

$$P\{(y = v)\} P\{(x = u) \mid (y = v)\} = P\{(y = v)\} P\{(x = u)\}$$

and hence, from the expansion of the joint probability law that

$$P\{(y = v)\} = P\{(y = v) \mid (x = u)\}.$$

If x is independent of y then it follows that y must be independent of x.

THEOREM. The expectation of the product of two discontinuous random variables is equal to the product of their expectations *if the variables are independent.*

Let the two random independent variables be x and y. Let the only possible values for x be $u_1, u_2, \ldots, u_n$, with corresponding probabilities $p_1, p_2, \ldots, p_n$ and for y, $v_1, v_2, \ldots, v_m$, with corresponding probabilities $p'_1, p'_2, \ldots, p'_m$. Let P_{ij} be the probability that x takes the value u_i while y takes the value v_j.

Then $$\mathscr{E}(x.y) = \sum_{i=1}^{n} \sum_{j=1}^{m} u_i v_j P_{ij}.$$

Now

$$P_{ij} = P\{(x = u_i)(y = v_j)\} = P\{(x = u_i)\}\, P\{(y = v_j) \mid (x = u_i)\}$$

and because given x and y are independent it must be that

$$P_{ij} = P\{(x = u_i)(y = v_j)\} = p_i p'_j.$$

Substituting in the expression for the expectation of $(x.y)$ we have

$$\mathscr{E}(x.y) = \sum_{i=1}^{n} \sum_{j=1}^{m} u_i v_j p_i p'_j = \sum_{i=1}^{n} u_i p_i \sum_{j=1}^{m} v_j p'_j = \mathscr{E}(x)\,\mathscr{E}(y).$$

Thus if the two random variables are independent, the expectation of their product is equal to the product of their expectations.

These two theorems, regarding the sum and product of expectations of two random variables, will hold good whether the variables are continuous or discontinuous. The theorem regarding the product may be extended, as before, to cover the use of k random independent variables.

THEOREM. If $x_1, x_2, \ldots, x_k$ are k independent random variables, then the expectation of their product is equal to the product of their expectations.

By repeated application of the theorem for two variables it is seen that

$$\mathscr{E}\left(\prod_{i=1}^{k} x_i\right) = \mathscr{E}(x_1)\,\mathscr{E}\left(\prod_{i=2}^{k} x_i\right) = \mathscr{E}(x_1)\,\mathscr{E}(x_2) \ldots \mathscr{E}(x_k) = \prod_{i=1}^{k} \mathscr{E}(x_i)$$

and the theorem is proved.

DEFINITION. The standard error of a random variable x is defined as

$$\sigma_x = (\mathscr{E}(x - \mathscr{E}(x))^2)^{\frac{1}{2}}.$$

Example. If k is a random variable having as elementary probability law the binomial law of probabilities, what is its standard error?

It has been found previously that

$$\mathscr{E}(k) = np, \quad \mathscr{E}(k^2) = n(n-1)p^2 + np,$$

where n and p have their usual meanings. By definition

$$\sigma_k^2 = \mathscr{E}(k - \mathscr{E}(k))^2 = \mathscr{E}(k^2) - (\mathscr{E}(k))^2$$

and hence

$$\sigma_k^2 = np(1-p)$$

as found in chapter IV.

Example. n random independent variables $x_1, x_2, \ldots, x_n$ have the same probability law about which nothing is known except that the first two moments exist, viz.

$$\mathscr{E}(x_i) = \xi, \quad \mathscr{E}(x_i - \mathscr{E}(x_i))^2 = \sigma^2 \quad \text{for } i = 1, 2, \ldots, n.$$

Find the expectation of the mean of the x's and its standard error.

For any x it is given that

$$\mathscr{E}(x) = \xi$$

and $$\sigma_x^2 = \mathscr{E}(x - \mathscr{E}(x))^2 = \mathscr{E}(x^2) - (\mathscr{E}(x))^2 = \sigma^2.$$

Let $$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Then $$\mathscr{E}(\bar{x}) = \mathscr{E}\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) = \frac{1}{n} \sum_{i=1}^{n} \mathscr{E}(x_i) = \frac{1}{n} \sum_{i=1}^{n} \xi = \xi.$$

Thus the expectation of the mean of a sample, that is the mean value of the sample mean in repeated sampling, is equal to the population mean. Similarly

$$\sigma_{\bar{x}}^2 = \mathscr{E}\left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mathscr{E}\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)\right)^2 = \frac{1}{n^2} \mathscr{E}\left(\sum_{i=1}^{n} (x_i - \mathscr{E}(x_i))\right)^2$$

$$= \frac{1}{n^2}\left[\mathscr{E} \sum_{i=1}^{n} (x_i - \mathscr{E}(x_i))^2 + 2\mathscr{E} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (x_i - \mathscr{E}(x_i))\,(x_j - \mathscr{E}(x_j))\right].$$

It will be noted that in finding the expectation of the mean of the n variables no use was made of the fact that the variables were given independent. Thus the fact that the expectation of the sample mean is the population mean is unaltered by dependence between the x's. The same is not true for the standard error of the mean because the cross-products of the right-hand side of the expression immediately above can vanish only if the variables are independent. Given that the variables are independent it follows that

$$\mathscr{E} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (x_i - \mathscr{E}(x_i))\,(x_j - \mathscr{E}(x_j))$$

$$= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathscr{E}(x_i - \mathscr{E}(x_i))\,\mathscr{E}(x_j - \mathscr{E}(x_j)) = 0.$$

Hence the (standard error)^2 of the mean of a sample of n is

$$\sigma_{\bar{x}}^2 = \frac{1}{n^2} \sum_{i=1}^{n} \mathscr{E}(x_i - \mathscr{E}(x_i))^2 = \sigma^2/n,$$

a fact which is well known.

This result is true for any n independent variables possessing the same probability law and for which

$$\mathscr{E}(x_i) = \xi \quad \text{and} \quad \sigma_{x_i}^2 = \sigma^2,$$

although it is most commonly made use of in the normal case.

DEFINITION. The correlation coefficient, ρ_{ij}, between any two random variables x_i and x_j is defined as

$$\rho_{ij} = \frac{\mathscr{E}[(x_i - \mathscr{E}(x_i))\,(x_j - \mathscr{E}(x_j))]}{\sigma_{x_i}\,\sigma_{x_j}},$$

where σ_{x_i} and σ_{x_j} are the standard errors of x_i and x_j as previously defined.

THEOREM. The standard error of any linear function

$$y = \sum_{i=1}^{n} \alpha_i x_i$$

of n random variables $x_1, x_2, \ldots, x_n$ is

$$\sigma_y^2 = \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \alpha_i \alpha_j \sigma_i \sigma_j \rho_{ij},$$

where the α's are constants, σ_i and σ_j are the standard errors of x_i and x_j respectively, and ρ_{ij} is the correlation coefficient as defined above.

This theorem will be true for both discontinuous and continuous random variables.

Let $\mathscr{E}(x_i) = a_i \quad \text{for } i = 1, 2, \ldots, n.$

Then $\mathscr{E}(y) = \mathscr{E} \sum_{i=1}^{n} \alpha_i x_i = \sum_{i=1}^{n} \alpha_i a_i.$

By definition

$$\sigma_y^2 = \mathscr{E}(y - \mathscr{E}(y))^2 = \mathscr{E}\left(\sum_{i=1}^{n} \alpha_i (x_i - a_i) \right)^2,$$

from which, by expanding the bracket,

$$\sigma_y^2 = \mathscr{E}\left[\sum_{i=1}^{n} \alpha_i^2 (x_i - a_i)^2 \right] + 2\mathscr{E}\left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \alpha_i \alpha_j (x_i - a_i)\,(x_j - a_j) \right].$$

Applying the theorem that the expectation of a sum equals the sum of expectations and remembering the definition of the correlation coefficient we have

$$\sigma_y^2 = \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \alpha_i \alpha_j \sigma_i \sigma_j \rho_{ij},$$

which proves the theorem.

Example. If $\alpha_1 = \alpha_2 = \ldots = \alpha_n = \dfrac{1}{n}$, $a_1 = a_2 = \ldots = a_n = a$, and if $\sigma_1 = \sigma_2 = \ldots = \sigma_n = \sigma$ then $y = \bar{x}$, $\mathscr{E}(y) = a$ and

$$\sigma_y^2 = \frac{\sigma^2}{n} + \frac{2\sigma^2}{n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \rho_{ij}.$$

If, further, the variables are assumed independent then

$$\sigma_y^2 = \sigma^2/n$$

as found in a previous example.

Example. Find the standard error of the sum and the difference of two random variables (i) if they are dependent, (ii) if they are independent.

For the sum of two variables let

$$\alpha_3 = \alpha_4 = \ldots = \alpha_n = 0, \quad \alpha_1 = \alpha_2 = 1.$$

Then $$y = x_1 + x_2$$

and $$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2\sigma_1 \sigma_2 \rho_{12}.$$

If x_1 and x_2 are independent then

$$\sigma_y^2 = \sigma_1^2 + \sigma_2^2.$$

For the difference of two variables let

$$\alpha_3 = \alpha_4 = \ldots = \alpha_n = 0, \quad \alpha_1 = 1, \quad \alpha_2 = -1.$$

Then $$y = x_1 - x_2$$

and $$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_1 \sigma_2 \rho_{12}.$$

If x_1 and x_2 are independent then

$$\sigma_y^2 = \sigma_1^2 + \sigma_2^2.$$

Exercise. Find the standard error of the sum of three random variables, $x_1 + x_2 + x_3$, if the variables are dependent and show how this simplifies if they are assumed independent.

Example. A bag contains N balls. Each ball has a number stamped on it, the numbers being $u_1, u_2, \ldots, u_N$. It may be assumed that it is equiprobable that any one ball will be drawn and the probability of drawing u_i is $1/N$ for $i = 1, 2, \ldots, N$. From this bag n balls are drawn and no ball is replaced after drawing. What is the standard error of the mean of the n numbers thus drawn?

Let the numbers drawn be $x_1, x_2, \ldots, x_n$. It is required therefore to find $\sigma_{\bar{x}}$, where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The expectation of any given number will be

$$\mathscr{E}(x_i) = u_1 \frac{1}{N} + u_2 \frac{1}{N} + \ldots + u_N \frac{1}{N} = \frac{1}{N} \sum_{t=1}^{N} u_t = \bar{u} \quad \text{(say)}.$$

Hence the expectation of $\bar{x}$ will be

$$\mathscr{E}(\bar{x}) = \mathscr{E}\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) = \frac{1}{n} \sum_{i=1}^{n} \mathscr{E}(x_i) = \bar{u}.$$

This result is what might have been expected. The standard error of $\bar{x}$ is not, however, so easily intuitive. Consider first the standard error, σ_i, of x_i.

$$\sigma_i^2 = \mathscr{E}(x_i - \mathscr{E}(x_i))^2 = \mathscr{E}(x_i - \bar{u})^2.$$

From the definition of an expectation it follows that

$$\sigma_i^2 = \mathscr{E}(x_i - \bar{u})^2 = \frac{1}{N} \sum_{t=1}^{N} (u_t - \bar{u})^2 = V_u \quad \text{(say)}.$$

V_u, the variance of the u's, is constant. From the theorem on the expectation of a linear function we may write immediately

$$\sigma_{\bar{x}}^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 + \frac{2}{n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sigma_i \sigma_j \rho_{ij}.$$

The first summation on the right-hand side may be evaluated but it remains to calculate the double summation.

$$\rho_{ij} \sigma_i \sigma_j = \mathscr{E}[(x_i - \mathscr{E}(x_i))(x_j - \mathscr{E}(x_j))]$$

by definition. Again appealing to the definition of an expectation, i.e. the summation of all possible values a random variable may

take multiplied by the probability that it takes each separate value, we have, treating the product as a unit

$$\mathscr{E}[(x_i-\mathscr{E}(x_i))(x_j-\mathscr{E}(x_j))] = \sum_{t=1}^{N-1}\sum_{l=t+1}^{N}(u_t-\bar{u})(u_l-\bar{u})\frac{(N-2)!\,2!}{N!}.$$

The (standard error)^2 of $\bar{x}$ accordingly will be

$$\sigma_{\bar{x}}^2 = \frac{V_u}{n}+\frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{t=1}^{N-1}\sum_{l=t+1}^{N}(u_t-\bar{u})(u_l-\bar{u})\frac{2}{N(N-1)}.$$

This expression may be simplified by the following device. It is clear that

$$\sum_{t=1}^{N}(u_t-\bar{u}) = 0.$$

Squaring each side

$$0 = \left[\sum_{t=1}^{N}(u_t-\bar{u})\right]^2 = \sum_{t=1}^{N}(u_t-\bar{u})^2+2\sum_{t=1}^{N-1}\sum_{l=t+1}^{N}(u_t-\bar{u})(u_l-\bar{u})$$

and therefore

$$-NV_u = 2\sum_{t=1}^{N-1}\sum_{l=t+1}^{N}(u_t-\bar{u})(u_l-\bar{u}).$$

By substitution the (standard error)^2 of $\bar{x}$ will reduce to

$$\sigma_{\bar{x}}^2 = \frac{V_u}{n}\left[1-\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{2}{n(N-1)}\right].$$

This last double summation is of a constant and it is only necessary therefore to enumerate the number of constants concerned. This will be

$$\frac{n!}{(n-2)!\,2!},$$

whence $\sigma_{\bar{x}}^2$ becomes $\sigma_{\bar{x}}^2 = \frac{V_u}{n}\left(\frac{N-n}{N-1}\right)$.

Note (i). When $n = 1$ the expression for $\sigma_{\bar{x}}^2$ reduces to the variance of the u's which might be expected.

Note (ii). If the x's were independent, that is if each ball had been replaced after being drawn and its number noted, then

$$\sigma_{\bar{x}}^2 = \frac{V_u}{n}$$

from the expression for the standard error of a linear function. If this expression is compared with that for the standard error of

the mean when the drawings are made without replacement, it will be seen that the latter is less, provided n is greater than unity.

THEOREM. Given that (i) $x_1, x_2, \ldots, x_n$, are n random independent variables, (ii) $\mathscr{E}(x_i) = a_i$ for $i = 1, 2, \ldots, n$, (iii) $\mathscr{E}(x_i - a_i)^2 = \sigma_i^2$ for $i = 1, 2, \ldots, n$, then

$$\mathscr{E}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right) = \frac{n-1}{n^2}\sum_{i=1}^{n}\sigma_i^2 + \frac{1}{n}\sum_{i=1}^{n}(a_i-\bar{a})^2,$$

where $$\bar{a} = \mathscr{E}(\bar{x}) \quad \text{and} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Possibly the simplest method of attack for this and similar problems is to employ the device of inserting the expectation of each variable within the bracket. Neglecting the factor $1/n$ for the time being, write

$$\mathscr{E}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right) = \mathscr{E}\sum_{i=1}^{n}[(x_i-a_i)-(\bar{x}-\bar{a})+(a_i-\bar{a})]^2.$$

The usefulness of this device is clear. In the expansion of the bracket a number of cross-products will appear. Provided x_i and x_j are independent, and they are so given, then

$$\mathscr{E}[(x_i-a_i)(x_j-a_j)] = 0$$

and the algebra is accordingly simplified. The expansion of the bracket becomes

$$\mathscr{E}\sum_{i=1}^{n}[(x_i-a_i)-(\bar{x}-\bar{a})+(a_i-\bar{a})]^2 = \mathscr{E}\sum_{i=1}^{n}(x_i-a_i)^2$$
$$+\mathscr{E}\sum_{i=1}^{n}(\bar{x}-\bar{a})^2 + \sum_{i=1}^{n}(a_i-\bar{a})^2 - 2\mathscr{E}\sum_{i=1}^{n}[(x_i-a_i)(\bar{x}-\bar{a})].$$

Using the theorem regarding the expectation of a sum it is seen that the evaluation of the first term is immediate. We need to consider the second and fourth terms.

$$\mathscr{E}\sum_{i=1}^{n}(\bar{x}-\bar{a})^2 = n\mathscr{E}(\bar{x}-\bar{a})^2 = \frac{1}{n}\mathscr{E}\left(\sum_{i=1}^{n}(x_i-a_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathscr{E}(x_i-a_i)^2.$$

Also

$$\mathscr{E}\sum_{i=1}^{n}[(x_i-a_i)(\bar{x}-\bar{a})] = \frac{1}{n}\mathscr{E}\left(\sum_{i=1}^{n}(x_i-a_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathscr{E}(x_i-a_i)^2.$$

Hence

$$\mathscr{E}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)=\sum_{i=1}^{n}\mathscr{E}(x_i-a_i)^2+\frac{1}{n}\sum_{i=1}^{n}\mathscr{E}(x_i-a_i)^2$$
$$+\sum_{i=1}^{n}(a_i-\bar{a})^2-\frac{2}{n}\sum_{i=1}^{n}\mathscr{E}(x_i-a_i)^2$$
$$=\frac{n-1}{n}\sum_{i=1}^{n}\sigma_i^2+\sum_{i=1}^{n}(a_i-\bar{a})^2,$$

or introducing the factor $1/n$ previously neglected

$$\mathscr{E}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)=\frac{n-1}{n^2}\sum_{i=1}^{n}\sigma_i^2+\frac{1}{n}\sum_{i=1}^{n}(a_i-\bar{a})^2.$$

COROLLARY. Suppose the n random independent variables x to follow the same probability law with mean a and standard deviation σ. This might be the case for a sample of n individuals which had been randomly and independently drawn from such a population. In this case

$$\mathscr{E}(x_i)=a_i=a,\quad \mathscr{E}(x_i-a_i)^2=\sigma_i^2=\sigma^2,\quad \text{for } i=1,2,\ldots,n,$$

and the theorem reduces to

$$\mathscr{E}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)=\frac{n-1}{n}\sigma^2.$$

Now $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ will be recognized as the sample (standard deviation)^2, s^2. It follows then that the mean value in repeated sampling of the square of the sample standard deviation is not equal to the square of the population standard deviation, but in fact

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

If, therefore, the sample standard deviation is used as an estimate of the population standard deviation, in the long run the tendency will be to underestimate it. If it is desired to obtain a sum of squares which in repeated sampling will average out to be σ, then it is clear from the equation

$$\mathscr{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)=\sigma^2$$

that the factor $1/(n-1)$ should replace the $1/n$ of the sample (standard deviation)^2. This new expression is not a sample

(standard deviation)^2 nor a population (standard deviation)^2; it is an expression which, averaged over a series of experiments, will approximate to the population (standard deviation)^2.

This elementary piece of algebra supplies the answer to the confused question 'Do I divide by n or $n-1$ to obtain the standard deviation?' If a measure of the scatter in the sample is required then the sample standard deviation

$$s = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{1}{2}}$$

must be calculated. If it is desired to estimate the population (standard deviation)^2 then the expression $s^2(n/n-1)$ may be calculated because in the long run it will be equal to σ^2.

Exercise. Given (i) n independent random variables each of which has the same probability law,

(ii) $\mathscr{E}(x_i) = a$, (iii) $\mathscr{E}(x_i-a_i)^2 = \sigma^2$,

(iv) $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$, calculate

$$\sigma_{s^2}^2 = \mathscr{E}(s^2-\mathscr{E}(s^2))^2.$$

[Note $\mathscr{E}(s^2)$ was found above.]

Example. (*Weldon's dice problem.*) n_1 white dice and n_2 red dice are shaken together and thrown on a table. The sums of the dots on the upper faces are noted. The red dice are then picked up and thrown again among the white dice left on the table. The sum of the dots on the upper faces is again noted. What is the correlation between the first and second sums?

Let the numbers on the upper faces of the white dice be $t_{11}, t_{12}, \ldots, t_{1n_1}$, the numbers on the upper faces of the red dice at the first throw be $t_{21}, t_{22}, \ldots, t_{2n_2}$, and at the second throw $t_{31}, t_{32}, \ldots, t_{3n_2}$. Further let

$$t_1 = \sum_{i=1}^{n_1} t_{1i}, \quad t_2 = \sum_{i=1}^{n_2} t_{2i}, \quad t_3 = \sum_{i=1}^{n_2} t_{3i}.$$

It is required to find the correlation between t_1+t_2 and t_1+t_3. Consider first just one die. If t_{xi} be the number of dots on its upper face after throwing, then

$$\mathscr{E}(t_{xi}) = \tfrac{1}{6}.1+\tfrac{1}{6}.2+\tfrac{1}{6}.3+\tfrac{1}{6}.4+\tfrac{1}{6}.5+\tfrac{1}{6}.6 = \tfrac{7}{2}$$

and $\mathscr{E}(t_{xi}^2) = \tfrac{1}{6}.1^2+\tfrac{1}{6}.2^2+\tfrac{1}{6}.3^2+\tfrac{1}{6}.4^2+\tfrac{1}{6}.5^2+\tfrac{1}{6}.6^2 = \tfrac{91}{6}$.

Therefore $\sigma_{tx i}^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$.

For the sum

$$\mathscr{E}(t_1+t_2) = \mathscr{E}\left(\sum_{i=1}^{n_1} t_{1i} + \sum_{i=1}^{n_2} t_{2i}\right) = \tfrac{7}{2}(n_1+n_2)$$

and similarly $\mathscr{E}(t_1+t_3) = \frac{7}{2}(n_1+n_2)$.

Applying the elementary theorems on expectations, and remembering that any one die is independent of any other die, it may be shown that

$$\sigma_{t_1+t_2}^2 = \tfrac{35}{12}(n_1+n_2) \quad \text{and} \quad \sigma_{t_1+t_3}^2 = \tfrac{35}{12}(n_1+n_2).$$

If ρ is the coefficient of correlation between t_1+t_2 and t_1+t_3, then by definition

$$\rho\sigma_{t_1+t_2}\sigma_{t_1+t_3} = \mathscr{E}[((t_1+t_2)-\tfrac{7}{2}(n_1+n_2))\,((t_1+t_3)-\tfrac{7}{2}(n_1+n_2))].$$

Replacing the individual sums on the right-hand side we shall have

$$\rho\sigma_{t_1+t_2}\sigma_{t_1+t_3} = \mathscr{E}\left[\left(\sum_{i=1}^{n_1}(t_{1i}-\tfrac{7}{2}) + \sum_{i=1}^{n_2}(t_{2i}-\tfrac{7}{2})\right) \times \left(\sum_{i=1}^{n_1}(t_{1i}-\tfrac{7}{2}) + \sum_{i=1}^{n_2}(t_{3i}-\tfrac{7}{2})\right)\right]$$

$$= \mathscr{E}\left[\sum_{i=1}^{n_1}(t_{1i}-\tfrac{7}{2})^2\right] = \tfrac{35}{12}n_1.$$

The correlation coefficient between the two sums (t_1+t_2) and (t_1+t_3) is therefore

$$\rho = \frac{n_1}{n_1+n_2}.$$

It may be noted that this is a simple example of a more general case. If X and Y are two random variables, each composed of the sum of two random variables

$$X = x+t, \quad Y = y+t,$$

then there will be a correlation between X and Y.

DEFINITION. If a random variable x may take only the values zero or unity then x is defined as a characteristic random variable.

If x is a characteristic random variable with probability p that it takes the value 1, and $1-p$ that it takes the value 0, then

$$\mathscr{E}(x) = 1.p+0.(1-p) = p.$$

A characteristic random variable may be seen, in this way, to have the interesting property that

$$\mathscr{E}(x) = \mathscr{E}(x^2) = \ldots = \mathscr{E}(x^k) = \ldots = p,$$

for
$$\mathscr{E}(x^k) = p.1^k + (1-p).0^k = p$$

and this will be true whatever p.

Example. Consider a series of trials, n in number, in each of which the constant probability of a success is p. If with each of these trials is associated a characteristic random variable x which will take the value 1 if the trial succeeds and 0 if it fails then
$$\mathscr{E}(nx) = n\mathscr{E}(x) = n(1.p + 0.(1-p)) = np.$$

Example on expectations. In the previous chapter we have discussed the enumeration of the patterns in which k balls will fall when dropped randomly in a box of N compartments. It is easy to show by direct argument that the average proportion of compartments filled in repeated sampling is

$$\mathscr{E}\left(\frac{k}{N}\right) = \left[1 - \left(1 - \frac{1}{N}\right)^k\right].$$

This result may also be achieved by application of the theorems of this chapter. The probability that exactly t compartments will be filled is
$$\frac{\Delta^t(0)^k}{t!}\,\frac{N!}{(N-t)!}\,\frac{1}{N^k}.$$

Hence

$$\mathscr{E}\left(\frac{k}{N}\right) = \frac{1}{N}\sum_{t=1}^{k} t\,\frac{\Delta^t(0)^k}{t!}\,\frac{N!}{(N-t)!}\,\frac{1}{N^k} = \frac{1}{N}\sum_{t=1}^{k}\frac{N\Delta}{N^k}\,\frac{\Delta^{t-1}(0)^k}{(t-1)!}\,\frac{(N-1)!}{(N-t)!}$$

$$= \frac{1}{N}\,\frac{N\Delta(1+\Delta)^{N-1}(0)^k}{N^k}.$$

If we use the linear difference notation and write

$$E = 1 + \Delta,$$

then
$$E^{N-1}(0)^k = (N-1)^k *$$

and
$$\mathscr{E}\left(\frac{k}{N}\right) = \frac{1}{N}\,\frac{N\Delta(N-1)^k}{N^k} = 1 - \left(1 - \frac{1}{N}\right)^k.$$

* Ex_s is defined as $Ex_s = x_{s+1}$. Hence
$$E^{N-1}t^k = E^{N-2}(t+1)^k = E^{N-3}(t+2)^k = \ldots = (t+N-1)^k.$$

Similarly the mean value of $(k/N)^2$ in repeated sampling will be

$$\mathscr{E}\left(\frac{k^2}{N^2}\right) = \frac{1}{N^2}\sum_{t=1}^{k} t^2 \frac{\Delta^t(0)^k}{t!}\frac{N!}{(N-t)!}\frac{1}{N^k}$$

$$= \frac{1}{N^{2+k}}[N(N-1)\Delta^2(1+\Delta)^{N-2}(0)^k + N\Delta(1+\Delta)^{N-1}(0)^k]$$

$$= \frac{N-1}{N}\left[\left(1-\frac{2}{N}\right)^k + 1 - 2\left(1-\frac{1}{N}\right)^k\right] + \frac{1}{N}\left[1-\left(1-\frac{1}{N}\right)^k\right]$$

from which it follows that the variance of k/N, say $\sigma^2_{k/N}$ is

$$\sigma^2_{k/N} = \left[\left(1-\frac{2}{N}\right)^k - \left(1-\frac{1}{N}\right)^{2k}\right] + \frac{1}{N}\left[\left(1-\frac{1}{N}\right)^k - \left(1-\frac{2}{N}\right)^k\right].$$

Exercise. Find the third and fourth moments about the mean in repeated sampling of the proportion k/N.

REFERENCES AND READING

Suitable exercises for most of the processes outlined in this chapter may be found in J. V. Uspensky, *Introduction to Mathematical Probability.*

Again, W. Whitworth, *Choice and Chance*, gives a choice of many ingenious examples.

CHAPTER XI

MOMENTS OF SAMPLING DISTRIBUTIONS

In the previous chapter there has been set out the mathematical technique whereby the expectation of a random variable, or of a function of random variables, may be calculated. One of the main uses to which the statistician puts this technique is for the calculation of the theoretical moments of sampling distributions. Such calculations are straightforward and are really only exercises on the use of the theorems already proved, but since they are of importance we shall consider them here in some detail. The connexion between the random variable of the probabilist and the sampling unit of the statistician is usually made in the following way. A single unit is randomly drawn from some population the probability distribution of which may be completely known, or may be incompletely specified. With this single unit we associate a random variable which has the same probability distribution as the parent population; this single unit will thus be one observed value of the given random variable. Hence, if we randomly draw a sample of units from a given population, we may associate a random variable with each element of the sample in order of drawing, and to find the mean value in repeated sampling of a function of the observed values it will only be necessary to discuss the mathematical expectation of the same function of the associated random variables. We shall begin by finding the moments of the sampling distribution of the means of samples the units of which have been randomly and independently drawn from an infinite population or, more precisely, from a finite population with replacement after drawing. In this latter case the population is effectively infinite for provided each unit is returned after drawing it is not possible to exhaust the population.

Sampling Moments of the Mean: Population Infinite

Assume that a sample of n units is randomly and independently drawn from a population the distribution of which is not specified but the first four moments of which are known to exist. Let these population moments be μ_1', μ_2, μ_3 and μ_4, where the μ's have their usual meaning. Associate with the sample units in order of drawing n random variables $x_1, x_2, \ldots, x_n$, and let

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

It is required to find $\mu_1'(\bar{x})$, $\mu_2(\bar{x})$, $\mu_3(\bar{x})$, $\mu_4(\bar{x})$.

It has already been shown in the previous chapter, that

$$\mu_1'(\bar{x}) = \mathscr{E}(\bar{x}) = \mu_1'$$

and

$$\mu_2(\bar{x}) = \mathscr{E}(\bar{x} - \mathscr{E}(\bar{x}))^2 = \mu_2/n$$

from which, if we apply the usual convention of writing $\mu_2 = \sigma^2$ we have that

$$\sigma(\bar{x}) = \sigma/\sqrt{n}.$$

The third and fourth moments follow in similar fashion but require a little more enumeration.

$$\mu_3(\bar{x}) = \mathscr{E}(\bar{x} - \mathscr{E}(\bar{x}))^3 = \frac{1}{n^3} \mathscr{E} \left[\sum_{i=1}^{n} (x_i - \mu_1') \right]^3$$

$$= \frac{1}{n^3} \left[\mathscr{E} \sum_{i=1}^{n} (x_i - \mu_1')^3 + 3\mathscr{E} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [(x_i - \mu_1')^2 (x_j - \mu_1') + (x_i - \mu_1') (x_j - \mu_1')^2] \right.$$

$$\left. + 6\mathscr{E} \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} (x_i - \mu_1') (x_j - \mu_1') (x_k - \mu_1') \right].$$

The sample units, and hence the random variables, are given independent. The second and third terms therefore vanish and we have

$$\mu_3(\bar{x}) = \frac{1}{n^3} \sum_{i=1}^{n} (x_i - \mu_1')^3 = \frac{\mu_3}{n^2}.$$

For $\mu_4(\bar{x})$ similarly

$$\mu_4(\bar{x}) = \mathscr{E}(\bar{x}-\mathscr{E}(\bar{x}))^4 = \frac{1}{n^4}\mathscr{E}\left[\sum_{i=1}^{n}(x_i-\mu_1')\right]^4$$

$$= \frac{1}{n^4}\Bigg[\mathscr{E}\sum_{i=1}^{n}(x_i-\mu_1')^4 + 4\mathscr{E}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}[(x_i-\mu_1')^3(x_j-\mu_1') + (x_i-\mu_1')(x_j-\mu_1')^3]$$

$$+ 6\mathscr{E}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i-\mu_1')^2(x_j-\mu_1')^2$$

$$+ 12\mathscr{E}\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{k=j+1}^{n}[(x_i-\mu_1')^2(x_j-\mu_1')(x_k-\mu_1') + (x_i-\mu_1')(x_j-\mu_1')^2(x_k-\mu_1') + (x_i-\mu_1')(x_j-\mu_1')(x_k-\mu_1')^2]$$

$$+ 24\mathscr{E}\sum_{i=1}^{n-3}\sum_{j=i+1}^{n-2}\sum_{k=j+1}^{n-1}\sum_{t=k+1}^{n}(x_i-\mu_1')(x_j-\mu_1')(x_k-\mu_1')(x_t-\mu_1')\Bigg].$$

Again, because of independence, all the terms except the first and third vanish and

$$\mu_4(\bar{x}) = \frac{\mu_4 + 3(n-1)\mu_2^2}{n^3}.$$

The $\beta_1(\bar{x})$ and $\beta_2(\bar{x})$ of the distribution of the means will be

$$\beta_1(\bar{x}) = \frac{\mu_3^2(\bar{x})}{\mu_2^3(\bar{x})} = \frac{\beta_1}{n}$$

and

$$\beta_2(\bar{x}) = \frac{\mu_4(\bar{x})}{\mu_2^2(\bar{x})} = 3 + \frac{\beta_2 - 3}{n}.$$

It is clear therefore that whatever the β_1 and β_2 of the parent population (provided they exist), as n increases the $\beta_1(\bar{x})$ and $\beta_2(\bar{x})$ will tend to those of the normal population. If the parent population is known to be normally distributed then $\beta_1 = 0$ and $\beta_2 = 3$ and therefore so do $\beta_1(\bar{x})$ and $\beta_2(\bar{x})$.

Sampling Moments of the Mean: Population Finite

While the concept of the infinite population presents no difficulties to the probabilist it is rare for the statistician to find a population which he could not count if he had sufficient time and patience. Also it is unusual for the statistician to be able to

sample with replacement. However, generally it is assumed, and it is often the case, that the population is large enough, and the sample small enough, for the sampling moments of the mean to be used assuming the population is infinite. We shall derive the sampling moments of the mean when the population is finite and the drawings are made without replacement, and the student may then judge for himself the degree of approximation involved.

Assume that the population consists of N elements or units, and that the characteristic u we are considering takes values $u_1, u_2, \ldots, u_N$ in the population. Let

$$\mu_1' = \frac{1}{N}\sum_{g=1}^{N} u_g, \qquad \mu_2 = \frac{1}{N}\sum_{g=1}^{N} (u_g - \mu_1')^2,$$

$$\mu_3 = \frac{1}{N}\sum_{g=1}^{N} (u_g - \mu_1')^3, \quad \mu_4 = \frac{1}{N}\sum_{g=1}^{N} (u_g - \mu_1')^4.$$

Suppose that a sample of n units is drawn from the population of N units, and associate with each unit of the sample, in order of drawing, a random variable. We have therefore n random variables $x_1, x_2, \ldots, x_n$ but they are no longer independent as in the previous case. We require to find

$$\mu_1'(\bar{x}), \quad \mu_2(\bar{x}), \quad \mu_3(\bar{x}), \quad \mu_4(\bar{x}).$$

In an example in the preceding chapter it was shown that

$$\mu_1'(\bar{x}) = \mathscr{E}(\bar{x}) = \mu_1',$$

$$\mu_2(\bar{x}) = \mathscr{E}(\bar{x} - \mathscr{E}(\bar{x}))^2 = \frac{\mu_2}{n}\frac{N-n}{N-1}.$$

The expansions for $\mu_3(\bar{x})$ and $\mu_4(\bar{x})$ will be the same as in the infinite case but when we come to take expectations the terms will no longer vanish because of the lack of independence. Thus we have that

$$\mu_3(\bar{x}) = \frac{1}{n^3}\Bigg[\sum_{i=1}^{n} \mathscr{E}(x_i - \mu_1')^3 + 3\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} [\mathscr{E}(x_i - \mu_1')^2 (x_j - \mu_1') + \mathscr{E}(x_i - \mu_1')(x_j - \mu_1')^2] + 6\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{t=j+1}^{n} \mathscr{E}(x_i - \mu_1')(x_j - \mu_1')(x_t - \mu_1')\Bigg].$$

We shall consider the evaluation of the middle term in detail, and leave the reader to fill in the calculations for the last term. It is required to find

$$\mathscr{E}(x_i-\mu_1')^2(x_j-\mu_1').$$

x_i and x_j are random variables, and we may make an appeal to the definition of an expectation and write

$$\mathscr{E}(x_i-\mu_1')^2(x_j-\mu_1') = \frac{1}{N(N-1)}\sum_{g=1}^{N-1}\sum_{h=g+1}^{N}[(u_g-\mu_1')^2(u_h-\mu_1') + (u_g-\mu_1')(u_h-\mu_1')^2].$$

The main difficulty of the problem rests in the evaluation of the sum on the right-hand side. Consider

$$\sum_{g=1}^{N}(u_g-\mu_1')^2\sum_{h=1}^{N}(u_h-\mu_1').$$

This product is equal to zero, since the second sum is zero. Hence if we expand the summations we have

$$0 = \sum_{g=1}^{N}(\mu_g-\mu_1')^3 + \sum_{g=1}^{N-1}\sum_{h=g+1}^{N}[(u_g-\mu_1')^2(u_h-\mu_1') + (u_g-\mu_1')(u_h-\mu_1')^2]$$

and

$$-N\mu_3 = \sum_{g=1}^{N-1}\sum_{h=g+1}^{N}[(u_g-\mu_1')^2(u_h-\mu_1') + (u_g-\mu_1')(u_h-\mu_1')^2].$$

It is clear then that if we consider

$$\mathscr{E}(x_i-\mu_1')^2(x_j-\mu_1')$$

together with $\mathscr{E}(x_i-\mu_1')(x_j-\mu_1')^2$

we shall have that

$$3\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}[\mathscr{E}(x_i-\mu_1')^2(x_j-\mu_1') + \mathscr{E}(x_i-\mu_1')(x_j-\mu_1')^2] = -6\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{N\mu_3}{N(N-1)} = -\frac{3n(n-1)}{N-1}\mu_3.$$

Again, by considering the product

$$\sum_{g=1}^{N}(u_g-\mu_1')\sum_{h=1}^{N}(u_h-\mu_1')\sum_{v=1}^{N}(u_v-\mu_1') = 0,$$

it may be shown that

$$6\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{t=j+1}^{n}\mathscr{E}(x_i-\mu_1')(x_j-\mu_1')(x_t-\mu_1') = \frac{2n(n-1)(n-2)}{(N-1)(N-2)}\mu_3,$$

from which it follows that

$$\mu_3(\bar{x}) = \frac{\mu_3}{n^3}\left[n - \frac{3n(n-1)}{N-1} + \frac{2n(n-1)(n-2)}{(N-1)(N-2)}\right].$$

This simplifies to

$$\mu_3(\bar{x}) = \frac{(N-n)(N-2n)}{n^2(N-1)(N-2)}\mu_3.$$

The calculations for $\mu_4(\bar{x})$ may be carried out along similar lines.

Writing down the expectations it is clear that in order to evaluate the sums it will be necessary to consider the expansions of

$$\sum_{g=1}^{N}(u_g-\mu_1')^3\sum_{h=1}^{N}(u_h-\mu_1') = 0,$$

$$\left[\sum_{g-1}^{N}(u_g-\mu_1')^2\right]^2 = N^2\mu_2^2,$$

$$\sum_{g=1}^{N}(u_g-\mu_1')^2\sum_{h=1}^{N}(u_h-\mu_1')\sum_{v=1}^{N}(u_v-\mu_1') = 0,$$

$$\left[\sum_{g=1}^{N}(u_g-\mu_1')\right]^4 = 0.$$

The reader should go through the algebra involved in order to get a facility in expansion and enumeration of the products of sums of variables. This algebra is quite straightforward and it is easy to show that

$$\mu_4(\bar{x}) = \frac{1}{n^4}\left[n\mu_4 - 4.\frac{n(n-1)}{N-1}\mu_4 + 3.\frac{n(n-1)}{N-1}[N\mu_2^2-\mu_4]\right.$$

$$+ 6.\frac{n(n-1)(n-2)}{(N-1)(N-2)}[2\mu_4 - N\mu_2^2]$$

$$\left.+\frac{n(n-1)(n-2)(n-3)}{(N-1)(N-2)(N-3)}[3N\mu_2^2-6\mu_4]\right],$$

whence, collecting terms and rearranging we have finally that

$$\mu_4(\bar{x}) = \frac{1}{n^3}\left[\mu_4\left(1-\frac{7(n-1)}{N-1}+\frac{12(n-1)(n-2)}{(N-1)(N-2)}-\frac{6(n-1)(n-2)(n-3)}{(N-1)(N-2)(N-3)}\right)\right.$$
$$\left.+N\mu_2^2\left(\frac{3(n-1)}{N-1}-\frac{6(n-1)(n-2)}{(N-1)(N-2)}+\frac{3(n-1)(n-2)(n-3)}{(N-1)(N-2)(N-3)}\right)\right].$$

If N is very large compared with n it is clear that

$$\mu_3(\bar{x}) \simeq \mu_3/n^2,$$

$$\mu_4(\bar{x}) \simeq \frac{\mu_4+3(n-1)\mu_2^2}{n^3},$$

i.e. $\mu_3(\bar{x})$ and $\mu_4(\bar{x})$ are very nearly the same as those obtained for the case when the population is infinite.

We shall now go on to a discussion of the first two sampling moments of the (standard deviation)^2, i.e. we shall find

$$\mathscr{E}(s^2) = \mathscr{E}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right) \quad \text{and} \quad \sigma_{s^2}^2 = \mathscr{E}(s^2-\mathscr{E}(s^2))^2.$$

First two Sampling Moments of the (Standard Deviation)^2: Population Infinite

As before it will be assumed that we consider a sample of n units randomly and independently drawn from some population of which it is known that the first four moments, $\mu_1', \mu_2, \mu_3, \mu_4$, exist. With each element of the sample we associate a random variable x_i, $(i = 1, 2, \ldots, n)$.

In the previous chapter it was shown that

$$\mathscr{E}(s^2) = \frac{n-1}{n}\mu_2,$$

that is to say, the mean value in repeated sampling of the sample (standard deviation)^2 is not equal to the population (standard deviation)^2. We shall refer to this fact later. The process of obtaining the second moment of s^2 is perhaps a little difficult as

regards enumeration but the method is the same as those used previously both in this and the last chapter. We shall expand $\sigma^2_{s^2}$,

$$\sigma^2_{s^2} = \mathscr{E}(s^2 - \mathscr{E}(s^2))^2 = \mathscr{E}(s^4) - \left(\frac{n-1}{n}\mu_2\right)^2$$

and carry out the enumeration of $\mathscr{E}(s^4)$ in three stages.

$$n^2\mathscr{E}(s^4) = \mathscr{E}\left[\left(\sum_{i=1}^{n}(x_i-\mu_1')^2\right)^2 + n^2(\bar{x}-\mu_1')^4 - 2n(\bar{x}-\mu_1')^2\sum_{i=1}^{n}(x_i-\mu_1')^2\right].$$

(i) $$\mathscr{E}\left(\sum_{i=1}^{n}(x_i-\mu_1')^2\right)^2$$

$$= \mathscr{E}\sum_{i=1}^{n}(x_i-\mu_1')^4 + 2\mathscr{E}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i-\mu_1')^2(x_j-\mu_1')^2$$

$$= n\mu_4 + n(n-1)\mu_2^2.$$

(ii) $$\mathscr{E}(n^2(\bar{x}-\mu_1')^4) = n^2\mathscr{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_1')\right]^4$$

$$= \frac{1}{n^2}\mathscr{E}\Bigg[\sum_{i=1}^{n}(x_i-\mu_1')^4 + 4\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}[(x_i-\mu_1')^3(x_j-\mu_1') + (x_i-\mu_1')(x_j-\mu_1')^3] + 6\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i-\mu_1')^2(x_j-\mu_1')^2 + 12\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{t=j+1}^{n}[(x_i-\mu_1')^2(x_j-\mu_1')(x_t-\mu_1') + (x_i-\mu_1')(x_j-\mu_1')^2(x_t-\mu_1') + (x_i-\mu_1')(x_j-\mu_1')(x_t-\mu_1')^2] + 24\sum_{i=1}^{n-3}\sum_{j=i+1}^{n-2}\sum_{t=j+1}^{n-1}\sum_{v=t+1}^{n}(x_i-\mu_1')(x_j-\mu_1')(x_t-\mu_1')(x_v-\mu_1')\Bigg].$$

Because of the independence of the x's we have then

$$\mathscr{E}(n^2(\bar{x}-\mu_1')^4) = \frac{1}{n^2}[n\mu_4 + 3n(n-1)\mu_2^2].$$

$$\text{(iii)}\quad \mathscr{E}\left(2n(\bar{x}-\mu_1')^2\sum_{i=1}^{n}(x_i-\mu_1')^2\right)$$

$$=\frac{2}{n}\mathscr{E}\left[\sum_{i=1}^{n}(x_i-\mu_1')^2\right]\left[\sum_{i=1}^{n}(x_i-\mu_1')\right]^2$$

$$=\frac{2}{n}\mathscr{E}\left[\sum_{i=1}^{n}(x_i-\mu_1')^4+2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i-\mu_1')^2(x_j-\mu_1')^2\right.$$

$$+2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}[(x_i-\mu_1')^3(x_j-\mu_1')+(x_i-\mu_1')(x_j-\mu_1')^3]$$

$$+2\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{k=j+1}^{n}[(x_i-\mu_1')^2(x_j-\mu_1')(x_k-\mu_1')$$

$$\left.+(x_i-\mu_1')(x_j-\mu_1')^2(x_k-\mu_1')+(x_i-\mu_1')(x_j-\mu_1')(x_k-\mu_1')^2]\right],$$

whence

$$\mathscr{E}\left(2n(\bar{x}-\mu_1')^2\sum_{i=1}^{n}(x_i-\mu_1')^2\right)=\frac{2}{n}[n\mu_4+n(n-1)\mu_2^2].$$

Substituting for these expressions in the expression for $\mathscr{E}(s^4)$ we have

$$\mathscr{E}(s^4)=\frac{\mu_4}{n}\left(1-\frac{1}{n}\right)^2+\mu_2^2\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}+\frac{3}{n^2}\right),$$

which gives us for $\sigma_{s^2}^2$, remembering $\beta_2=\mu_4/\mu_2^2$,

$$\sigma_{s^2}^2=\frac{\mu_2^2}{n^3}(n-1)^2\left[\beta_2-\frac{n-3}{n-1}\right].$$

Estimate of Population (Standard Deviation)^2

In the previous chapter it was noted that the sample standard deviation can only be used in describing something about the sample and that as soon as it is necessary to estimate the standard deviation in the population it is desirable to consider not s^2, but a quantity s_e^2 (say), where

$$\mathscr{E}(s_e^2)=\mathscr{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)=\sigma^2,$$

because the mean value in repeated sampling of s_e^2 is equal to σ^2.

The (standard error)^2 of s_e^2 will be, from the immediately preceding analysis

$$\sigma_{s_e^2}^2 = \frac{\mu_2^2}{n}\left[\beta_2 - \frac{n-3}{n-1}\right].$$

It will be noted for both s^2 and s_e^2 that

(1) when n becomes very large

$$\sigma_{s^2}^2 \text{ and } \sigma_{s_e^2}^2 \text{ both} \to \frac{\mu_2^2}{n}[\beta_2 - 1].$$

(2) when the parent population is normal and therefore $\beta_2 = 3$ we have that

$$\sigma_{s_e^2}^2 = \frac{\mu_2^2}{n}\left[3 - \frac{n-3}{n-1}\right]$$

and when in addition n is large

$$\sigma_{s^2}^2 \text{ and } \sigma_{s_e^2}^2 \text{ both} \to \frac{2\mu_2^2}{n}.$$

This last expression is often useful when carrying out a rough test of significance in one's head. The expressions for $\beta_1(s_e^2)$ and $\beta_2(s_e^2)$ may be calculated in the same way as we have calculated $\sigma_{s_e^2}^2$, and the student should try to work these out. It may be shown, for the parent population normally distributed and for n, the size of sample, large, that

$$\beta_1(s_e^2) \doteq \frac{8}{n-1}, \quad \beta_2(s_e^2) \doteq 3 + \frac{12}{n-1}.$$

First Two Sampling Moments of s_e: Population Infinite

For the student who is not yet accustomed to the ideas of analysis of variance it may seem a little artificial that we have treated s^2 and s_e^2 in some detail, but have made no mention of s and s_e which, being measures of scale, are collective characters which are used in the development of statistical theory from the very beginning. The reason why we do not treat of s and s_e is not far to seek; the square root sign makes them difficult to handle mathematically. We shall find here the first two moments of the sampling distribution of s_e but we shall do so in a very approximate way and rely for our justification on the fact that

the expressions obtained do agree well with numerical sampling experiments. By definition

$$s_e = \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{1}{2}}.$$

Write $$s_e^2 = \sigma^2 + \delta s_e^2,$$

i.e. $$s_e = (\sigma^2+\delta s_e^2)^{\frac{1}{2}} = \sigma\left(1+\frac{\delta s_e^2}{\sigma^2}\right)^{\frac{1}{2}},$$

and assume δs_e is small by comparison with σ, which will be so if n, the sample size, is large.

Expand the right-hand side as a series, and take expectations.

$$\mathscr{E}(s_e) = \sigma\left[1+\tfrac{1}{2}\mathscr{E}\left(\frac{\delta s_e}{\sigma}\right)^2 + \frac{\frac{1}{2}(-\frac{1}{2})}{1.2.}\mathscr{E}\left(\frac{\delta s_e}{\sigma}\right)^4 + \ldots\right].$$

The expectations within the bracket may be evaluated directly.

$$\mathscr{E}(s_e^2) = \sigma^2$$

from previous work, while from above

$$\mathscr{E}(s_e^2) = \sigma^2 + \mathscr{E}(\delta s_e^2)$$

from which it follows that

$$\mathscr{E}(\delta s_e^2) = 0.$$

Also $$\mathscr{E}(s_e^2-\sigma^2)^2 = \sigma_{s_e^2}^2 = \frac{\sigma^4}{n}(\beta_2-1) = \mathscr{E}(\delta s_e^4).$$

This is certainly only true for large n but if n is not large then the original assumption will not hold good either. Substituting for these expressions we have

$$\mathscr{E}(s_e) = \sigma\left[1-\frac{(\beta_2-1)}{8n}+\ldots\right],$$

$$\sigma_{s_e}^2 = \mathscr{E}(s_e-\mathscr{E}(s_e))^2 \doteq \sigma^2\left[1-\left(1-\frac{(\beta_2-1)}{8n}\right)^2\right] \doteq \frac{\beta_2-1}{4n}\sigma^2$$

and hence $$\sigma_{s_e} \doteq \sigma\sqrt{\left(\frac{\beta_2-1}{4n}\right)}.$$

When the parent population is normally distributed then $\beta_2 = 3$ and

$$\sigma_{s_e} \doteq \frac{\sigma}{\sqrt{(2n)}}.$$

The student must beware of an indiscriminate use of these sampling moments of s_e. Nevertheless, in spite of their mathematical limitations, they are useful if only because the distribution of s_e tends, with n increasing, to be normal more quickly than does the distribution of s_e^2.

It is known that if the original population is normally distributed then $(n-1)\,s_e^2/\sigma^2$ is distributed as χ^2 with $n-1$ degrees of freedom. We shall discuss the χ^2 distribution as an example in the theory of characteristic functions but it is useful here as an exercise to find the first two sampling moments of χ^2, which we shall do in a quite general way, for the case of a grouped frequency distribution. Some variation of our previous technique will be necessary, and we shall now make use of the concept of the characteristic random variable. This concept will be found useful in all sampling problems where it is necessary to consider groups or strata and the student should try to make himself familiar with its use. We shall assume that a sample of N observations is randomly and independently drawn from a population which may be classified into k groups. Suppose that the chance of an individual being drawn from the ith group is p_i for $i = 1, 2, \ldots, k$, that $\sum_{i=1}^{k} p_i = 1$ and that the number in the sample actually drawn from the ith group is n_i. It is obvious that

$$\sum_{i=1}^{k} n_i = N, \quad \mathscr{E}(n_i) = Np_i.$$

Write

$$\delta n_i = n_i - Np_i.$$

We shall begin by finding $\sigma^2(\delta n_i)$ and $\rho(\delta n_i\, \delta n_j)$, for

$$i, j = 1, 2, \ldots, k,$$

and for convenience we shall consider only two groups, the ith and the jth, throughout. The argument will obviously hold good for any pair. Associate with each unit of the sample of N, in order of drawing, two series of independent characteristic random variables, α_l and β_t, for $l, t = 1, 2, \ldots, N$. The series of variables α_l will have the property that they will take the value 1 when the sample unit falls into the ith group but that they will be zero otherwise. Similarly the variables β_t will take the value 1

when the sample unit falls in the jth group and be zero otherwise. It is clear from this definition that

$$\sum_{l=1}^{N} \alpha_l = n_i, \quad \sum_{t=1}^{N} \beta_t = n_j.$$

Using only the variables α_l we may find $\sigma^2(\delta n_i)$.

$$\sigma^2(\delta n_i) = \mathscr{E}(\delta n_i - \mathscr{E}(\delta n_i))^2 = \mathscr{E}(\delta n_i)^2 - (\mathscr{E}(\delta n_i))^2.$$

From definition

$$\mathscr{E}(\delta n_i) = \mathscr{E}(n_i) - Np_i = 0,$$

$$\mathscr{E}(\delta n_i)^2 = \mathscr{E}(n_i - Np_i)^2 = \mathscr{E}(n_i^2) - (Np_i)^2.$$

Again from definition

$$\mathscr{E}(n_i^2) = \mathscr{E}\left(\sum_{l=1}^{N} \alpha_l\right)^2 = \mathscr{E}\left[\sum_{l=1}^{N} \alpha_l^2 + 2 \sum_{l=1}^{N-1} \sum_{h=l+1}^{N} \alpha_l \alpha_h\right].$$

We assumed that the characteristic random variables were independent, whence the right-hand term will be the product of two expectations.

$$\mathscr{E}(\alpha_l^2) = 1 \cdot p_i + 0(1 - p_i) = p_i, \quad \mathscr{E}(\alpha_l \alpha_h) = \mathscr{E}(\alpha_l)\mathscr{E}(\alpha_h) = p_i^2$$

and therefore $$\mathscr{E}(n_i^2) = Np_i + N(N-1)p_i^2$$

and $$\sigma^2(\delta n_i) = Np_i(1 - p_i).$$

Similarly for the correlation coefficient $\rho(\delta n_i \delta n_j)$. From the definition of the correlation coefficient we have

$$\rho(\delta n_i \delta n_j) = \frac{\mathscr{E}(\delta n_i \delta n_j) - \mathscr{E}(\delta n_i)\mathscr{E}(\delta n_j)}{\sigma_{\delta n i} \sigma_{\delta n j}} = \frac{\mathscr{E}(\delta n_i \delta n_j)}{\sigma_{\delta n i} \sigma_{\delta n j}},$$

since the separate expectations are each zero. The denominator is already known and it remains to calculate $\mathscr{E}(\delta n_i \delta n_j)$.

$$\mathscr{E}(\delta n_i \delta n_j) = \mathscr{E}(n_i n_j) - N^2 p_i p_j.$$

Using now the two series of variables

$$\mathscr{E}(n_i n_j) = \mathscr{E}\left(\sum_{l=1}^{N} \alpha_l \sum_{t=1}^{N} \beta_t\right) = \mathscr{E}\left(\sum_{l=1}^{N} \alpha_l \beta_l + \sum_{l=1}^{N-1} \sum_{t=l+1}^{N} (\alpha_l \beta_t + \alpha_t \beta_l)\right).$$

From definition it is certain that the first summation must be zero, i.e.

$$\sum_{l=1}^{N} \alpha_l \beta_l = 0,$$

for when $\alpha = 1, \beta = 0$ and vice versa. Hence

$$\mathscr{E}(n_i n_j) = N(N-1)\,p_i p_j$$

and

$$\mathscr{E}(\delta n_i \delta n_j) = -Np_i p_j.$$

Substituting in the expression for $\rho(\delta n_i \delta n_j)$ it is seen that

$$\rho(\delta n_i \delta n_j) = -\sqrt{\left(\frac{p_i p_j}{(1-p_i)(1-p_j)}\right)}.$$

These preliminary calculations will serve to make the reader familiar with the way in which the concept of the characteristic random variable is used. χ^2 may be defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{\delta n_i^2}{Np_i}$$

and we must therefore consider $\mathscr{E}(\delta n_i^2)$, $\mathscr{E}(\delta n_i^2 \delta n_j^2)$ and $\mathscr{E}(\delta n_i^4)$ if we are to find the first two sampling moments of this criterion.

We have already shown that

$$\mathscr{E}(\delta n_i^2) = \sigma_{\delta ni}^2 = Np_i(1-p_i),$$

from which it follows that

$$\mathscr{E}(\chi^2) = \sum_{i=1}^{k} (1-p_i) = k-1,$$

but the expectations of the higher moments of δn_i and $\delta n_i \delta n_j$ may cause difficulty.

$$\mathscr{E}(\delta n_i^4) = \mathscr{E}(n_i - Np_i)^4 = \mathscr{E}\left(\sum_{l=1}^{N} (\alpha_l - p_i)\right)^4$$

$$= \mathscr{E}\left(\sum_{l=1}^{N} (\alpha_l - p_i)^4\right) + 6\,\mathscr{E} \sum_{l=1}^{N-1} \sum_{t=l+1}^{N} (\alpha_l - p_i)^2 (\alpha_t - p_i)^2.$$

The other cross products vanish because of the independence of the variables. The $\mathscr{E}(\alpha_l - p_i)^4$ comes immediately on expansion.

$$\mathscr{E}(\alpha_l - p_i)^4 = \mathscr{E}(\alpha_l^4) - 4p_i\,\mathscr{E}(\alpha_l^3) + 6p_i^2\,\mathscr{E}(\alpha_l^2) - 4p_i^3\,\mathscr{E}(\alpha_l) + p_i^4$$

$$= p_i - 4p_i^2 + 6p_i^3 - 3p_i^4 = p_i(1-p_i)(1-3p_i+3p_i^2),$$

and we obtain similarly $\mathscr{E}(\alpha_l - p_i)^2$. Substitution of these values in the expression for $\mathscr{E}(\delta n_i^4)$ gives

$$\mathscr{E}(\delta n_i^4) = Np_i(1-p_i)(1+3(N-2)p_i(1-p_i)).$$

The $\mathscr{E}(\delta n_i^2\,\delta n_j^2)$ can be calculated in a similar way, but it will be necessary to take care in the enumeration of terms. For the student who is not sure of himself when dealing with summation signs it is perhaps better to expand

$$\mathscr{E}(\delta n_i^2\,\delta n_j^2) = \mathscr{E}[(n_i - Np_i)^2\,(n_j - Np_j)^2]$$

and consider each term separately.

For example,

$$\mathscr{E}(n_i^2 n_j^2) = \mathscr{E}\left[\left(\sum_{l=1}^{N}\alpha_l\right)^2\left(\sum_{t=1}^{N}\beta_t\right)^2\right]$$

$$= \mathscr{E}\Bigg[\sum_{l=1}^{N}\alpha_l^2\beta_l^2 + \sum_{l=1}^{N-1}\sum_{t=l+1}^{N}(\alpha_l^2\beta_t^2 + \alpha_t^2\beta_l^2)$$

$$+ 2\sum_{l=1}^{N-1}\sum_{t=l+1}^{N}[\alpha_l\alpha_t\beta_t^2 + \alpha_l\alpha_t\beta_l^2 + \alpha_l^2\beta_l\beta_t + \alpha_t^2\beta_l\beta_t]$$

$$+ 2\sum_{l=1}^{N-2}\sum_{t=l+1}^{N-1}\sum_{h=t+1}^{N}[\alpha_l\alpha_t\beta_h^2 + \alpha_l\alpha_h\beta_t^2 + \alpha_h\alpha_t\beta^2$$
$$+ \alpha_h^2\beta_l\beta_t + \alpha_t^2\beta_l\beta_h + \alpha_t^2\beta_h\beta_t]$$

$$+ 4\sum_{l=1}^{N-1}\sum_{t=l+1}^{N}(\alpha_l\alpha_l\beta_t\beta_l)$$

$$+ 4\sum_{l=1}^{N-2}\sum_{t=l+1}^{N-1}\sum_{h=t+1}^{N}(\alpha_l\alpha_l\beta_t\beta_h + \alpha_t\alpha_l\beta_l\beta_h + \ldots \text{etc. (6 terms)})$$

$$+ \sum_{l=1}^{N-3}\sum_{t=l+1}^{N-2}\sum_{h=t+1}^{N-1}\sum_{v=h+1}^{N}[\alpha_l\alpha_l\beta_h\beta_v + \alpha_h\alpha_v\beta_t\beta_l + \ldots \text{etc. (24 terms)}]\Bigg],$$

whence

$$\mathscr{E}(n_i^2 n_j^2) = N(N-1)\,p_ip_j + N(N-1)\,(N-2)\,p_ip_j(p_i+p_j)$$
$$+ N(N-1)\,(N-2)\,(N-3)\,p_i^2p_j^2.$$

Similarly

$$\mathscr{E}(n_i^2 n_j) = N(N-1)\,p_ip_j + N(N-1)\,(N-2)\,p_i^2p_j,$$
$$\mathscr{E}(n_i n_j^2) = N(N-1)\,p_ip_j + N(N-1)\,(N-2)\,p_ip_j^2.$$

We have already evaluated $\mathscr{E}(n_in_j)$ and $\mathscr{E}(n_j^2)$, so that on substitution we have that

$$\mathscr{E}(\delta n_i^2\,\delta n_j^2) = Np_ip_j[(N-2)\,(1-p_i-p_j+3p_ip_j)+1].$$

From the definition of a (standard error)^2 it follows that

$$\sigma_{\chi^2}^2 = \mathscr{E}(\chi^2 - \mathscr{E}(\chi^2))^2 = \mathscr{E}(\chi^2)^2 - (k-1)^2.$$

Substitute for χ^2

$$\mathscr{E}(\chi^2)^2 = \mathscr{E}\left(\sum_{i=1}^{k} \frac{\delta n_i^2}{Np_i}\right)^2 = \mathscr{E}\left(\sum_{i=1}^{k} \frac{\delta n_i^4}{N^2 p_i^2} + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \frac{\delta n_i^2\, \delta n_j^2}{N^2 p_i p_j}\right).$$

The expectations of both these terms have been evaluated and it only remains therefore to substitute the values obtained and to show that

$$\sigma_{\chi^2}^2 = 2(k-1)\left(1-\frac{1}{N}\right) + \frac{1}{N}\sum_{i=1}^{k}\frac{1}{p_i} - \frac{k^2}{N}.$$

The algebra involved is not heavy but the student may perhaps get into difficulties if he does not resort to the by now familiar trick of getting rid of the double summation sign, e.g.

$$\left[\sum_{i=1}^{k}(1-p_i)\right]^2 = \sum_{i=1}^{k}(1-p_i)^2 + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}(1-p_i)(1-p_j)$$

and

$$\left[\sum_{i=1}^{k} p_i\right]^2 = \sum_{i=1}^{k} p_i^2 + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} p_i p_j,$$

remembering that

$$\sum_{i=1}^{k} p_i = 1.$$

The usual values taken for the moments of χ^2 are

$$\mathscr{E}(\chi^2) = k-1, \quad \sigma_{\chi^2}^2 = 2(k-1).$$

It will be noted that by taking these values we are neglecting terms of order $1/N$ in the expression for $\sigma_{\chi^2}^2$. Generally this approximation is not important and the student may convince himself that this is the case by working out some numerical examples.

Numerical examples

Compare the true value and the approximate values of $\sigma_{\chi^2}^2$ for the cases:

(i) $k = 10$, $N = 20$, $p_i = \frac{1}{10}$ for $i = 1, 2, \ldots, 10$.

(ii) $k = 10$, $N = 50$, $p_1 = p_{10} = 0{\cdot}02$, $p_2 = p_9 = 0{\cdot}04$, $p_3 = p_8 = 0{\cdot}08$, $p_4 = p_7 = 0{\cdot}14$, $p_5 = p_6 = 0{\cdot}22$.

REFERENCES AND READING

There are many papers both in the journal *Biometrika* and elsewhere where the sampling moments of different distributions are derived. We may mention A. E. R. Church, *Biometrika*, XVII, p. 79; J. M. le Roux, *Biometrika*, XXIII, p. 134; J. Neyman, *Biometrika*, XVII, p. 472; J. B. S. Haldane, *Biometrika*, XXXIII, p. 234, but the list is not by any means exhaustive. The student wishing further exercises on expectations should consult these and other papers and work through the algebra.

For examples in the use of the characteristic random variable the student might consult J. Neyman, *J. Amer. Statist. Assoc.* XXXIII, p. 101, 'Contribution to the theory of sampling human populations', and the appendix to F. N. David, *Statist. Res. Mem.* II, p. 69, 'Limiting distributions connected with certain methods of sampling human populations'.

CHAPTER XII

RANDOM VARIABLES. INEQUALITIES. LAWS OF LARGE NUMBERS. LEXIS THEORY

By application of the elementary theorems regarding the addition and multiplication of expectations most problems can be solved. It will have been noted that the theorems are quite general and do not depend for their application on the random variable following a particular probability law. Following along the same lines, and without specifying anything about x, other than that it is a discontinuous or continuous random variable, several inequalities have been devised which enable limits to be set for the probability of x being less than a given value. Most of these inequalities spring from, or are generated from, Markoff's lemma.

Markoff's Lemma

It is assumed that a certain random variable x may take only positive or zero values. If $a = \mathscr{E}(x)$ and t is any given number greater than unity then

$$P\{x \geqslant at^2\} \leqslant 1/t^2.$$

Let x take values in ascending order

$$0 \leqslant u_1 \leqslant u_2 \leqslant u_3 < \ldots < u_n < u_{n+1} < \ldots < u_N$$

and let $p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_N$ be the corresponding probabilities that x takes the given values. The proof may conveniently be divided into three parts.

I. $at^2 < u_1$

From the definition of expectation

$$\mathscr{E}(x) = a = \sum_{i=1}^{N} u_i p_i > u_1 \sum_{i=1}^{N} p_i = u_1.$$

Since t is an integer greater than unity, if $a > u_1$, then at^2 *a fortiori* $> u_1$, and cannot be less than u_1.

II. $at^2 > u_N$

If $at^2 > u_N$ then $P\{x \geqslant at^2\} = 0$, which is certainly less than $1/t^2$.

III. $u_1 < at^2 < u_N$

If at^2 lies between u_1 and u_N then it must be possible to find two values of x, say u_n and u_{n+1}, between which at^2 will lie. Assume therefore that

$$u_n < at^2 \leqslant u_{n+1}.$$

Writing down the expectation of x we have

$$\mathscr{E}(x) = a = u_1p_1 + u_2p_2 + \ldots + u_np_n + u_{n+1}p_{n+1} + \ldots + u_Np_N$$

$$a \geqslant u_{n+1}p_{n+1} + \ldots + u_Np_N > at^2(p_{n+1} + \ldots + p_N) = at^2P\{x \geqslant at^2\}.$$

It follows that $$P\{x \geqslant at^2\} < 1/t^2.$$

ILLUSTRATION. x is a binomial variable with expectation equal to np.

Let $t = \sqrt{3}$. Then $$P\{x \geqslant 3np\} < \tfrac{1}{3}.$$

It will be seen from this that the limit set to the probability by the inequality is not very restrictive, but it must be remembered that the lemma will apply to any random variable about which the only thing known is its mean value. If the probability law of a variable is known then there is no need to calculate the probability as given by Markoff's lemma because the exact value for any required probability can be found.

Finer limits can be obtained by the use of the Bienaymé-Tchebycheff inequality which makes use of both the mean value and the standard error but again since this inequality will be applicable to any random variable which has a mean value and a standard error too much cannot be expected of it.

BIENAYMÉ-TCHEBYCHEFF INEQUALITY

If x is a random variable of any distribution whatever, then provided $\mathscr{E}(x) = a$ and $\mathscr{E}(x-a)^2 = \sigma^2$ exist and $t > 1$

$$P\{|x-a| < t\sigma\} > 1 - 1/t^2.$$

The proof of this inequality follows directly from the Markoff lemma.

Write $$y = (x-a)^2.$$

Then $$\mathscr{E}(y) = \mathscr{E}(x-a)^2 = \sigma^2.$$

Applying Markoff's lemma it will be seen that

$$P\{y \geqslant t^2\sigma^2\} < 1/t^2 \quad \text{or} \quad P\{(x-a)^2 \geqslant t^2\sigma^2\} < 1/t^2$$

and therefore $\quad 1 - P\{|\, x-a \,| \geqslant t\sigma\} > 1 - 1/t^2.$

It is obvious that

$$P\{|\, x-a \,| \geqslant t\sigma\} + P\{|\, x-a \,| < t\sigma\} = 1$$

and the inequality

$$P\{|\, x-a \,| < t\sigma\} > 1 - 1/t^2$$

is proved.

ILLUSTRATION. For illustration let us again consider the binomial variable, x, and further, let $t = 3$.

$$\mathscr{E}(x) = np, \quad \mathscr{E}(x-\mathscr{E}(x))^2 = npq$$

and $$P\{|\, x-np \,| < 3\sqrt{(npq)}\} > \tfrac{8}{9}.$$

The Bienaymé-Tchebycheff inequality leads naturally to the mathematical Law of Large Numbers. This last is a name given to a series of theorems which although differing in their proofs do not differ radically in their conclusions.

It may be shown for a broad class of linear functions,

$$y = F(x_1, x_2, \ldots, x_n),$$

that the standard error of y tends to zero as n, the number of variables x, increases without limit. The simplest case of this will be when the x's are all independent, when the standard error of each x has the same value, σ, and when $y = \bar{x}$. Then the standard error of y is $\sigma/\sqrt{n}$ and it is seen that this tends to zero as n tends to infinity. But the standard errors of each x need not necessarily be equal. The same result can be reached by supposing simply that all the x's are bounded. That is to say that there exists a certain number m, such that $|\, x \,|$ cannot exceed m, and therefore the standard error of y cannot exceed $m/\sqrt{n}$.

Also the standard error of a linear function may tend to zero even if the x's are not all independent. This will be the case when each random variable x is correlated with the one which immediately precedes it and the one which immediately follows it, the others being independent or at least uncorrelated. Such successions of x's were considered by Markoff and called chains.

If the standard error of any function y, of the n random variables x, tends to zero as n tends to ∞, then the Law of Large Numbers will apply to the function, y.

THEOREM. (*Law of Large Numbers.*) If y is a function of n random variables, $x_1, x_2, \ldots, x_n$, and if

$$\mathscr{E}(y) = a \quad \text{and} \quad \mathscr{E}(y - \mathscr{E}(y)) = \sigma^2,$$

then provided $\sigma^2 \to 0$ as $n \to \infty$, for any two positive numbers ϵ and η, where ϵ and η may be as small as desired, it is possible to find a number n_0, such that for $n > n_0$

$$P\{|y - a| < \epsilon\} > 1 - \eta.$$

The inequality of Bienaymé-Tchebycheff applied to y will give, for any t greater than unity,

$$P\{|y - a| < \sigma t\} > 1 - 1/t^2.$$

Write $$1/t^2 = \eta$$

and the inequality becomes

$$P\{|y - a| < \sigma/\sqrt{\eta}\} > 1 - \eta.$$

Now, assuming $\sigma^2 \to 0$ as $n \to \infty$, whatever the numbers ϵ and η, where ϵ and η are as small as desired, it will be possible to find a number n_0 so large that if $n > n_0$ then

$$\sigma < \epsilon\sqrt{\eta}.$$

It follows therefore that

$$P\{|y - a| < \epsilon\} > P\{|y - a| < \sigma/\sqrt{\eta}\} > 1 - \eta$$

and the inequality is proved.

Example. m dice are thrown. If $\bar{x}$ is the mean of the sum of the dots on their upper faces find

$$P\{|\bar{x} - \tfrac{7}{2}| < \tfrac{1}{2}\} > 1 - 4\sigma_{\bar{x}}^2.$$

Let $x_i (i = 1, 2, \ldots, m)$ be the number of dots on the upper face of the ith die. Hence

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i.$$

From first principles it may be shown that

$$\mathscr{E}(x_i) = \tfrac{7}{2} \quad \text{and} \quad \mathscr{E}(\bar{x}) = \tfrac{7}{2}.$$

Also since the x's must be independent

$$\sigma_{x_i}^2 = \tfrac{35}{12} \quad \text{and} \quad \sigma_{\bar{x}}^2 = \frac{35}{12m}.$$

If $m \to \infty$ then $\sigma_{\bar{x}} \to 0$.

The probability it is required to calculate therefore is

$$P\{|\bar{x} - \tfrac{7}{2}| < \tfrac{1}{2}\} > 1 - \frac{35}{3m}$$

by an application of the Law of Large Numbers. As m increases this probability will tend to unity. It may be noted that for the probability to exist m must be greater than 11.

THEOREM. (*Generalized Tchebycheff Inequality.*) Assume that there are n independent random variables $x_1, x_2, \ldots, x_n$ and that

$$\mathscr{E}(x_i) = a_i, \quad \mathscr{E}(x_i - a_i)^2 = \sigma_i^2, \quad \text{for } i = 1, 2, \ldots, n.$$

Then, provided $t^2 < n$, where t is a number at choice

$$P\left\{|\bar{x} - \mathscr{E}(\bar{x})| < \frac{1}{t}\sqrt{\left(\sum_{i=1}^{n} \sigma_i^2/n\right)}\right\} > 1 - \frac{t^2}{n}.$$

The proof follows directly along the lines of the two preceding theorems.

THEOREM. (*Poisson's Law of Large Numbers.*) It is assumed that the probability for the success of an event varies from trial to trial. In n successive trials the successive probabilities are $p_1, p_2, \ldots, p_n$. If there be k successes in n trials then

$$P\left\{\left|\frac{k}{n} - \mathscr{E}\left(\frac{k}{n}\right)\right| < \frac{1}{t}\sqrt{\left(\sum_{i=1}^{n} p_i q_i/n\right)}\right\} > 1 - \frac{t^2}{n},$$

where t is a number at choice and $t^2 < n$.

Assume that characteristic random variables $x_1, x_2, \ldots, x_n$ are attached one to each trial. The proof of the theorem then follows from an application of the Bienaymé-Tchebycheff inequality.

The above theorems are only a few of the many which could be quoted and which all express the same conclusion; that, given a function of n random variables, the difference between the observed value of the function and its expectation will become small as n increases provided the standard error of the function tends to zero. Mathematically the conclusions cannot be queried, but the question may be raised as to whether they are of any

practical importance. It is held by some that these laws can be made to justify a given definition of probability but it is doubtful if this can be so. In no practical problem can the conditions under which the material is collected be kept constant and n can never tend to infinity. Possibly the most that can be drawn from the theorems is the reminder that the larger the sample under consideration, all other things being equal, the smaller will be the difference between the sample estimate and its expected value. We note, however, that there will always be a difference in practice.

Lexis Theory

We now leave the theorems regarding the Laws of Large Numbers and turn to the further applications of the simple theorems on expectations. It will be supposed that there is a number of independent trials, N, which may be divided into n sets of s so that $N = ns$.

If the binomial theorem on probabilities is applicable to these observations then it is necessary for the probability to be constant throughout the set of N trials and therefore throughout each of the n sets of s. Such a set of N trials is sometimes spoken of as a Bernoulli series. If the probability is not constant throughout the set of N observations then two ways in which it may vary will be considered. Suppose first that the probability varies from trial to trial within a set of s observations but that the variation is the same within each set. That is to say, if the probability for the fifth event in the first set is p_5, then this will be the probability for the fifth event in each set. This type of variation is known as Poisson.* Secondly, suppose that the probability is the same within a set of s observations but varies from set to set; the variation is then known as Lexis. The theory, commonly called Lexis theory, which we shall now develop, deals with the separation of these three types of variation.

Lexis theory is commonly applied to birth and death rates and it is not inappropriate therefore to illustrate the difference between these three types of variation in this way. The probability of death at a given age among (say) university students

* This should not be confused with Poisson's limit to the binomial.

may be assumed to be constant and it is unlikely that the estimate of probability would vary to any marked extent if a large number of students were arbitrarily divided into different sets. We should be justified in this case in assuming that the binomial (or Bernoulli) series would be valid.

Next consider a town divided into different homogeneous districts. We might assume that the probability of death at a given age would be the same for each district but that the probability of death would be different for different age-groups. This would be an example of Poisson variation.

Finally we might consider n different age groups in a single district. The probability of death at a given age may be assumed constant for s persons of the same age, but it will be different for different age groups. This would be Lexis variation.

Consider therefore N random independent variables, x, divided into n sets of s in the following way:

$$\begin{array}{llll} x_{11}, & x_{12}, & \ldots, & x_{1s} \qquad \frac{1}{s}\sum_{j=1}^{s} x_{1j} = \bar{x}_1. \\ x_{21}, & x_{22}, & \ldots, & x_{2s} \qquad \frac{1}{s}\sum_{j=1}^{s} x_{2j} = \bar{x}_2. \\ \vdots & \vdots & \vdots & \vdots \qquad \vdots \\ x_{i1}, & x_{i2}, & \ldots, & x_{is} \qquad \frac{1}{s}\sum_{j=1}^{s} x_{ij} = \bar{x}_i. \\ \vdots & \vdots & \vdots & \vdots \qquad \vdots \\ x_{n1}, & x_{n2}, & \ldots, & x_{ns} \qquad \frac{1}{s}\sum_{j=1}^{s} x_{nj} = \bar{x}_n. \end{array}$$

If x_{ij} is the jth variable in the ith set, let

$$\mathscr{E}(x_{ij}) = a_{ij}, \quad \mathscr{E}(x_{ij}-a_{ij})^2 = \sigma_{ij}^2,$$
$$\mathscr{E}(\bar{x}_i) = \bar{a}_i, \quad \text{and} \quad \mathscr{E}(\bar{x}_i-\bar{a}_i)^2 = \sigma_i^2$$
$$\text{for } i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, s.$$

We have shown previously that

$$\mathscr{E}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right) = \frac{n-1}{n}\sum_{i=1}^{n}\sigma_i^2 + \sum_{i=1}^{n}(a_i-\bar{a})^2,$$

where a_i is the expectation of x_i and σ_i its standard error.

Applying this theorem to the present variables we have

$$\mathscr{E}\left(\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2\right)=\frac{s-1}{s}\sum_{j=1}^{s}\sigma_{ij}^2+\sum_{j=1}^{s}(a_{ij}-a_i)^2$$

and summing for each set

$$\mathscr{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2\right)=\frac{s-1}{s}\sum_{i=1}^{n}\sum_{j=1}^{s}\sigma_{ij}^2+\sum_{i-1}^{n}\sum_{j-1}^{s}(a_{ij}-\bar{a}_i)^2.$$

Again applying the theorem we have

$$\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right)=\frac{n-1}{n}\sum_{i=1}^{n}\sigma_i^2+\sum_{i=1}^{n}(\bar{a}_i-\bar{a})^2,$$

where $\bar{x}$ is the mean of all the observations and $\bar{a}$ is its expectation. We now proceed to establish a relation between σ_i^2 and σ_{ij}^2.

$$\sigma_i^2=\mathscr{E}(\bar{x}_i-\mathscr{E}(\bar{x}_i))^2=\frac{1}{s^2}\mathscr{E}\left(\sum_{j-1}^{s}(x_{ij}-a_{ij})\right)^2$$

$$=\frac{1}{s^2}\mathscr{E}\sum_{j-1}^{s}(x_{ij}-a_{ij})^2=\frac{1}{s^2}\sum_{j-1}^{s}\sigma_{ij}^2.$$

The two fundamental equations can be combined by means of this relationship to give a single equation; eliminating σ_i^2 and σ_{ij}^2,

$$\frac{1}{s(s-1)}\left[\mathscr{E}\left(\sum_{i-1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2\right)-\sum_{i=1}^{n}\sum_{j=1}^{s}(a_{ij}-\bar{a}_i)^2\right]$$

$$=\frac{n}{n-1}\left[\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right)-\sum_{i=1}^{n}(\bar{a}_i-\bar{a})^2\right].$$

It remains to make an approximation, and replace expectations by observed values. This can only be an approximation but provided n and s are both large it will probably be adequate. In any case, since we are aiming at applying the theorem to observations, it will be the best that can be done.

We now distinguish between three types of variation.

I. Bernoulli

For the Bernoulli law of variation to hold

$$a_{ij}=\bar{a}_i=\bar{a}\quad\text{for } i=1,2,\ldots,n,\quad\text{and}\quad j=1,2,\ldots,s,$$

that is the expectations of the random variables in each set are

equal to one another and are also equal to the expectations of the random variables in any other set. It follows that

$$\frac{1}{s(s-1)}\mathscr{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2\right)=\frac{n}{n-1}\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right).$$

Hence if calculations are made on n sets of s and it is shown that

$$\frac{1}{s(s-1)}\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2=\frac{n}{n-1}\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2,$$

then we should be justified in assuming the Bernoulli (or binomial) law of variation. When this equation holds it is said that there is normal dispersion.

II. Poisson

For the Poisson law of variation to hold

$$\bar{a}_i=a_{ij} \quad \text{but} \quad a_i=\bar{a},$$

whence

$$\frac{1}{s(s-1)}\left[\mathscr{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2\right)-\sum_{i=1}^{n}\sum_{j=1}^{s}(a_{ij}-\bar{a}_i)^2\right]$$
$$=\frac{n}{n-1}\mathscr{E}\left[\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right]$$

and therefore

$$\frac{1}{s(s-1)}\mathscr{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2\right)>\frac{n}{n-1}\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right).$$

If from the observations it is found that

$$\frac{1}{s(s-1)}\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2>\frac{n}{n-1}\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2$$

then it would be justifiable to assume Poisson variation and we should say that there was sub-normal dispersion.

III. Lexis

For the Lexis law of variation to hold

$$a_{ij}=\bar{a}_i \quad \text{but} \quad \bar{a}_i\neq\bar{a}.$$

A similar reasoning to the above will give that

$$\frac{1}{s(s-1)}\left[\mathscr{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2\right)\right]<\frac{n}{n-1}\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right)$$

and if this was found to be true when expectations are replaced by observed values the dispersion would be said to be super-normal.

It is customary to calculate what is known as the Lexis Ratio on probabilities. This ratio may be found simply by assuming that each of the x's of the original set-up are characteristic random variable, capable therefore of taking the values 0 or 1 only. Let the probability that x_{ij} takes the value 1 be p_{ij}. We begin with the equation

$$\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right)=\frac{n-1}{ns^2}\sum_{i=1}^{n}\sum_{j=1}^{s}\sigma_{ij}^2+\sum_{i=1}^{n}(\bar{a}_i-\bar{a})^2,$$

which follows directly from the immediately preceding analysis.

$$\mathscr{E}(x_{ij})=a_{ij}=p_{ij}.$$

Let $$\mathscr{E}(\bar{x}_i)=\bar{a}_i=p_i \quad \text{and} \quad \mathscr{E}(\bar{x})=\bar{a}=p,$$

from which it follows that

$$\mathscr{E}(x_{ij}-a_{ij})^2=\sigma_{ij}^2=p_{ij}-p_{ij}^2, \quad \mathscr{E}(\bar{x}_i-\bar{a}_i)^2=\sigma_i^2=p_i-p_i^2.$$

Substitution in the equation gives

$$\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right)=\frac{n-1}{ns^2}\sum_{i=1}^{n}\sum_{j=1}^{s}(p_{ij}-p_{ij}^2)+\sum_{i=1}^{n}(p_i-p)^2.$$

Now $$\sum_{j=1}^{s}(p_{ij}-p_i)^2=\sum_{j=1}^{s}p_{ij}^2-sp_i^2$$

and therefore $$\sum_{j=1}^{s}p_{ij}^2=\sum_{j=1}^{s}(p_{ij}-p_i)^2+sp_i^2.$$

Similarly $$\sum_{i=1}^{n}p_i^2=\sum_{i=1}^{n}(p_i-p)^2+np^2.$$

Again substituting, in turn, in the equation we have

$$\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right)=\frac{(n-1)\,p(1-p)}{s}-\frac{n-1}{ns^2}\sum_{i=1}^{n}\sum_{j=1}^{s}(p_{ij}-p_i)^2$$
$$+\frac{ns-n+1}{ns}\sum_{i=1}^{n}(p_i-p)^2.$$

Finally, if we write

$$\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right)=(n-1)\,\sigma^2$$

the equation becomes

$$\sigma^2 = \frac{p(1-p)}{s} - \frac{1}{ns^2}\sum_{i=1}^{n}\sum_{j=1}^{s}(p_{ij}-p_i)^2 + \frac{ns-n+1}{ns(n-1)}\sum_{i=1}^{n}(p_i-p)^2.$$

If the probability is constant throughout the trials, as in the case of binomial probabilities, then we have for the (standard error)^2 of the probability $\sigma^2 = pq/s$

which is familiar. For Poisson σ^2 will be less than pq/s while for Lexis it will be greater.

If σ' is the actual standard error estimated from the observations and if $\sigma_B = \sqrt{(pq/s)}$ is the standard error of the probability assuming it is constant from trial to trial then the Lexis Ratio L is defined as $L = \sigma'/\sigma_B$.

It will be seen that if L is unity then the probability may be assumed to be constant throughout the observations. If $L<1$ it may be assumed that the probability is varying within the set but varies in the same way from set to set. If $L>1$ the probability may be assumed constant within the set but variable from set to set. We have at present no means of judging the significance of the departure of L from unity, although the student will realize at a later stage that the χ^2 distribution may be used.

Example. Rietz gives the following example of the death rates of white infants under one year of age in the U.S.A.

State	Births	Deaths per 1000
California	50,707	70
Indiana	57,915	78
Kentucky	53,658	77
Minnesota	51,452	66
N. Carolina	51,832	74
Wisconsin	54,472	79
Mean	53,339	74

The numbers of births in each state are approximately equal, and the application of Lexis theory would seem to be appropriate. The average number may be taken equal to s, the supposed number in each set, and the standard error assuming the probability constant will be

$$\sigma_B = \sqrt{\left(\frac{0{\cdot}074 \times 0{\cdot}926}{53{,}339}\right)} = 1{\cdot}13 \text{ (per thousand).}$$

The standard error estimated from the observations is 5·14, and the Lexis Ratio is therefore

$$L = \frac{5{\cdot}14}{1{\cdot}13} = 4{\cdot}5.$$

This is considerably different from unity and we may draw the inference that it is likely that the infant mortality rate is significantly different from State to State.

Exercise. The death rate in Germany per 1000 inhabitants is given for the years 1877–86 in the table below. Assume that 45,000,000 was the size of the population of Germany within the period 1877–86 and study the dispersion within the table. (B.Sc. London 1938.)

Year	1877	1878	1879	1880	1881	1882	1883	1884	1885	1886
Death-rate	28·0	27·8	27·2	27·5	26·9	27·2	27·3	27·4	27·2	27·6

Exercise. The proportions of males born in Vienna during the years 1908 and 1909 are given below:

Proportion of male births

Month ...	Jan.	Feb.	Mar.	Apr.	May	June
1908	0·522	0·513	0·514	0·525	0·513	0·514
1909	0·514	0·509	0·599	0·510	0·514	0·509
Month ...	July	Aug.	Sept.	Oct.	Nov.	Dec.
1908	0·519	0·521	0·511	0·520	0·512	0·514
1909	0·513	0·528	0·518	0·513	0·518	0·503

Assume that the number of births in Vienna was 3903 for each of these 24 months and study the dispersion within the table. (B.Sc. London 1938.)

There is one aspect of the analysis of this type of variation which might be mentioned. It will be convenient to refer back to the original scheme for N random independent variables, the fundamental equation for which was shown to be

$$\frac{1}{s(s-1)}\left[\mathscr{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2\right)-\sum_{i=1}^{n}\sum_{j=1}^{s}(a_{ij}-\bar{a}_i)^2\right]$$

$$=\frac{n}{n-1}\left[\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right)-\sum_{i=1}^{n}(\bar{a}_i-\bar{a})^2\right].$$

If the x's are regarded as units which have been randomly and independently drawn from the same population then

$$a_{ij} = \bar{a}_i = \bar{a}$$

and $$\frac{1}{s(s-1)}\mathscr{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2\right) = \frac{n}{n-1}\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right)$$

or $$\frac{1}{n(s-1)}\mathscr{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{s}(\bar{x}_{ij}-\bar{x}_i)^2\right) = \frac{s}{n-1}\mathscr{E}\left(\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2\right).$$

Since the N units are assumed homogeneous it will be clear that each side of this equation represents an estimate of the total variance in the population. If we write V for this total variance, then it will be recognized that without dividing the material into sets

$$V = \frac{1}{ns-1}\mathscr{E}\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x})^2$$

and that this V will also be equal to either side of the last equation. It follows therefore that if the material is homogeneous the following estimates of variance will all have the same expectation (replacing observed values for expectations in the equation),

$$V = \frac{1}{ns-1}\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x})^2. \qquad V_{i\bar{x}} = \frac{s}{n-1}\sum_{i=1}^{n}(\bar{x}_i-\bar{x})^2.$$

$$V_i = \frac{1}{n(s-1)}\sum_{i=1}^{n}\sum_{j=1}^{s}(x_{ij}-\bar{x}_i)^2. \qquad V_{j\bar{x}} = \frac{n}{s-1}\sum_{j=1}^{s}(\bar{x}_j-\bar{x})^2.$$

$$V_j = \frac{1}{s(n-1)}\sum_{j=1}^{s}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2.$$

Now $V_{i\bar{x}}$ measures the variation between the arithmetic means of the sets, and V_i the total variation within the sets. It follows that if the quantity $V_{i\bar{x}}/V_i$ is calculated on actual material we should obtain some idea of the variation between sets as opposed to the variation within sets. An exact test of significance is available for this, if the further assumption is made that the random variables are normally distributed.

REFERENCES AND READING

For further reading and exercises on the inequalities and the laws of large numbers the reader cannot do better than consult J. V. Uspensky, *Introduction to Mathematical Probability*, where these subjects are treated in some detail.

Lexis theory is developed fully in J. L. Coolidge, *Introduction to Mathematical Probability*.

From the statistical point of view the clearest exposition of Lexis theory will be found in the writings of H. L. Rietz and we may refer the student to Chapter VI of his book, *Mathematical Statistics*.

CHAPTER XIII

SIMPLE ESTIMATION. MARKOFF THEOREM ON LEAST SQUARES

Much of modern statistical technique is directed towards drawing valid inferences from a sample about the population from which that sample was drawn. In the early days of the subject the samples drawn were so large that the collective characters, such as the mean and standard deviation, of the sample could justly be inferred to be adequate estimates of the collective characters in the population. With the exploitation of small samples it was recognized that the sample collective characters need not necessarily be the best estimates of the population collective characters, and it became necessary to lay down certain rules which it is considered a collective character calculated from the sample must obey in order to be considered a true estimate of a collective character in the population.

Possibly no branch of statistical technique has been the subject of more controversy than the theory of estimation. We do not propose to enter into the details of this controversy and shall restrict ourselves to a statement of first principles.* These we shall formalize in the following way. It will be assumed that there is a collective character θ, of a population π, which it is desired to estimate. n drawings are made randomly and independently from π resulting in a number of observations $x_1, x_2, \ldots, x_n$.

DEFINITION. A function $F(x_1, x_2, \ldots, x_n)$ is an unbiased estimate of θ if, whatever the properties of the population π,

$$\mathscr{E}[F(x_1, x_2, \ldots, x_n)] \equiv \theta.$$

As an example of an unbiased estimate it will be remembered that the expectation of the sample mean, that is the mean value of the means in repeated sampling, is equal to the population mean.

* I do not wish to be dogmatic in any way and am prepared to admit that anyone may lay down any principles he chooses.

An example of a biased estimate is found in

$$\mathscr{E}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)=\frac{n-1}{n}\sigma^2.$$

The mean value of the sample variance in repeated sampling is not equal to the population variance. Thus the sample variance used as an estimate of the population variance will tend to underestimate it, the bias in this case being $-\sigma^2/n$.

In general there will be a large number of functions F which will satisfy the relationship

$$\mathscr{E}(F)\equiv\theta$$

and a further rule is necessary in order to choose between them.

DEFINITION. A function $F(x_1, x_2, \ldots, x_n)$ is the 'best' unbiased estimate of θ if, whatever the properties of the population, π, it satisfies the relation

$$\mathscr{E}(F)\equiv\theta$$

and $\mathscr{E}(F-\theta)^2$ is a minimum.

Example. The best *linear* unbiased estimate of the mean of the population, π, given a sample of n individuals randomly and independently drawn from π, is the sample mean. Let the population mean be ξ, and to the n sample values attach random variables

$$x_1, \quad x_2, \quad \ldots, \quad x_n.$$

We shall consider a linear function of these x's, say

$$F = a_1x_1+a_2x_2+\ldots+a_nx_n,$$

and find the conditions that the a's must satisfy in order that F shall be a best linear unbiased estimate of ξ.

Condition I. $\mathscr{E}(F)\equiv 0$.

$$\mathscr{E}(F)=\mathscr{E}\sum_{i=1}^{n}a_ix_i=\sum_{i=1}^{n}a_i\mathscr{E}(x_i)=\xi\sum_{i=1}^{n}a_i\equiv\xi,$$

and it follows that

$$\sum_{i=1}^{n}a_i=1.$$

Condition II. $\mathscr{E}(F-\theta)^2$ is a minimum subject to $\mathscr{E}(F)\equiv\theta$.

$$\mathscr{E}(F-\theta)^2=\sigma^2\sum_{i=1}^{n}a_i^2,$$

where σ is the population standard deviation.

It is necessary for $\sigma^2 \sum_{i=1}^{n} a_i^2$ to be a minimum, subject to the restriction that $\sum_{i=1}^{n} a_i = 1$, and we now find the a's to satisfy these two conditions.

Let α be Lagrange's undetermined multiplier and construct the function

$$\phi = \sigma^2 \sum_{i=1}^{n} a_i^2 - 2\alpha \sum_{i=1}^{n} a_i.$$

$$\frac{\partial \phi}{\partial a_i} = 2\sigma^2 a_i - 2\alpha = 0 \quad \text{and} \quad \alpha = \sigma^2 a_i \quad \text{for } i = 1, 2, \ldots, n.$$

Summing for all values of a_i, it is seen that $n\alpha = \sigma^2$ and therefore that

$$a_i = 1/n \quad \text{for } i = 1, 2, \ldots, n.$$

This is true for any value of i and the best linear unbiased estimate of the population mean will be

$$F = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}.$$

Similarly, although the process involves somewhat lengthy algebra, it may be shown that $\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is the best quadratic unbiased estimate of σ^2.

Example. It is given that x_1, x_2, x_3 are three random variables and that

$$\mathscr{E}(x_1) = \xi, \quad \mathscr{E}(x_2) = \xi - 2, \quad \mathscr{E}(x_3) = \xi + 1.$$

$$\mathscr{E}(x_1 - \mathscr{E}(x_1))^2 = 4, \quad \mathscr{E}(x_2 - \mathscr{E}(x_2))^2 = 1, \quad \mathscr{E}(x_3 - \mathscr{E}(x_3))^2 = 0{\cdot}01.$$

Further, x_1 is independent of x_2 and x_3, but x_2 and x_3 are correlated and have a correlation coefficient equal to $-0{\cdot}25$. Deduce a formula for the best linear unbiased estimate of ξ and calculate its standard error.

Let F be a linear function

$$F = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4.$$

We shall deduce the values of the a's necessary to fulfil the given conditions. First

$$\mathscr{E}(F) \equiv \xi \equiv a_1 \mathscr{E}(x_1) + a_2 \mathscr{E}(x_2) + a_3 \mathscr{E}(x_3) + a_4,$$

whence on substituting the given values for the expectations we find two relations,

$$a_1+a_2+a_3 = 1, \quad -2a_2+a_3+a_4 = 0.$$

Next

$$\mathscr{E}(F-\mathscr{E}(F))^2 = a_1^2\mathscr{E}(x_1-\mathscr{E}(x_1))^2+a_2^2\mathscr{E}(x_2-\mathscr{E}(x_2))^2 + a_3^2\mathscr{E}(x_3-\mathscr{E}(x_3))^2+2a_2a_3\mathscr{E}(x_2-\mathscr{E}(x_2))(x_3-\mathscr{E}(x_3)).$$

The other cross-products vanish because x_1 is given independent of x_2 and x_3. Hence

$$\mathscr{E}(F-\mathscr{E}(F))^2 = a_1^2\sigma_1^2+a_2^2\sigma_2^2+a_3^2\sigma_3^2+2a_2a_3\rho_{23}\sigma_2\sigma_3.$$

Writing $a_1 = 1-(a_2+a_3)$, and differentiating partially with respect to a_2 and a_3 we obtain, after substitution of the given values,

$$a_2 = 0{\cdot}033, \quad a_3 = 0{\cdot}965 \quad \text{and} \quad a_1 = 0{\cdot}002,$$

and from the second relation of the a's therefore that

$$a_4 = -0{\cdot}899.$$

σ_F^2 is found to be 0·0103 and the best linear unbiased estimate of ξ is

$$F = 0{\cdot}002x_1+0{\cdot}033x_2+0{\cdot}965x_3-0{\cdot}899.$$

It is interesting to note how the standard error of a given x affects the size of the coefficient of x as determined by this method. x_3 has a very much smaller standard error than either x_1 or x_2, and consequently that variable plays by far the largest part in determining F.

This estimate F is one which would not be easy to determine *a priori* on intuitive grounds. If no attention is paid to the standard errors of the x's it might have seemed natural to take

$$a_1 = a_2 = a_3 = a_4 = \tfrac{1}{3}.$$

If the a's had been so chosen, then

$$\mathscr{E}(F) = \tfrac{1}{3}\mathscr{E}(x_1+x_2+x_3+1) = \tfrac{1}{3}(\xi+\xi-2+\xi+1+1) = \xi$$

and F would be an unbiased estimate of ξ. Suppose we now examine its (standard error)^2.

$$\sigma_F^2 = \mathscr{E}(F-\mathscr{E}(F))^2 = \tfrac{1}{9}(\sigma_1^2+\sigma_2^2+\sigma_3^2+2\rho_{23}\sigma_2\sigma_3) = 0{\cdot}556.$$

The standard error of this unbiased estimate is therefore some seven times greater than that estimate which we defined as the 'best'.

The general theory of estimation covers the estimation of any collective character in a population, π, from the evidence of the sample. This is, however, rather a wide field and we shall therefore consider further only the estimation of best linear unbiased estimates from observations which are randomly and independently drawn from a given population or series of populations. That is to say we shall consider only the case where

$$F(x_1, x_2, \ldots, x_n)$$

is a linear function of the x's, and where the x's themselves may be considered as random independent variables. Functions of this type are easily determined by an application of a generalization of the theorem on least squares usually attributed to Markoff. The theorem will be stated for s parameters, but because the proof is rather long and has already been set out fully elsewhere, we shall prove it for the case of two parameters only. This last was the case considered by Markoff.

Generalized Markoff Theorem on Least Squares

Consider n populations $\pi_1, \pi_2, \ldots, \pi_i, \ldots, \pi_n$. Out of each population an individual is randomly drawn and some given character measured. Suppose that on the individual drawn from the ith population, π_i, the measured character is x_i $(i = 1, 2, \ldots, n)$. It is required to estimate θ, where θ is some collective character of the n populations π_i.

If (i) $x_1, x_2, \ldots, x_n$ are independent,

(ii) the expectation of each x_i $(i = 1, 2, \ldots, n)$ is known to be a linear function of $s \leqslant n$ unknown parameters, $p_j (j = 1, 2, \ldots, s)$, but known coefficients, a_{ij}, i.e.

$$\mathscr{E}(x_i) = a_{i1}p_1 + a_{i2}p_2 + \ldots + a_{is}p_s,$$

(iii) out of the n equations, $\mathscr{E}(x_i)$ $(i = 1, 2, \ldots, n)$, it is possible to select at least one system of s equations which is soluble with respect to the p's, that is if at least one of the determinants of the sth order of the matrix M is different from zero, where

$$M = \begin{vmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1s} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{ns} \end{vmatrix},$$

(iv) the standard error σ_i of x_i is known to satisfy the relation

$$\sigma_i^2 = \frac{\sigma^2}{P_i} \quad \text{for } i = 1, 2, \ldots, n,$$

where σ may be unknown but the P's must be known positive constants,

then it may be shown that

(1) the best unbiased estimate of any linear function of the p's,

$$\theta = b_1 p_1 + b_2 p_2 + \ldots + b_s p_s,$$

where the b's are known, is obtained by substituting in the expression for θ instead of p_j $(j = 1, 2, \ldots, s)$, the values

$$q_j^0 \quad (j = 1, 2, \ldots, s),$$

obtained by minimizing the sum of squares

$$S = \sum_{i=1}^{n} (x_i - a_{i1} q_1 - a_{i2} q_2 - \ldots - a_{is} q_s)^2 P_i$$

with regard to the q's considered as independent variables. That is to say the best linear unbiased estimate of θ will be

$$F = b_1 q_1^0 + b_2 q_2^0 + \ldots + b_s q_s^0.$$

(2) the estimate of the (standard error)^2 of F is given by the expression

$$\mu_F^2 = \frac{S_0}{n-s} \sum_{i=1}^{n} \frac{\lambda_i^2}{P_i},$$

where S_0 is the minimum value of S, i.e.

$$S_0 = \sum_{i=1}^{n} (x_i - a_{i1} q_1^0 - a_{i2} q_2^0 - \ldots - a_{is} q_s^0)^2 P_i$$

and λ_i is the coefficient of x_i in

$$F = b_1 q_1^0 + b_2 q_2^0 + \ldots + b_s q_s^0.$$

Stated fully in this way the Markoff theorem might appear cumbersome to use. Actually it is not and it will be found to be of great practical utility. The proof of the theorem for the case of two parameters is relatively easy.

Proof of Markoff Theorem for Two Parameters

θ is required to be a linear function of two parameters, say generally

$$\theta = b_0 + b_1 p_1 + b_2 p_2,$$

but since the b's are assumed known we may write

$$\theta = b_1 p_1 + b_2 p_2$$

by suitable modification of the p's. It is required to show first of all that

$$F = b_1 q_1^0 + b_2 q_2^0$$

is the best linear unbiased estimate of θ, where the q^0's are as defined above. Let

$$F = \sum_{i=1}^{n} \lambda_i x_i.$$

Then

$$\mathscr{E}(F) = \mathscr{E} \sum_{i=1}^{n} \lambda_i x_i = \sum_{i=1}^{n} \lambda_i \mathscr{E}(x_i)$$

$$= \sum_{i=1}^{n} \lambda_i (a_{i1} p_1 + a_{i2} p_2) \equiv b_1 p_1 + b_2 p_2 = \theta.$$

A restriction on the λ's will be therefore that

$$b_1 = \sum_{i=1}^{n} \lambda_i a_{i1}, \quad b_2 = \sum_{i=1}^{n} \lambda_i a_{i2}.$$

The fundamental sum of squares, S, is

$$S = \sum_{i=1}^{n} (x_i - a_{i1} q_1 - a_{i2} q_2)^2 P_i.$$

It is necessary to minimize this with respect to the q's in order to obtain q_1^0 and q_2^0.

Differentiating S partially with respect to q_1 and q_2 we obtain the equations

$$\sum_{i=1}^{n} a_{i1} x_i P_i = q_1^0 \sum_{i=1}^{n} a_{i1}^2 P_i + q_2^0 \sum_{i=1}^{n} a_{i1} a_{i2} P_i,$$

$$\sum_{i=1}^{n} a_{i2} x_i P_i = q_1^0 \sum_{i=1}^{n} a_{i1} a_{i2} P_i + q_2^0 \sum_{i=1}^{n} a_{i2}^2 P_i,$$

whence
$$q_1^0 = \frac{\sum_{i=1}^{n} a_{i1}a_{i2}P_i \sum_{i=1}^{n} a_{i2}x_iP_i - \sum_{i=1}^{n} a_{i2}^2P_i \sum_{i=1}^{n} a_{i1}x_iP_i}{\left(\sum_{i=1}^{n} a_{i1}a_{i2}P_i\right)^2 - \sum_{i=1}^{n} a_{i1}^2P_i \sum_{i=1}^{n} a_{i2}^2P_i}$$

and
$$q_2^0 = \frac{\sum_{i=1}^{n} a_{i1}a_{i2}P_i \sum_{i=1}^{n} a_{i1}x_iP_i - \sum_{i=1}^{n} a_{i1}^2P_i \sum_{i=1}^{n} a_{i2}x_iP_i}{\left(\sum_{i=1}^{n} a_{i1}a_{i2}P_i\right)^2 - \sum_{i=1}^{n} a_{i1}^2P_i \sum_{i=1}^{n} a_{i2}^2P_i}.$$

We have now to show from a consideration of the λ's that

$$F = b_1q_1^0 + b_2q_2^0,$$

where q_1^0 and q_2^0 have the same values as above.

We may do this in the following way. Since the x's are independent
$$\mathscr{E}(F - \mathscr{E}(F))^2 = \sigma_F^2 = \sum_{i=1}^{n} \sigma^2 \frac{\lambda_i^2}{P_i}.$$

In order that F shall be a best linear unbiased estimate of θ, σ_F^2 must be a minimum, subject to the restrictions found above, which were
$$b_1 = \sum_{i=1}^{n} \lambda_i a_{i1}, \quad b_2 = \sum_{i=1}^{n} \lambda_i a_{i2}.$$

σ^2 is constant and it is sufficient therefore to construct a function

$$\phi = \sum_{i=1}^{n} \frac{\lambda_i^2}{P_i} - 2\alpha_1 \sum_{i=1}^{n} \lambda_i a_{i1} - 2\alpha_2 \sum_{i=1}^{n} \lambda_i a_{i2},$$

where α_1 and α_2 are Lagrange's undetermined multipliers. Differentiating partially with respect to λ_i, equating to zero and solving for λ_i, we have, after eliminating λ_i from the restrictions, the equations

$$b_1 = \alpha_1 \sum_{i=1}^{n} a_{i1}^2P_i + \alpha_2 \sum_{i=1}^{n} a_{i1}a_{i2}P_i,$$

$$b_2 = \alpha_1 \sum_{i=1}^{n} a_{i1}a_{i2}P_i + \alpha_2 \sum_{i=1}^{n} a_{i2}^2P_i.$$

Solving for α_1 and α_2, substituting back in the equation for λ_i, and finally putting the value for λ_i in the expression

$$F = \sum_{i=1}^{n} \lambda_i x_i,$$

we find that

$$F = b_1 \frac{\sum_{i=1}^{n} a_{i1}a_{i2}P_i \sum_{i=1}^{n} a_{i2}x_iP_i - \sum_{i=1}^{n} a_{i2}^2 P_i \sum_{i=1}^{n} a_{i1}x_iP_i}{\left(\sum_{i=1}^{n} a_{i1}a_{i2}P_i\right)^2 - \sum_{i=1}^{n} a_{i1}^2 P_i \sum_{i=1}^{n} a_{i2}^2 P_i}$$

$$+ b_2 \frac{\sum_{i=1}^{n} a_{i1}a_{i2}P_i \sum_{i=1}^{n} a_{i1}x_iP_i - \sum_{i=1}^{n} a_{i1}^2 P_i \sum_{i=1}^{n} a_{i2}x_iP_i}{\left(\sum_{i=1}^{n} a_{i1}a_{i2}P_i\right)^2 - \sum_{i=1}^{n} a_{i1}^2 P_i \sum_{i=1}^{n} a_{i2}^2 P_i}.$$

The multipliers of the b's will be recognized as the quantities which we have shown to be equal to q_1^0 and q_2^0. It follows that

$$F = b_1 q_1^0 + b_2 q_2^0$$

and the first part of the theorem is proved.

It remains to show that

$$\mu_F^2 = \text{estimate of } \sigma_F^2 = \frac{S_0}{n-2} \sum_{i=1}^{n} \frac{\lambda_i^2}{P_i}.$$

From first principles we have shown that

$$\mathscr{E}(\mu_F^2) = \sigma_F^2 = \sigma^2 \sum_{i=1}^{n} \frac{\lambda_i^2}{P_i}$$

and it will be sufficient therefore to show that

$$(n-2)\,\sigma^2 = \mathscr{E}(S_0).$$

S_0 is defined as

$$S_0 = \sum_{i=1}^{n} (x_i - a_{i1}q_1^0 - a_{i2}q_2^0)^2 P_i.$$

This may be rearranged as

$$S_0 = \sum_{i=1}^{n} ((x_i - a_{i1}p_1 - a_{i2}p_2) - a_{i1}(q_1^0 - p_1) - a_{i2}(q_2^0 - p_2))^2 P_i.$$

Expanding the right-hand side, and taking expectations on both sides, while remembering that

$$\mathscr{E}(q_1^0) = p_1, \quad \mathscr{E}(q_2^0) = p_2$$

a straightforward but lengthy algebraic process reduces to

$$\mathscr{E}(S_0) = n\sigma^2 - \sigma^2 - \sigma^2 = (n-2)\,\sigma^2$$

and the second part of the theorem is proved.

Application of Markoff Theorem to Linear Regression

A random variable y is known to be correlated with another variable x and the regression of y on x may be assumed to be linear. It may further be assumed that the standard error of y for a given x is constant, and does not depend on the value of x. Values of x, n in number, are selected systematically beforehand, and the value of y taken at any of these x's is assumed to be independent of the value of y taken at any other of the x's.

In this case the y's take the place of the x's of the theorem. The conditions of the theorem are satisfied and it is required to find the best linear unbiased estimate of

$$\theta = Y(X) = p_1 + p_2 X,$$

where p_1 and p_2 are unknown parameters and X is the population mean.

The statement that the regression of y on x is linear is equivalent to writing

$$\mathscr{E}(y_i) = p_1 + p_2 x_i \quad i = 1, 2, \ldots, n.$$

Referring back to the proof of the theorem and noting that $P_i = 1$, $b_1 = 1$, $b_2 = X$, $a_{i1} = 1$, $a_{i2} = x_i$, and that y_i is the x_i of the theorem it will be seen that

$$F = \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i}{\left(\sum_{i=1}^{n} x_i\right)^2 - n \sum_{i=1}^{n} x_i^2} + X \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - n \sum_{i=1}^{n} x_i y_i}{\left(\sum_{i=1}^{n} x_i\right)^2 - n \sum_{i=1}^{n} x_i^2},$$

or, if $\bar{x}$, $\bar{y}$, s_x, s_y, and r_{xy} have their usual meanings, we may write

$$F = \bar{y} + r_{xy} \frac{s_y}{s_x} (X - \bar{x}),$$

which is the familiar formula for the regression of y on x. The (standard error)^2 of F may be obtained in the same way.

$$\mu_F^2 = \frac{S_0}{n-2} \sum_{i=1}^{n} \frac{\lambda_i^2}{P_i}.$$

Substituting for q_1^0 and q_2^0 in S_0 it may be shown that

$$S_0 = ns_y^2(1-r_{xy}^2)$$

and

$$\sum_{i=1}^{n} \lambda_i^2 = \frac{1}{n}\left[1+\frac{(X-\bar{x})^2}{s_x^2}\right],$$

whence

$$\mu_F^2 = \frac{s_y^2(1-r_{xy}^2)}{n-2}\left[1+\frac{(X-\bar{x})^2}{s_x^2}\right].$$

It has been stated that the x's are at choice. In order therefore to make μ_F^2 as small as possible two courses are open: first, the mean of the sample chosen may be made exactly equal to X, or, secondly, s_x^2 may be made as large as possible. The first course is not often practicable. If all the x's are chosen in a cluster about X, then, while $X-\bar{x}$ will be small (it will rarely be possible to make it exactly zero) so also will s_x, and the resulting ratio may be large. It is possible, however, to carry out the second course by choosing pairs of values at the same distance (approximately) on either side of X but as far away as possible. In this way $\bar{x}$ will be close to X but s_x^2 will be large and the ratio $(X-\bar{x})/s_x$ will therefore be small.

It should be noted that no assumption whatever of normality has been made.

Simplification of the Calculations by means of Determinants

For the case of two parameters the algebra involved in the application of Markoff's theorem is not heavy, but it rapidly becomes so when the number of parameters increases. Since it is desired to discuss applications of the theorem for the case of three or more parameters the solutions of the equations in the form of determinants are set out below. These determinants reduce the application of the theorem to an arithmetical process. The reader may check the truth of the statements from the proof of the theorem for two parameters.

We refer to the notation given in the statement of the generalization of the Markoff theorem.

If

$$H_0 = \sum_{i=1}^{n} P_i x_i^2,\ H_j = \sum_{i=1}^{n} P_i x_i a_{ij} \text{ for } j = 1, 2, \ldots, s,$$

and $G_{jk} = \sum_{i=1}^{n} P_i a_{ij} a_{ik}$ for $j = 1, 2, \ldots, s$, and $k = 1, 2, \ldots, s$,

then $$F = -\Delta_\theta / \Delta,$$

where

$$\Delta_\theta = \begin{vmatrix} 0 & b_1 & b_2 & \ldots & b_s \\ H_1 & G_{11} & G_{12} & \ldots & G_{1s} \\ H_2 & G_{21} & G_{22} & \ldots & G_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ H_s & G_{s1} & G_{s2} & \ldots & G_{ss} \end{vmatrix} \quad \text{and } \Delta = \begin{vmatrix} G_{11} & G_{12} & \ldots & G_{1s} \\ G_{21} & G_{22} & \ldots & G_{2s} \\ \vdots & \vdots & & \vdots \\ G_{s1} & G_{s2} & \ldots & G_{ss} \end{vmatrix}.$$

Also $$\mu_F^2 = -\frac{\Delta_0 \Delta_1}{(n-s)\Delta^2},$$

where Δ is as already defined, and Δ_0 and Δ_1 are

$$\Delta_0 = \begin{vmatrix} H_0 & H_1 & H_2 & \ldots & H_s \\ H_1 & G_{11} & G_{12} & \ldots & G_{1s} \\ H_2 & G_{21} & G_{22} & \ldots & G_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ H_s & G_{s1} & G_{s2} & \ldots & G_{ss} \end{vmatrix}$$

and

$$\Delta_1 = \begin{vmatrix} 0 & b_1 & b_2 & \ldots & b_s \\ b_1 & G_{11} & G_{12} & \ldots & G_{1s} \\ b_2 & G_{21} & G_{22} & \ldots & G_{2s} \\ \vdots & \vdots & \vdots & & \vdots \\ b_s & G_{s1} & G_{s2} & \ldots & G_{ss} \end{vmatrix}.$$

Application of Markoff Theorem to Partial Regression

We may apply Markoff's theorem to deduce formulae for the estimates and the variances of the estimates of

(i) a partial regression coefficient in an equation with two independent variables, e.g. $z = A + Bx + Cy$,

(ii) the ordinate z of the regression plane corresponding to $x = \xi$, and $y = \eta$.

It is necessary to begin by deciding in Markoff terminology what it is that it is desired to estimate.

If the equation $$z = A + Bx + Cy$$

is considered then it would appear that for the two different cases

$$\text{(i) } \theta = A \text{ or } B \text{ or } C, \qquad \text{(ii) } \theta = A + B\xi + C\eta.$$

It will be necessary to discuss (ii) only because (i) will be solved automatically in the process of the solution for (ii). All the conditions of the Markoff theorem can be made to be satisfied with the exception of the restriction

$$\sigma_i^2 = \sigma^2/P_i.$$

No indication is given in the statement of the problem as to the nature of σ_i^2, and therefore of P_i, and we must needs perforce put $P_i = 1$ for all i. In so doing we shall not invalidate the unbiasedness of the estimate but we can no longer speak of it as the 'best' estimate. If

$$\theta = A + B\xi + C\eta,$$

then, in the terminology of the theorem, $b_1 = 1$, $b_2 = \xi$, $b_3 = \eta$. We have now to consider the expectation of z.

We imagine that we have n sets of observations x_i, y_i, z_i.

$$\mathscr{E}(z_i) = A + Bx_i + Cy_i \quad \text{for } i = 1, 2, \ldots, n.$$

All the quantities necessary to evaluate the determinants of the previous section are available.

$$H_0 = \sum_{i=1}^{n} z_i^2, \quad H_1 = \sum_{i=1}^{n} z_i, \quad H_2 = \sum_{i=1}^{n} x_i z_i, \quad H_3 = \sum_{i=1}^{n} y_i z_i,$$

$$G_{11} = n, \quad G_{22} = \sum_{i=1}^{n} x_i^2, \quad G_{33} = \sum_{i=1}^{n} y_i^2,$$

$$G_{12} = \sum_{i=1}^{n} x_i, \quad G_{23} = \sum_{i=1}^{n} x_i y_i, \quad G_{31} = \sum_{i=1}^{n} y_i.$$

The determinants for F will be

$$\Delta_\theta = \begin{vmatrix} 0 & 1 & \xi & \eta \\ \Sigma z_i & n & \Sigma x_i & \Sigma y_i \\ \Sigma z_i x_i & \Sigma x_i & \Sigma x_i^2 & \Sigma x_i y_i \\ \Sigma z_i y_i & \Sigma y_i & \Sigma x_i y_i & \Sigma y_i^2 \end{vmatrix}$$

and

$$\Delta = \begin{vmatrix} n & \Sigma x_i & \Sigma y_i \\ \Sigma x_i & \Sigma x_i^2 & \Sigma x_i y_i \\ \Sigma y_i & \Sigma x_i y_i & \Sigma y_i^2 \end{vmatrix}.$$

The summation in each case being understood to extend from $i = 1$ to $i = n$. Replace these variables by a new set of variables X_i, Y_i and Z_i, such that

$$X_i = x_i - \bar{x}, \quad Y_i = y_i - \bar{y}, \quad Z_i = z_i - \bar{z},$$

and denote the standard deviations of X, Y, and Z by

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad S_Y^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2, \quad S_Z^2 = \frac{1}{n}\sum_{i=1}^{n} Z_i^2.$$

The determinants become, writing ξ' for $\xi - \bar{x}$, and η' for $\eta - \bar{y}$,

$$\Delta_\theta = \begin{vmatrix} 0 & 1 & \xi' & \eta' \\ 0 & n & 0 & 0 \\ nr_{XZ}S_XS_Z & 0 & nS_X^2 & nr_{XY}S_XS_Y \\ nr_{YZ}S_YS_Z & 0 & nr_{XY}S_XS_Y & nS_Y^2 \end{vmatrix}$$

and

$$\Delta = \begin{vmatrix} n & 0 & 0 \\ 0 & nS_X^2 & nr_{XY}S_XS_Y \\ 0 & nr_{XY}S_XS_Y & nS_Y^2 \end{vmatrix}.$$

Easy algebra will give that

$$F = -\frac{\Delta_\theta}{\Delta} = \bar{z} + (\xi - \bar{x})\frac{S_z}{S_x}\left[\frac{r_{xz} - r_{xy}r_{yz}}{1 - r_{xy}^2}\right] + (\eta - \bar{y})\frac{S_z}{S_y}\left[\frac{r_{yz} - r_{xy}r_{xz}}{1 - r_{xy}^2}\right].$$

It remains to estimate A, B and C. The determinant Δ will be the same as before for each parameter, but Δ_θ will be different.

For A, Δ_θ will be

$$\begin{vmatrix} 0 & 1 & 0 & 0 \\ \Sigma z_i & n & \Sigma x_i & \Sigma y_i \\ \Sigma x_i z_i & \Sigma x_i & \Sigma x_i^2 & \Sigma x_i y_i \\ \Sigma y_i z_i & \Sigma y_i & \Sigma x_i y_i & \Sigma y_i^2 \end{vmatrix},$$

which in terms of the variables X, Y, and Z is zero. It follows that A is the constant involving the means in the general expression already calculated above and that the estimate of A is therefore

$$\text{estimate of } A = \bar{z} - \bar{x}\frac{S_z}{S_x}\left[\frac{r_{xz} - r_{xy}r_{yz}}{1 - r_{xy}^2}\right] - \bar{y}\frac{S_z}{S_y}\left[\frac{r_{yz} - r_{xy}r_{xz}}{1 - r_{xy}^2}\right].$$

The reader may prove this by working out the determinant Δ_θ in terms of the original variables as given above.

For B, Δ_θ will be

$$\Delta_\theta = \begin{vmatrix} 0 & 0 & 1 & 0 \\ \Sigma z_i & n & \Sigma x_i & \Sigma y_i \\ \Sigma z_i x_i & \Sigma x_i & \Sigma x_i^2 & \Sigma x_i y_i \\ \Sigma z_i y_i & \Sigma y_i & \Sigma x_i y_i & \Sigma y_i^2 \end{vmatrix},$$

from which the estimate of B is found to be

$$\text{estimate of } B = \frac{S_z}{S_x}\left(\frac{r_{xz} - r_{xy}r_{yz}}{1 - r_{xy}^2}\right).$$

Similarly for C we have

$$\text{estimate of } C = \frac{S_z}{S_y}\left(\frac{r_{yz} - r_{xy}r_{xz}}{1 - r_{xy}^2}\right).$$

We now proceed to the evaluation of the variance of the estimate of

$$\theta = A + B\xi + C\eta.$$

The variance, μ_F^2, the estimate of σ_F^2, is

$$\mu_F^2 = -\frac{\Delta_1 \Delta_0}{\Delta^2 (n-3)},$$

where Δ_1 and Δ_0 have been defined and Δ already calculated in finding F. In the terminology of the present problem

$$\Delta_1 = \begin{vmatrix} 0 & 1 & \xi & \eta \\ 1 & n & \Sigma x_i & \Sigma y_i \\ \xi & \Sigma x_i & \Sigma x_i^2 & \Sigma x_i y_i \\ \eta & \Sigma y_i & \Sigma x_i y_i & \Sigma y_i^2 \end{vmatrix}$$

and

$$\Delta_0 = \begin{vmatrix} \Sigma z_i^2 & \Sigma z_i & \Sigma z_i x_i & \Sigma z_i y_i \\ \Sigma z_i & n & \Sigma x_i & \Sigma y_i \\ \Sigma z_i x_i & \Sigma x_i & \Sigma x_i^2 & \Sigma x_i y_i \\ \Sigma z_i y_i & \Sigma y_i & \Sigma x_i y_i & \Sigma y_i^2 \end{vmatrix}.$$

These determinants may be simplified as before by transforming the variables, and μ_F^2 is found to be

$$\mu_F^2 = \frac{S_z^2}{(n-3)(1 - r_{xy}^2)}(1 - r_{xy}^2 - r_{yz}^2 - r_{xz}^2 + 2r_{xy}r_{yz}r_{zx})$$
$$\times\left[1 + \frac{1}{1 - r_{xy}^2}\left(\left(\frac{\xi - \bar{x}}{S_x}\right)^2 - 2r_{xy}\left(\frac{\xi - \bar{x}}{S_x}\right)\left(\frac{\eta - \bar{y}}{S_y}\right) + \left(\frac{\eta - \bar{y}}{S_y}\right)^2\right)\right].$$

It may be noted here, as in the case of the regression line, that if x and y are at choice, the estimate of z with a small standard error will be obtained by making the standard errors of x and y as large as possible, but in so doing taking care to balance the values of the observations about ξ and η so that $\xi - \bar{x}$ and $\eta - \bar{y}$ are as small as possible.

It is left to the reader as an exercise to calculate the estimates of the variances of A, B, and C. By evaluating the appropriate determinants these will be found to be

$$\mu_A^2 = \frac{S_z^2}{(n-3)(1-r_{xy}^2)}(1-r_{xy}^2-r_{yz}^2-r_{xz}^2+2r_{xy}r_{yz}r_{xz}),$$

$$\mu_B^2 = \frac{S_z^2}{(n-3)S_x^2(1-r_{xy}^2)^2}(1-r_{xy}^2-r_{yz}^2-r_{xz}^2+2r_{xy}r_{yz}r_{xz}),$$

$$\mu_C^2 = \frac{S_z^2}{(n-3)S_y^2(1-r_{xy}^2)^2}(1-r_{xy}^2-r_{yz}^2-r_{xz}^2+2r_{xy}r_{yz}r_{xz}).$$

Numerical Application

An example of the numerical application of the regression-line formulae will be found in the following problem.

Table I gives the distribution of yield estimates of sugar beet per acre on 100 fields of 50 acres each. The estimates are made by eye some time before the harvest and are expressed in terms of conventional marks varying from 1 to 10. Table II represents the experience with similar marking in the previous year, the markings x being correlated with the actual yields y in tons per acre obtained on harvesting.

It may be assumed that the regression of y on x is linear and that the y arrays are homoscedastic. Use the data given in the tables to calculate the best linear unbiased estimate, F, of the total yield Y on an area of 5000 acres with estimated yields as in Table I, and find the standard error of Y. (M.Sc. London, 1937.)

It is clear that this is a case for the application of Markoff's theorem. It is given that

(1) the regression of y on x is linear, i.e. $\mathscr{E}(y) = p_1 + p_2 x$,

(2) the standard deviation of y in arrays is constant, i.e.

$$P_i = 1 \quad \text{for } i = 1, 2, \ldots, 10.$$

TABLE I

Estimate of yield per acre in marks	Number of fields
1	1
2	3
3	7
4	5
5	10
6	19
7	30
8	10
9	7
10	8
Total	100

TABLE II. *Correlation between estimated yield x and actual yield y*

y \ x	1	2	3	4	5	6	7	8	9	10	x / Totals
18	.	.	.	.	.	.	.	.	1	4	5
16	.	.	.	.	.	1	2	1	3	3	10
14	.	.	.	1	3	7	5	4	3	.	23
12	.	.	4	10	4	8	3	1	.	.	30
10	.	.	5	8	9	4	1	.	.	.	27
8	.	2	7	7	5	.	.	.	.	.	21
6	1	7	3	1	.	.	.	.	.	.	12
4	5	3	1	.	.	.	.	.	.	.	9
y / Totals	6	12	20	27	21	20	11	6	7	7	137

Assume that there are n_1 acres with mark 1, n_2 acres with mark 2, ..., n_{10} acres with mark 10.

It follows that

$$\theta = \sum_{X=1}^{10} n_X(p_1 + p_2 X).$$

From previous work with the regression line we may at once write down

$$F = \sum_{X=1}^{10} n_X \left(\bar{y} + (X - \bar{x}) r_{xy} \frac{S_y}{S_x}\right)$$

and

$$\mu_F^2 = \frac{1}{n-2} \frac{S_y^2}{S_x^2} (1 - r_{xy}^2) \left[S_x^2 \left(\sum_{X=1}^{10} n_X \right)^2 + \left(\sum_{X=1}^{10} n_X X - \bar{x} \sum_{X=1}^{10} n_X \right)^2 \right].$$

From the data of Table II

$$F = \sum_{X=1}^{10} n_X[10{\cdot}788 + 1{\cdot}372(X - 4{\cdot}971)].$$

n_X are the figures on the right-hand side of Table I multiplied by 50. Hence giving X values in turn 1, 2, ..., 10 and multiplying by n_X it is found that

$$F = 64{,}089 \text{ tons.}$$

It is left to the reader to find the numerical value of μ_F^2.

REFERENCES AND READING

The proof of the generalized Markoff theorem on least squares will be found in J. Neyman and F. N. David, 'Extension of the Markoff theorem on least squares', *Statist. Res. Mem.* II, p. 105.

The original Markoff theorem may be found in A. A. Markoff, *Wahrscheinlichkeitsrechnung*. Leipzig and Berlin, 1912.

A. C. Aitken has carried the generalization further than as given in this chapter in A. C. Aitken (1935). *Proc. Roy. Soc. Edinb.* LV.

CHAPTER XIV

FURTHER APPLICATIONS OF MARKOFF'S METHOD. SAMPLING HUMAN POPULATIONS

Every ten years in normal peacetime conditions it is the custom in the United Kingdom to carry out a complete enumeration of persons in the British Isles. In addition to the counting of heads, various questions are asked such as the age and sex of the individual and so on. The information thus obtained is tabulated and collated and provides information regarding the distribution of individuals which is vital for the governing of the country. Occasionally, however, even such complete enumerations as this may give rise to misleading conclusions. A complete census of individuals was taken in September 1939 after war had been declared and large-scale evacuation had taken place. During late 1939 and early 1940 there was a steady drift back to the towns and many men were called to the armed forces; as a consequence when air raids began in July 1940 the statistician and administrator had no really precise idea of the size of the populations which would be, and were, exposed to risk in the large towns.

It is not suggested that the peacetime census figures will be subject to such vast fluctuations as those of the National Register of 1939. Nevertheless, ten years is a long time and under the press of modern conditions it may well be that the legislator will need to supplement the ten-yearly figures by a sample census taken at more frequent intervals. An example of this may be found in the sampling scheme carried out in 1946 to obtain information regarding the size of families.

There are other arguments which may be put forward in favour of conducting sample enumerations. If the country is to have a planned economy then it will be necessary to find out what is the minimum consumption by individuals of certain goods. Sampling surveys to this end were carried out during the war by several Government Departments such as the Board of Trade and the Ministry of Food. Only in this way will it be possible in a time of scarcity to ensure that no one goes without but that

there is no surplus and therefore no waste of manpower occurs. Thus it would seem that the process of sampling human populations, well known to statisticians long before the war, is likely to be used frequently by legislators and planners.

It is customary to begin by fixing the size of the sample to be collected, and this is decided most often not on statistical principles but by the amount of money available to be spent on the collection of the material and the urgency with which the answer is required. A large sample will be costly both in money and in the time necessary to analyse the results. The statistician may state therefore what would seem to be a minimum figure, but the actual determination of the size of sample will not be entirely in his hands.

Suppose it is decided to collect a sample of n out of a total of N where both n and N are likely to be large numbers. For example, if N is composed of units which are industrial firms then N may be of the order of 50,000 and n of the order of 1000. The ratio of n/N will therefore be 1/50. The question now arises as to how the sample should be selected. The obvious way would seem to be to select n firms at random from the N firms, or if the list of N firms is in random order to choose every fiftieth firm for investigation. This procedure is sound enough, for the n firms would give an estimate of the population of N firms which would be unbiased. It would, however, in many cases lead to rather a large standard error of the estimate.

Consider, for example, firms employed in building houses. The number of firms which employ only one man runs into thousands, while the number of firms which employ thousands of men is very small. If it is desired to find out something about the output per man-hour it would be misleading to sample the population of firms at random. The number of large firms are few and would have little chance of being included in the sample; conversely, if they were included they might upset the estimate from a small sample radically. It is clear therefore that before sampling takes place the material must be divided into groups, commonly denoted as strata, such that the material within each of these strata is as far as possible homogeneous. For instance, in the case of the building firms the obvious method of dichotomy would be to take the first stratum of firms employing one building

operative only, and so on. If there were obvious differences between such firms then it might be necessary to divide the strata into substrata, but the method of stratification will usually be clear.

If we have a population N divided into k strata containing $M_1, M_2, \ldots, M_k$ individuals such that

$$N = \sum_{i=1}^{k} M_i,$$

then the intuitive choice would be to take numbers from each stratum proportional to its size, subject to the restriction that the total sample size is to be n. Thus if $m_1, m_2, \ldots, m_k$ are the sub-samples chosen from the k strata then

$$\sum_{i=1}^{k} m_i = n \quad \text{and} \quad m_i = \frac{M_i}{N} n.$$

There is, however, one further point which may be considered. The variation of individuals within some strata may be less than the variation in others. Consequently a smaller sub-sample need be taken from the strata in which the variation is small, and a larger sub-sample from the strata in which the variation is large. The methods of the Markoff theorem may be used to show that the choice of the number of individuals proportional to the number in the stratum provides an unbiased estimate of the collective character in the population, but that if the variance within the stratum is known then it is not the best estimate which can be made.

Suppose a population π is divided into k strata, $\pi_1, \pi_2, \ldots, \pi_k$, and that the ith stratum π_i of the population π contains M_i individuals ($i = 1, 2, \ldots, k$).

Let u_{ij} be the measured character of the jth individual of the ith stratum, let

$$\bar{u}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} u_{ij}$$

and let σ_i^2 be

$$\sigma_i^2 = \frac{1}{M_i} \sum_{j=1}^{M_i} u_{ij}^2 - \bar{u}_i^2.$$

It is required to estimate

$$\theta = \sum_{i=1}^{k} \sum_{j=1}^{M_i} u_{ij} = \sum_{i=1}^{k} M_i \bar{u}_i.$$

Let m_i be the size of the sub-sample selected from the stratum π_i, and let x_{ij} be the value of the jth element of this sub-sample. Further, let the best linear unbiased estimate of θ be F, where F is a function

$$F = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \lambda_{ij} x_{ij}$$

and the λ's are constants to be determined.

It is clear that

$$\mathscr{E}(x_{ij}) = \bar{u}_i,$$

whence, remembering that the expectation of F must be identically equal to θ, we have

$$\sum_{i=1}^{k} \bar{u}_i \left(\sum_{j=1}^{m_i} \lambda_{ij} - M_i \right) \equiv 0.$$

$\bar{u}_i$ is constant for any given stratum and it follows therefore that the λ_{ij} must satisfy the condition

$$\sum_{j=1}^{m_i} \lambda_{ij} = M_i$$

for all i. The (standard error)^2 of F follows directly from definition,

$$\sigma_F^2 = \mathscr{E}(F - \theta)^2.$$

In any given stratum the sampling will be without replacement. Write for convenience

$$\lambda_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \lambda_{ij} = \frac{M_i}{m_i}.$$

By definition

$$\sigma_F^2 = \mathscr{E}\left[\sum_{i=1}^{k} \sum_{j=1}^{m_i} \lambda_{ij}(x_{ij} - \bar{u}_i) \right]^2$$

$$= \mathscr{E}\left[\sum_{i=1}^{k} \sum_{j=1}^{m_i} \lambda_{ij}^2 (x_{ij} - \bar{u}_i)^2 + \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq t}}^{m_i} \sum_{t=1}^{m_i} \lambda_{ij} \lambda_{it} (x_{ij} - \bar{u}_i)(x_{it} - \bar{u}_i) \right.$$

$$\left. + 2 \sum_{i=1}^{k-1} \sum_{s=i+1}^{k} \sum_{j=1}^{m_i} \sum_{t=1}^{m_s} \lambda_{ij} \lambda_{st} (x_{ij} - \bar{u}_i)(x_{st} - \bar{u}_s) \right].$$

The first part of this expression is immediately evaluated.

$$\mathscr{E} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \lambda_{ij}^2 (x_{ij} - \bar{u}_i)^2 = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \lambda_{ij}^2 \sigma_i^2.$$

The second term involving the cross-products is more difficult but only in so far as the enumeration of the individual terms is

concerned. The third term is zero for obviously x_{ij} is independent of x_{st} for all $i \neq s$, for the drawing of an individual from one stratum cannot affect the chances of an individual being drawn from another. It will follow that

$$\mathscr{E} \sum_{j=1}^{m_i} \sum_{t=1}^{m_s} \lambda_{ij}\lambda_{st}(x_{ij}-\bar{u}_i)(x_{st}-\bar{u}_s) = 0.$$

The second term will not be zero. The drawing of one individual from the ith stratum will affect the chances of another individual from the same stratum of being drawn and therefore the drawings of elements from the ith stratum cannot be considered as independent. Consider the expectation of a typical term of the summation, say

$$\mathscr{E}(\lambda_{ij}\lambda_{it}(x_{ij}-\bar{u}_i)(x_{it}-\bar{u}_i)).$$

By an appeal to first principles it is clear that

$$\mathscr{E}(x_{ij}-\bar{u}_i)(x_{it}-\bar{u}_i) = \sum_{h=1}^{M_i-1} \sum_{l=h+1}^{M_i} (u_{ih}-\bar{u}_i)(u_{il}-\bar{u}_i)\frac{2!(M_i-2)!}{M_i!}.$$

This may be simplified by the device used in Chapter XI and we have

$$0 = \left(\sum_{h=1}^{M_i}(u_{ih}-\bar{u}_i)\right)^2$$

$$= \sum_{h=1}^{M_i}(u_{ih}-\bar{u}_i)^2 + 2\sum_{h=1}^{M_i-1}\sum_{l=h+1}^{M_i}(u_{ih}-\bar{u}_i)(u_{il}-\bar{u}_i)$$

from which it is easy to see that

$$-M_i\sigma_i^2 = 2\sum_{h=1}^{M_i-1}\sum_{l=h+1}^{M_i}(u_{ih}-\bar{u}_i)(u_{il}-\bar{u}_i)$$

and that $\mathscr{E}(\lambda_{ij}\lambda_{it}(x_{ij}-\bar{u}_i)(x_{it}-\bar{u}_i)) = -\lambda_{ij}\lambda_{it}\dfrac{\sigma_i^2}{M_i-1}$.

Using this relation, we have for the expression of σ_F^2,

$$\sigma_F^2 = \left[\sum_{i=1}^{k}\sigma_i^2\sum_{j=1}^{m_i}\lambda_{ij}^2 + \sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq t}}^{m_i}\sum_{t=1}^{m_i}\lambda_{ij}\lambda_{it}\left(-\frac{\sigma_i^2}{M_i-1}\right)\right].$$

Making use of the fact that

$$\left(\sum_{j=1}^{m_i}\lambda_{ij}\right)^2 = \sum_{j=1}^{m_i}\lambda_{ij}^2 + \sum_{\substack{j=1\\ j\neq t}}^{m_i}\sum_{t=1}^{m_i}\lambda_{ij}\lambda_{it}$$

the summation signs of the second term on the right-hand side can be reduced to two and σ_F^2 rewritten

$$\sigma_F^2 = \left[\sum_{i=1}^{k} \sigma_i^2 \left(\frac{M_i}{M_i - 1}\right) \sum_{j=1}^{m_i} \lambda_{ij}^2 - \sum_{i=1}^{k} \frac{\sigma_i^2}{M_i - 1} \left(\sum_{j=1}^{m_i} \lambda_{ij}\right)^2\right].$$

We now introduce λ_i. It is easy to show that

$$\sum_{j=1}^{m_i} (\lambda_{ij} - \lambda_i)^2 = \sum_{j=1}^{m_i} \lambda_{ij}^2 - \frac{1}{m_i} \left(\sum_{j=1}^{m_i} \lambda_{ij}\right)^2,$$

whence on substitution into the expression for σ_F^2 we obtain finally

$$\sigma_F^2 = \sum_{i=1}^{k} \left[\sigma_i^2 \left(m_i \frac{M_i - m_i}{M_i - 1} \lambda_i^2 + \frac{M_i}{M_i - 1} \sum_{j=1}^{m_i} (\lambda_{ij} - \lambda_i)^2\right)\right].$$

It is clear that the λ's which will satisfy the given condition and also minimize σ_F^2 will be

$$\lambda_{ij} = \lambda_i = \frac{M_i}{m_i}.$$

Substituting in these values for F and σ_F^2, if

$$\bar{x}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij},$$

then

$$F = \sum_{i=1}^{k} M_i \bar{x}_i$$

and

$$\sigma_F^2 = \sum_{i=1}^{k} \left[\sigma_i^2 \frac{M_i^2}{m_i} \frac{M_i - m_i}{M_i - 1}\right].$$

The variance of the estimate F will therefore depend on the size of the sub-sample taken from each stratum, for σ_i and M_i are fixed numbers descriptive of the population. If the whole population was enumerated then there would be no estimation involved, the exact value of F would be θ and the variance of F would be zero.

The size of the population is fixed and is equal to

$$N = \sum_{i=1}^{k} M_i.$$

We have said it is customary to fix before sampling the approxi-

mate size of the sample to be drawn from the population. Let n therefore be a fixed number where

$$n = \sum_{i=1}^{k} m_i.$$

Let $$S_i^2 = \frac{M_i \sigma_i^2}{M_i - 1}$$

and rewrite σ_F^2, somewhat arbitrarily, in the following way:

$$\sigma_F^2 = \frac{N-n}{n} \sum_{i=1}^{k} M_i S_i^2 + \sum_{i=1}^{k} m_i \left[\frac{M_i S_i}{m_i} - \frac{\sum_{i=1}^{k} M_i S_i}{n} \right]^2 - \frac{N}{n} \sum_{i=1}^{k} M_i \left[S_i - \frac{\sum_{i=1}^{k} M_i S_i}{N} \right]^2.$$

It may be verified algebraically that this reduces to the expression for σ_F^2 already found. If the size of the sub-sample drawn from the stratum is proportional to the total number within the stratum, i.e. if m_i is proportional to M_i, then

$$\sigma_F^2 = \frac{N-n}{n} \sum_{i=1}^{k} M_i S_i^2.$$

If m_i is proportional to $M_i S_i$, then

$$\sigma_F^2 = \frac{N-n}{n} \sum_{i=1}^{k} M_i S_i^2 - \frac{N}{n} \sum_{i=1}^{k} M_i \left(S_i - \frac{\sum_{i=1}^{k} M_i S_i}{N} \right)^2.$$

This implies that where the size of the sample is fixed a smaller value will be obtained for the variance of F if the numbers chosen for the sample from each stratum are proportional to the number within the stratum and to the standard error. Neyman has pointed out that, while the standard error within each stratum is not known generally, improvement in the accuracy of the estimate will be obtained if it is decided to begin sampling proportionately to the number in the stratum only, and then, by estimating the standard error within each stratum from the first observations taken, to readjust the numbers of the sub-sample so that they are proportional to $M_i S_i$.

Example. In some countries annual surveys of the crops grown on farms are made. For this purpose each year a stratified random sample of 1000 farms might be taken to provide data whereby yields may be estimated, etc. We shall suppose that the whole population of farms is divided into strata according to their acreage. It is desired to estimate the total acreage under wheat and it may be assumed that the standard error of this acreage does not vary to any marked extent within farms of a given acreage from year to year. It will be legitimate therefore to use the estimates of S_i which have been calculated from previous years. These are given in the following table. (M_i in hundreds.)

Acreage code no.	S_i	M_i	Acreage code no.	S_i	M_i
1	7·3	180	6	9·5	20
2	1·3	100	7	4·5	15
3	5·1	110	8	10·7	3
4	2·1	70	9	6·6	6
5	9·2	50	10	11·2	2

Calculate the numbers which should be drawn from each stratum assuming (1) $m_i \propto M_i$, (2) $m_i \propto M_i S_i$, and find the variance of the estimate F in each case.

The equations of the theory just set out will enable the numerical calculations to be carried out without any further algebra. It is given that

$$n = 1000, \quad N = 55{,}600,$$

and hence the numbers to be drawn from each stratum will be:

Acreage code no.	1	2	3	4	5	6	7	8	9	10	Total
$m_i \propto M_i$	324	180	198	126	90	36	27	5	11	4	1001
$m_i \propto M_i S_i$	443	44	189	50	155	64	23	11	13	8	1000

The effect of sampling proportionately to $M_i S_i$ is, as might have been expected, to increase the size of the sub-samples to be drawn from those strata with large standard deviations.

σ_F^2 follows directly from the expression given or may be calculated from

$$\sigma_F^2 = \sum_{i=1}^{k} M_i S_i^2 \left(\frac{M_i - m_i}{m_i} \right),$$

where m_i $(i = 1, 2, \ldots, k)$ are first the numbers proportional to M_i and second the numbers proportional to $M_i S_i$. It is left to the student to carry through the arithmetical processes necessary to show that

$$\sigma_{\bar{x}}^2(m_i \propto M_i S_i) < \sigma_{\bar{x}}^2(m_i \propto M_i).$$

Restricted Stratified Sampling

The theory of stratified sampling, which has just been set out, is well known and has been in use for some years. This method should be adopted in all cases where direct information is available about the desired character once the sampling element has been drawn. There are, however, cases where, even after the sample has been drawn, information is not readily available or it is perhaps costly in time and money to obtain. For example, a great many people may be found on inquiry to have an objection to telling the investigator the amount of their weekly wages. If, therefore, these people have been drawn as part of the random sample, trouble is caused if they refuse information. This is a real difficulty and it is difficult to overcome.

If information cannot be obtained about a character X without difficulty, it has been suggested that it should be sought about another character Y which it is reasonable to suppose is highly correlated with X. For example, if X is the amount of the weekly wage, then X may be correlated with Y, the rateable value of the house in which the individual lives. Thus we should stratify the population according to Y and then sample within each stratum to obtain information about X. The situation is, however, complicated in that the distribution of Y within a given population may not often be known. The proposed method consists therefore of drawing a large sample randomly from the population and stratifying it according to the values of Y. From this stratified sample a further sample is drawn from which information is sought about X.

It is open to question whether this method is more efficient than the normal method of stratified sampling, but if the correlation between X and Y is close, then in some circumstances this proposed method may be of value.

Let the proportions of Y in the strata forming the complete population, π, be $p_1, p_2, \ldots, p_k$. If $\bar{X}_i$ denotes the population mean

of the X's for the ith stratum ($i = 1, 2, \ldots, k$), then we shall assume that it is required to estimate

$$\bar{X} = \sum_{i=1}^{k} p_i \bar{X}_i.$$

Let the first sample to be drawn be S_1 consisting of N elements. These elements or individuals are stratified with respect to Y. If n_i of these individuals fall in the ith stratum, then $r_i = n_i/N$ will be an estimate of p_i, where

$$p_i = \mathscr{E}(r_i) = \mathscr{E}\left(\frac{n_i}{N}\right).$$

Let the second sample, drawn from S_1, be S_2. Let S_2 consist of n individuals of which m_i are drawn from the ith stratum of S_1.

If x_{ij} denotes the jth individual drawn from the ith stratum then we shall write

$$\frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij} = \bar{x}_i.$$

We require to estimate $\bar{X}$ so we next consider a linear function

$$F = \sum_{i=1}^{k} \sum_{l=1}^{k} \sum_{j=1}^{m_l} \lambda_{ilj} r_i x_{lj}$$

the λ's of which will be determined by the Markoff method. Noting that

$$\mathscr{E}(F) \equiv \bar{X}$$

we have also

$$\mathscr{E}(F) = \sum_{i=1}^{k} p_i \bar{X}_i \equiv \sum_{i=1}^{k} \sum_{l=1}^{k} \sum_{j=1}^{m_l} \lambda_{ilj} \mathscr{E}(r_i x_{lj}).$$

Since S_1 was drawn without any consideration of the values of S_2, it follows that, even if S_2 is drawn from S_1, the variable x_{lj} will be independent of r_i. Hence

$$\sum_{i=1}^{k} \sum_{l=1}^{k} \sum_{j=1}^{m_l} \lambda_{ilj} \mathscr{E}(r_i x_{lj})$$

$$= \sum_{i=1}^{k} \sum_{l=1}^{k} \sum_{j=1}^{m_l} \lambda_{ilj} \mathscr{E}(r_i)\, \mathscr{E}(x_{lj}) = \sum_{i=1}^{k} \sum_{l=1}^{k} \sum_{j=1}^{m_l} \lambda_{ilj} p_i \bar{X}_l$$

and it will follow that

$$\sum_{i=1}^{k} p_i \left(\sum_{l=1}^{k} \bar{X}_l \sum_{j=1}^{m_l} \lambda_{ilj} - \bar{X}_i \right) \equiv 0.$$

The necessary and sufficient condition that this shall hold good identically is that

$$\sum_{l=1}^{k} \bar{X}_l \sum_{j=1}^{m_l} \lambda_{ilj} - \bar{X}_i \equiv 0 \quad \text{for } i = 1, 2, \ldots, k.$$

Rewrite this in the following way:

$$\sum_{l=1}^{i-1} \bar{X}_l \sum_{j=1}^{m_l} \lambda_{ilj} + \bar{X}_i \left(\sum_{j=1}^{m_i} \lambda_{iij} - 1 \right) + \sum_{l=i+1}^{k} \bar{X}_l \sum_{j=1}^{m_l} \lambda_{ilj} \equiv 0.$$

In order to satisfy the identity we have that

$$\sum_{j=1}^{m_l} \lambda_{ilj} = 0 \quad \text{for } i \neq l \quad (i = 1, 2, \ldots, k;\ l = 1, 2, \ldots, k),$$

$$\sum_{j=1}^{m_i} \lambda_{iij} = 1 \quad \text{for } i = 1, 2, \ldots, k.$$

It is necessary now to find the λ's which satisfy these conditions and which make σ_F^2 a minimum, where

$$\sigma_F^2 = \mathscr{E}\left(\sum_{i=1}^{k} \sum_{l=1}^{k} \sum_{j=1}^{m_l} \lambda_{ilj} r_i x_{lj} - \sum_{i=1}^{k} p_i \bar{X}_i \right)^2.$$

For convenience write

$$\xi_i = \sum_{l=1}^{k} \sum_{j=1}^{m_l} \lambda_{ilj} x_{lj}$$

and calculate its expectation.

$$\mathscr{E}(\xi_i) = \sum_{l=1}^{k} \bar{X}_l \sum_{j=1}^{m_l} \lambda_{ilj} = \bar{X}_i \sum_{j=1}^{m_i} \lambda_{iij} = \bar{X}_i.$$

We may rewrite σ_F^2.

$$\sigma_F^2 = \mathscr{E}\left(\sum_{i=1}^{k} (r_i \xi_i - p_i \bar{X}_i) \right)^2$$

$$= \mathscr{E}\left[\sum_{i=1}^{k} (r_i \xi_i - p_i \bar{X}_i)^2 + 2 \sum_{i=1}^{k-1} \sum_{t=i+1}^{k} (r_i \xi_i - p_i \bar{X}_i)(r_t \xi_t - p_t \bar{X}_t) \right].$$

It will be convenient to calculate this in two stages and we shall therefore focus attention first on $\mathscr{E}\left(\sum_{i=1}^{k} (r_i \xi_i - p_i \bar{X}_i)^2 \right)$.

$$\mathscr{E}(r_i \xi_i - p_i \bar{X}_i)^2 = \mathscr{E}((r_i - p_i)^2 \xi_i^2) + \mathscr{E}(p_i^2 (\xi_i - \bar{X}_i)^2),$$

since r_i and ξ_i are independent. We require therefore to evaluate

$$\mathscr{E}\left(\sum_{i=1}^{k}(r_i\xi_i - p_i\bar{X}_i)^2\right) = \sum_{i=1}^{k}\mathscr{E}(r_i-p_i)^2\,\mathscr{E}(\xi_i^2) + \sum_{i=1}^{k}p_i^2\,\mathscr{E}(\xi_i-\bar{X}_i)^2.$$

At this stage, in order to simplify the algebra, it will be assumed that the parent population was large enough for any individual drawn to be independent of any other. This assumption is not very restricting for the case of human populations which are usually so large that it is virtually true.

We now introduce the device of the characteristic random variable in order to calculate $\mathscr{E}(r_i^2)$ and $\mathscr{E}(r_ir_j)$ which will be needed for the evaluation of the second part of σ_F^2. Two strata only, the ith and the gth, need be discussed, for the same arguments will hold good for other strata. Associate with the sample of N individuals two series of characteristic random variables α_f and β_v, for $f, v = 1, 2, \ldots, N$. These variables will have the properties

$\alpha_f = 1$ if the individual falls in the ith stratum and zero otherwise.

$\beta_v = 1$ if the individual falls in the gth stratum and zero otherwise.

It is obvious since

$$\sum_{f=1}^{N}\alpha_f = n_i \quad \text{and} \quad \sum_{v=1}^{N}\beta_v = n_g$$

that

$$\frac{1}{N}\sum_{f=1}^{N}\alpha_f = r_i \quad \text{and} \quad \frac{1}{N}\sum_{v=1}^{N}\beta_v = r_g.$$

We have said that $\mathscr{E}(n_i) = Np_i$

and it is clear that

$$\mathscr{E}(n_i^2) = \mathscr{E}\left(\sum_{f=1}^{N}\alpha_f\right)^2$$

$$= \mathscr{E}\left(\sum_{f=1}^{N}\alpha_f^2 + 2\sum_{f=1}^{N-1}\sum_{s=f+1}^{N}\alpha_f\alpha_s\right) = Np_i + N(N-1)p_i^2,$$

whence

$$\mathscr{E}(r_i^2) = \frac{p_iq_i}{N} + p_i^2.$$

The expectation of the product $r_i r_g$ follows in a similar way.

$$\mathscr{E}(r_i r_g) = \frac{1}{N^2}\mathscr{E}(n_i n_g) = \frac{1}{N^2}\mathscr{E}\left(\sum_{f=1}^{N}\alpha_f \sum_{v=1}^{N}\beta_v\right)$$

$$= \frac{1}{N^2}\mathscr{E}\left(\sum_{f=1}^{N}\alpha_f\beta_f + \sum_{f\neq v}\sum(\alpha_f\beta_v + \alpha_v\beta_f)\right) = p_i p_g\left(1-\frac{1}{N}\right).$$

We return to the evaluation of the first term of $\sigma_{\bar{F}}^2$. It is simple to show that

$$\mathscr{E}(\xi_i - \bar{X}_i)^2 = \sum_{l=1}^{k}\sigma_l^2\sum_{j=1}^{m_l}\lambda_{ilj}^2,$$

since the x's are assumed all independent and hence, on substitution,

$$\mathscr{E}\left(\sum_{i=1}^{k}(r_i\xi_i - p_i\bar{X}_i)^2\right)$$

$$= \sum_{i=1}^{k}\frac{p_i q_i}{N}\left(\sum_{l=1}^{k}\sigma_l^2\sum_{j=1}^{m_l}\lambda_{ilj}^2 + \bar{X}_i^2\right) + \sum_{i=1}^{k}p_i^2\sum_{l=1}^{k}\sigma_l^2\sum_{j=1}^{m_l}\lambda_{ilj}^2,$$

and the first term is evaluated.

We now turn to the second term of the expression for $\sigma_{\bar{F}}^2$, which may be rewritten

$$\mathscr{E}((r_i\xi_i - p_i\bar{X}_i)(r_t\xi_t - p_t\bar{X}_t)) = \mathscr{E}(r_i r_t)\,\mathscr{E}(\xi_i\xi_t) - p_i p_t\bar{X}_i\bar{X}_t,$$

since r_i and ξ_t, r_t and ξ_t are independent.

$$\mathscr{E}(\xi_i\xi_t) = \mathscr{E}\left(\sum_{l=1}^{k}\sum_{j=1}^{m_l}\lambda_{ilj}\lambda_{tlj}x_{lj}^2 + \sum_{l=1}^{k}\sum_{j=1}^{m_l}\sum_{g=1}^{k}\sum_{u=1}^{m_q}\lambda_{ilj}\lambda_{tgu}x_{lj}x_{gu}\right),$$

which on substituting for $\mathscr{E}(x_{lj})$ and $\mathscr{E}(x_{lj}^2)$ and using the conditions for unbiasedness reduces to

$$\mathscr{E}(\xi_i\xi_t) = \sum_{l=1}^{k}\sigma_l^2\sum_{j=1}^{m_l}\lambda_{ilj}\lambda_{tlj} + \bar{X}_i\bar{X}_t$$

and $\mathscr{E}((r_i\xi_i - p_i\bar{X}_i)(r_t\xi_t - p_t\bar{X}_t))$

$$= \left(1-\frac{1}{N}\right)p_i p_t\left(\sum_{l=1}^{k}\sigma_l^2\sum_{j=1}^{m_l}\lambda_{ilj}\lambda_{tlj} + \bar{X}_i\bar{X}_t\right) - p_i p_t\bar{X}_i\bar{X}_t.$$

The two terms comprising the expression for $\sigma_{\bar{F}}^2$ are thus evaluated and

$$\sigma_{\bar{F}}^2 = \sum_{i=1}^{k}\left[\left(p_i^2 + \frac{p_i q_i}{N}\right)\sum_{l=1}^{k}\sigma_l^2\sum_{j=1}^{m_l}\lambda_{ilj}^2\right] + \sum_{i=1}^{k}\frac{p_i q_i}{N}\bar{X}_i^2$$

$$+ \frac{2}{N}\sum_{i=1}^{k-1}\sum_{t=i+1}^{k}p_i p_t\left\{(N-1)\sum_{l=1}^{k}\sigma_l^2\sum_{j=1}^{m_l}\lambda_{ilj}\lambda_{tlj} - \bar{X}_i\bar{X}_t\right\}.$$

Minimum value of σ_F^2

The conditions for the unbiasedness of the estimate F have been found and the expression for σ_F^2. It remains to find what values of the λ's will satisfy the conditions for F and at the same time make σ_F^2 a minimum. Consider a function

$$\phi = \sigma_F^2 - 2\sum_{i=1}^{k}\sum_{l=1}^{k}\alpha_{il}\sum_{j=1}^{m_l}\lambda_{ilj},$$

where the α's are Lagrange's undetermined multipliers. By differentiating ϕ with respect to λ_{ilj} and equating the result to zero we have

$$\alpha_{il} = \frac{1}{N}p_i\sigma_l^2\left(\lambda_{ilj} + (N-1)\sum_{t=1}^{k}p_t\lambda_{tlj}\right).$$

Summing for all j and substituting the conditions for unbiasedness

$$m_l\alpha_{il} = \frac{N-1}{N}p_ip_l\sigma_l^2 \qquad \text{for } i \neq l,$$

$$m_l\alpha_{ll} = \frac{1}{N}p_l\sigma_l^2 + \frac{N-1}{N}p_l^2\sigma_l^2 \quad \text{for } i = l.$$

It is easy to show by substituting these last values in the expression for α_{il} that

$$\lambda_{ilj} = \lambda_{.lj} \qquad \text{when } i \neq l,$$

$$\lambda_{llj} = \lambda_{.lj} + \frac{1}{m_l} \quad \text{when } i = l,$$

where

$$\lambda_{.lj} = (N-1)\left(\frac{p_l}{m_l} - \sum_{t=1}^{k}p_t\lambda_{tlj}\right).$$

It follows by splitting the right-hand sum into the three parts

$$\sum_{t=1}^{l-1}\lambda_{tlj}p_t, \quad \lambda_{llj}p_l, \quad \sum_{t=l+1}^{k}\lambda_{tlj}p_t,$$

that

$$\lambda_{ilj} = \lambda_{.lj} = 0 \quad \text{for } i \neq l,$$

$$\lambda_{llj} = \frac{1}{m_l} \qquad \text{for } i = l.$$

The best linear unbiased estimate of F will be therefore

$$F = \sum_{l=1}^{k} r_l \bar{x}_l$$

and
$$\sigma_F^2 = \sum_{i=1}^{k} \left(p_i^2 + \frac{p_i q_i}{N}\right) \frac{\sigma_i^2}{m_i} + \frac{1}{N} \sum_{i=1}^{k} p_i (\bar{X}_i - \bar{X})^2.$$

Choice of m_i

The term in σ_F^2 containing m_i and therefore at choice for altering σ_F^2 is the first sum. n is the total size of the sample S_2, i.e.

$$n = \sum_{i=1}^{k} m_i.$$

The sum containing m_i in the expression for σ_F^2 may be rearranged to give

$$\sum_{i=1}^{k} \left(p_i^2 + \frac{p_i q_i}{N}\right) \frac{\sigma_i^2}{m_i} = \frac{1}{n} \left(\sum_{i=1}^{k} \sigma_i \sqrt{\left(p_i^2 + \frac{p_i q_i}{N}\right)} \right)^2$$
$$+ \sum_{i=1}^{k} m_i \left(\frac{\sigma_i \sqrt{(p_i^2 + p_i q_i/N)}}{m_i} - \frac{\sum_{i=1}^{k} \sigma_i \sqrt{(p_i^2 + p_i q_i/N)}}{n} \right)^2.$$

This may be checked by expanding the right-hand side of the equation. The minimum value of this expression will be reached when the value of the second term is zero, that is when

$$m_i = n \frac{\sigma_i \sqrt{(p_i^2 + p_i q_i/N)}}{\sum_{i=1}^{k} \sigma_i \sqrt{(p_i^2 + p_i q_i/N)}}$$

or, since p_i and q_i are proportions less than unity and N is a large number, when

$$m_i = n \frac{\sigma_i p_i}{\sum_{i=1}^{k} \sigma_i p_i},$$

when the expression for σ_F^2 becomes

$$\sigma_F^2 = \frac{1}{n} \left(\sum_{i=1}^{k} \sigma_i p_i \right)^2 + \frac{1}{N} \sum_{i=1}^{k} p_i (\bar{X}_i - \bar{X})^2.$$

An alternate method to the above is to use the method of Lagrange's Undetermined Multipliers and minimize σ_F^2 subject to the restriction that the sum of the m_i is equal to n.

COROLLARY. The above analysis may be carried a stage further by deciding, if the total sum of money to be spent is fixed, what proportion of the sum should be spent on collecting the first sample S_1. Let C be the total sum to be spent on the inquiry, let A be the cost per individual of collecting information about X and B the cost per individual of collecting information about Y. Then

$$C = An + BN.$$

Let L_A and L_B be the smallest numbers consistent with the relation

$$L_A . B = L_B . A,$$

and for convenience write

$$\sigma_F^2 = \frac{a^2}{n} + \frac{b^2}{N},$$

where $a^2 = \left(\sum_{i=1}^{k} \sigma_i p_i \right)^2$ and $b^2 = \sum_{i=1}^{k} p_i (\bar{X}_i - \bar{X})^2$.

Since n and N are integers which minimize σ_F^2 it must follow that

$$\frac{a^2}{n + L_B} + \frac{b^2}{N - L_A} > \frac{a^2}{n} + \frac{b^2}{N} < \frac{a^2}{n - L_B} + \frac{b^2}{N + L_A},$$

from which it may be deduced that

$$\frac{1 + L_B/n}{1 - L_A/N} > \frac{a^2 . L_B}{n^2} \frac{N^2}{b^2 . L_A} > \frac{1 - L_B/n}{1 + L_A/N}.$$

We therefore take $n \propto a\sqrt{L_B}$ and $N \propto b\sqrt{L_A}$ and decide that the sum of money should be so divided that the samples S_1 and S_2 should be of numbers respectively

$$N = \frac{bC}{a\sqrt{(AB)} + bB} \quad \text{and} \quad n = \frac{aC}{b\sqrt{(AB)} + aA}.$$

REFERENCES AND READING

All the material included in this chapter may be found in J. Neyman, 'On two aspects of the Representative Method', *J. R. Statist. Soc.* XCVII, p. 558 and J. Neyman, 'Contribution to the theory of sampling human populations', *J. Amer. Statist. Assoc.* XXXIII, p. 101.

The problem of sampling human populations is an important one but in addition the student who works carefully through this chapter will find he has learnt a great deal about the application of the fundamental theorems on expectations.

For those who would wish to read further about the sampling problem there is a paper by F. Yates, 'A review of Recent Statistical Developments in Sampling and Sampling Surveys', *J. R. Statist. Soc.* CIX, p. 12, which discusses the problem from the practical angle and gives a useful list of references.

CHAPTER XV

CHARACTERISTIC FUNCTIONS. ELEMENTARY THEOREMS

The characteristic function of a random variable is a useful device for the calculation of theoretical moments and cumulants of probability laws and, by means of the inversion theorem, of the probability laws themselves. It is possible that its application will not lead to the solution of problems which could not have been solved by other methods, but it is elegant mathematically and for some types of problem considerably shortens the necessary calculations.

The theory of characteristic functions will be treated here in very elementary fashion, and it will not be possible to offer proofs for all the theorems. However, it is the application of these theorems in which the student will principally be interested. Such proofs as are omitted will be found in other texts.

DEFINITION. $\phi_x(t)$ is defined as the characteristic function of the random variable, x, or of the probability law of the random variable, x, if

$$\phi_x(t) = \mathscr{E}(e^{itx}).$$

This characteristic function will always exist since

$$|e^{itx}| = |(\cos^2 tx + \sin^2 tx)^{\frac{1}{2}}| = 1,$$

and it may be shown that there will be a 1:1 correspondence between a probability law and its characteristic function. The theorem, which is fundamental to the theory, we state without proof.

THEOREM. To any probability law, $p(x)$, there corresponds a uniquely defined characteristic function, and conversely. By definition

$$\phi_x(t) = \mathscr{E}(e^{itx}) = \sum_x p(x)\, e^{itx} \qquad \text{if the variable is discontinuous}$$

$$= \int p(x)\, e^{itx}\, dx \qquad \text{if the variable is continuous,}$$

the summation, and the integral sign being taken over the whole range of possible values of x.

Example. Find the characteristic function of the random variable, x, whose probability law is the binomial, i.e.

$$p_x(k) = \frac{n!}{k!(n-k)!} p^k q^{n-k}.$$

From the definition

$$\phi_x(t) = \mathscr{E}(e^{itx}) = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} p^x q^{n-x} e^{itx}$$

$$= \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} (pe^{it})^x q^{n-x} = (q+pe^{it})^n.$$

The properties of the characteristic function may be stated in the form of a set of theorems, the proof of which follows directly from the definition.

THEOREM. If a is a constant, then $\phi_{ax}(t) = \phi_x(at)$. For,

$$\phi_{ax}(t) = \mathscr{E}(e^{itax}) = \mathscr{E}(e^{i(at)x}) = \phi_x(at).$$

THEOREM. If $x_1, x_2, \ldots, x_n$ are random independent variables, then

$$\phi_{\sum_{j=1}^{n} x_j}(t) = \prod_{j=1}^{n} \phi_{x_j}(t).$$

For by definition

$$\phi_{\sum_{j=1}^{n} x_j}(t) = \mathscr{E}\left(\exp\left[it \sum_{j=1}^{n} x_j\right]\right) = \mathscr{E}\left(\prod_{j=1}^{n} e^{itx_j}\right).$$

Since the variables are independent the expectation of the product will equal the product of the expectations. Hence

$$\mathscr{E}\left(\prod_{j=1}^{n} e^{itx_j}\right) = \prod_{j=1}^{n} \mathscr{E}(e^{itx_j}) = \prod_{j=1}^{n} \phi_{x_j}(t)$$

and the theorem is proved.

THEOREM. If $a_1, a_2, \ldots, a_n$ are constants, and $x_1, x_2, \ldots, x_n$ are random independent variables, then

$$\phi_{\sum_{j=1}^{n} a_j x_j}(t) = \prod_{j=1}^{n} \phi_{x_j}(a_j t).$$

For $$\phi_{\sum_{j=1}^{n} a_j x_j}(t) = \prod_{j=1}^{n} \phi_{a_j x_j}(t) = \prod_{j=1}^{n} \phi_{x_j}(a_j t).$$

COROLLARY. If $a_1 = a_2 = \ldots = a_n = 1/n$, then

$$\phi_{\bar{x}}(t) = \prod_{j=1}^{n} \phi_{x_j}(t/n),$$

and further, if $x_1, x_2, \ldots, x_n$ all follow the same probability law,

$$\phi_{\bar{x}}(t) = (\phi_{x_j}(t/n))^n.$$

Example. Suppose that there are N random independent variables $x_1, x_2, \ldots, x_N$ which follow the same binomial law of probability

$$p_x(k) = \frac{n!}{k!(n-k)!} p^k q^{n-k}.$$

What is the characteristic function of their mean?

We have shown in a previous example that

$$\phi_x(t) = (q + p e^{it})^n$$

if x follows a binomial law as given above. If

$$\bar{x} = \frac{1}{N} \sum_{j=1}^{N} x_j,$$

then $$\phi_{\bar{x}}(t) = (\phi_x(t/N))^N = (q + p e^{it/N})^{nN},$$

which is the characteristic function of a variable following a generalized binomial probability law.

Example. Find the characteristic function of a variable whose probability law is normal.

It is given that

$$P\{a < x < \beta\} = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \int_{\alpha}^{\beta} \exp\left[-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2\right] dx$$

$$\text{for } -\infty < \alpha, \beta < +\infty.$$

From definition

$$\phi_x(t) = \mathscr{E}(e^{itx}) = \frac{1}{\sigma\sqrt{(2\pi)}} \int_{-\infty}^{+\infty} \exp\left[itx - \frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2\right] dx$$

$$= \frac{\exp\left[it\xi - \frac{1}{2}(t^2\sigma^2)\right]}{\sigma\sqrt{(2\pi)}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2\sigma^2}(x - (\xi + it\sigma^2))^2\right] dx.$$

Hence $$\phi_x(t) = \exp\left[it\xi - \tfrac{1}{2}(t^2\sigma^2)\right].$$

COROLLARY. If $\xi = 0$ and $\sigma = 1$, then

$$\phi_x(t) = e^{-\frac{1}{2}t^2}.$$

We may use the more general result to prove the following important theorem.

THEOREM. If $x_1, x_2, \ldots, x_n$ are independent random variables each following a probability law

$$P\{\alpha < x_j < \beta\} = \frac{1}{\sigma_j \sqrt{(2\pi)}} \int_\alpha^\beta \exp\left[-\frac{1}{2}\left(\frac{x_j - \xi_j}{\sigma_j}\right)^2\right] dx_j$$

$$\text{for } j = 1, 2, \ldots, n \quad \text{and} \quad -\infty < \alpha, \beta < +\infty,$$

then, whatever the numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$, the probability law of

$$y = \sum_{j=1}^{n} \lambda_j x_j$$

will be a normal distribution with mean $\sum_{j=1}^{n} \lambda_j \xi_j$ and variance

$$\sigma^2 = \sum_{j=1}^{n} \lambda_j^2 \sigma_j^2.$$

$$\phi_y(t) = \prod_{j=1}^{n} \phi_{x_j}(\lambda_j t) = \prod_{j=1}^{n} \exp\left[i\lambda_j t \xi_j - \frac{\lambda_j^2 t^2 \sigma_j^2}{2}\right]$$

$$= \exp\left[it \sum_{j=1}^{n} \lambda_j \xi_j - \frac{t^2}{2} \sum_{j=1}^{n} \lambda_j^2 \sigma_j^2\right].$$

The right-hand expression will be recognized as the characteristic function of a normal variate having mean $\sum_{j=1}^{n} \lambda_j \xi_j$ and variance $\sum_{j=1}^{n} \lambda_j^2 \sigma_j^2$ and the theorem is proved. It may be noted that, using elementary theorems on expectations, we could have shown that

$$\mathscr{E}(y) = \sum_{j=1}^{n} \lambda_j \xi_j \quad \text{and} \quad \mathscr{E}(y - \mathscr{E}(y))^2 = \sum_{j=1}^{n} \lambda_j^2 \sigma_j^2.$$

The characteristic function has enabled us here to take a step forward, for in addition to the mean and the variance we are able to specify the actual probability law.

Example. $x_1, x_2, \ldots, x_n$ are random independent variables each following a probability law

$$p(x_j) = \frac{e^{-\lambda_j} \lambda_j^{x_j}}{x_j!} \quad \text{for } j = 1, 2, \ldots, n.$$

$$x_j = 0, 1, 2, \ldots, +\infty.$$

What is the probability law of their sum?

$$\phi_x(t) = \mathscr{E}(e^{itx}) = \sum_{x=0}^{\infty} \frac{\lambda^x e^{itx-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x e^{itx}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^{it})^x}{x!}$$

$$= \exp[-\lambda + \lambda e^{it}].$$

This is true for any x when the appropriate subscripts are added. Hence for $\sum_{j=1}^{n} x_j$ we have

$$\phi_{\sum_{j=1}^{n} x_j}(t) = \prod_{j=1}^{n} \phi_{x_j}(t) = \prod_{j=1}^{n} \exp[-\lambda_j + \lambda_j e^{it}]$$

$$= \exp\left[(e^{it}-1) \sum_{j=1}^{n} \lambda_j\right].$$

It follows that the sum of a number of random variables, each of which is distributed as Poisson, is also a Poisson variable with probability law

$$p\left(\sum_{j=1}^{n} x_j\right) = \frac{\exp\left[-\sum_{j=1}^{n} \lambda_j\right]\left(\sum_{j=1}^{n} \lambda_j\right)^{\sum_{j=1}^{n} x_j}}{\left(\sum_{j=1}^{n} x_j\right)!}.$$

These examples illustrate the derivation of probability laws by using their characteristic functions. The sum of a number of binomial variates has been shown to be distributed according to the binomial law, a number of Poisson variates as Poisson, and a number of normal variates as a normal distribution.

Consider one further distribution, that is the distribution of the sum of a number of independent random variables

$$X_1, \quad X_2, \quad \ldots, \quad X_n,$$

when

$$X_j = x_j^2 \quad \text{and} \quad P\{a < x_j < b\} = \frac{1}{\sqrt{(2\pi)}} \int_a^b \exp[-\tfrac{1}{2}x_j^2]\, dx_j$$

$$\text{for } j = 1, 2, \ldots, n, \quad -\infty < a, b < +\infty.$$

Let us take first of all one typical random variable, x. If this is normally distributed then a simple substitution will show that $x^2 = X$ has the distribution

$$P\{A < X < B\} = \frac{1}{\sqrt{(2\pi)}} \int_A^B X^{-\frac{1}{2}} e^{-\frac{1}{2}X} dX \quad \text{for } 0 < X < +\infty.$$

The characteristic function of any X will be

$$\phi_X(t) = \frac{1}{\sqrt{(2\pi)}} \int_0^{+\infty} X^{-\frac{1}{2}} \exp\left[-\frac{X}{2}(1-2it)\right] dX = (1-2it)^{-\frac{1}{2}}.$$

It is desired to find the distribution of

$$\sum_{j=1}^{n} X_j = \sum_{j=1}^{n} x_j^2.$$

The characteristic function of $\sum_{j=1}^{n} X_j$ will be

$$\phi_{\sum_{j=1}^{n} X_j}(t) = [\phi_{X_j}(t)]^n = (1-2it)^{-\frac{1}{2}n}.$$

This will be recognized as being the characteristic function of a variable

$$\chi^2 = \sum_{j=1}^{n} X_j = \sum_{j=1}^{n} x_j^2,$$

where the distribution of χ^2 is

$$P\{\chi_1^2 < \chi^2 < \chi_2^2\} = \frac{1}{2^{\frac{1}{2}n}\Gamma(n/2)} \int_{\chi_1^2}^{\chi_2^2} (\chi^2)^{\frac{1}{2}n-1} e^{-\frac{1}{2}\chi^2} d(\chi^2)$$

$$\text{for } 0 < \chi^2 < +\infty$$

for

$$\phi_{\chi^2}(t) = \frac{1}{2^{\frac{1}{2}n}\Gamma(n/2)} \int_0^{+\infty} (\chi^2)^{\frac{1}{2}n-1} e^{-\frac{1}{2}\chi^2 + it\chi^2} d(\chi^2) = (1-2it)^{-\frac{1}{2}n}.$$

We may proceed from here to show that the sum of any number of independent χ^2 is also distributed as χ^2. For if

$$\phi_{\chi^2}(t) = (1-2it)^{-\frac{1}{2}n}$$

is the characteristic function of χ^2 distributed with n degrees of freedom, then the characteristic function of

$$Y = \sum_{k=1}^{s} \chi_k^2,$$

where χ_k^2 is distributed with n_k degrees of freedom, will be

$$\phi_Y(t) = \prod_{k=1}^{s} (1-2it)^{-\frac{1}{2}n_k} = (1-2it)^{-\frac{1}{2}\sum_{k=1}^{s} n_k}.$$

From a comparison of $\phi_{\chi^2}(t)$ and $\phi_Y(t)$, the probability law of Y will be

$$P\{A < Y < B\} = \frac{1}{2^{\frac{1}{2}\sum_{k=1}^{s} n_k}\,\Gamma\left(\frac{1}{2}\sum_{k=1}^{s} n_k\right)} \int_A^B Y^{\frac{1}{2}\sum_{k=1}^{s} n_k - 1} e^{-\frac{1}{2}Y}\, dY$$

$$\text{for } 0 < Y < +\infty,$$

which will be recognized as another χ^2 distribution. The reader should check this by writing down the characteristic function of Y from the distribution and comparing with the characteristic function already derived.

These examples are sufficient to show how by straightforward application of the definition of the characteristic function the distributions can be obtained of various combinations of random variables following given probability laws. We may now proceed further and discuss the application of the characteristic function to certain limit theorems which have already been proved earlier. In order to do this it will be necessary to make use of a theorem which we shall state without proof.

THEOREM. Let $p_1, p_2, \ldots, p_n, \ldots$ represent a sequence of probability laws and $\phi_1(t), \phi_2(t), \ldots, \phi_n(t), \ldots$ be their corresponding characteristic functions. If $\phi_n(t)$ tends to a limit, $\phi_0(t)$, uniformly in any finite interval, then p_n tends to a limit p_0 and the characteristic function of p_0 will be $\phi_0(t)$.

We can use this theorem to prove the theorem that the normal curve is the limiting distribution of the binomial as the n of the binomial law increases without limit (see Chapter v). We shall begin by defining a 'reduced' or 'standardized' variable.

DEFINITION. The random variable x with expectation equal to m and standard error equal to σ is said to have been 'reduced' or 'standardized' when it is referred to a zero mean with unit standard deviation; i.e. the reduced variable X is

$$X = \frac{x - m}{\sigma}.$$

The characteristic function of a reduced variable will be

$$\phi_{(x-m)/\sigma} = \mathscr{E}(e^{it[(x-m)/\sigma]}) = e^{-itm/\sigma}\,\mathscr{E}e^{itx/\sigma} = e^{-itm/\sigma}\phi_x(t/\sigma).$$

THEOREM. If k is a random variable distributed according to the binomial law, then whatever α

$$P\left\{\frac{k-np}{\sqrt{(npq)}}<\alpha\right\}\to\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\alpha}e^{-\frac{1}{2}x^2}dx \quad \text{as} \quad n\to\infty,$$

where

$$P\{k=k_1\}=\frac{n!}{k_1!(n-k_1)!}p^{k_1}q^{n-k_1} \quad \text{for } k=0,1,2,\ldots,n.$$

Let

$$X=\frac{k-np}{\sqrt{(npq)}}.$$

X is then a standardized binomial variable and there will be a sequence of probability laws $p_n(X), p_{n+1}(X), \ldots$ corresponding to increasing n. If it can be shown that the characteristic functions of these probability laws tend to a limit as $n\to\infty$, then by the previous theorem it may be assumed that the probability laws also tend to a limit and this limit will have as characteristic function the limit of the sequence of characteristic functions.

From previous definitions

$$\begin{aligned}\phi_X(t)&=\exp\left[-\frac{itnp}{\sqrt{(npq)}}\right]\phi_k\left(\frac{t}{\sqrt{(npq)}}\right)\\&=\left(p\exp\left[\frac{itq}{\sqrt{(npq)}}\right]+q\exp\left[-\frac{itp}{\sqrt{(npq)}}\right]\right)^n.\end{aligned}$$

The interior of the right-hand bracket may be expanded into two exponential series to give

$$\begin{aligned}\Phi&=\left(p\exp\left[\frac{itq}{\sqrt{(npq)}}\right]+q\exp\left[-\frac{itp}{\sqrt{(npq)}}\right]\right)\\&=p+q-\frac{t^2}{2!\,npq}(pq^2+qp^2)+\left(\frac{it}{\sqrt{(npq)}}\right)^3\\&\quad\times\frac{1}{3!}\left[pq^3\exp\left(\theta_1\frac{itq}{\sqrt{(npq)}}\right)-qp^3\exp\left(-\theta_2\frac{itp}{\sqrt{(npq)}}\right)\right],\end{aligned}$$

where $0<\theta_1,\theta_2<1$. Remembering $p+q=1$ this may be written

$$\Phi=1-\frac{t^2}{2n}+\frac{t^3.R}{n^{\frac{3}{2}}},$$

where

$$R = \frac{i^3}{3!(pq)^{\frac{3}{2}}}\left[pq^3 \exp\left(\theta_1 \frac{itq}{\sqrt{(npq)}}\right) - qp^3 \exp\left(-\theta_2 \frac{itp}{\sqrt{(npq)}}\right)\right].$$

As $n \to \infty$, $|R| < M$, where M is some fixed number arbitrarily chosen. The characteristic function $\phi_X(t)$ is

$$\phi_X(t) = \left(1 - \frac{t^2}{2n} + \frac{t^3 R}{n^{\frac{3}{2}}}\right)^n = \left(1 - \frac{t^2}{2n}\right)^n \left(1 + \frac{t^3 R}{n^{\frac{3}{2}}\left(1 - \frac{t^2}{2n}\right)}\right)^n.$$

We are now in a position to investigate the limit of $\phi_X(t)$ as $n \to \infty$. Consider first

$$z = \left(1 + \frac{t^3 R}{n^{\frac{3}{2}}\left(1 - \frac{t^2}{2n}\right)}\right)^n.$$

Take logarithms and expand the right-hand side as a series.

$$\log z = n\left[\frac{t^3 R}{n^{\frac{3}{2}}\left(1 - \frac{t^2}{2n}\right)} - \frac{t^6 R}{2n^3\left(1 - \frac{t^2}{2n}\right)^2} + \ldots\right].$$

This is a convergent series each term of which tends separately to zero, as $n \to \infty$. Hence $\log z$ tends to zero as n tends to infinity and therefore z tends to one as n tends to infinity. Now consider the term

$$y = \left(1 - \frac{t^2}{2n}\right)^n = \left(1 - \frac{t^2}{2n}\right)^{\frac{2n}{t^2}\frac{t^2}{2}}.$$

y tends to $\exp(-\frac{1}{2}t^2)$ as n increases without limit. It follows that

$$\lim_{n \to \infty} \phi_X(t) = e^{-\frac{1}{2}t^2}.$$

This is recognized as the characteristic function of a variable whose probability law is a normal distribution with zero mean and unit standard deviation. Hence if

$$p(k) = \frac{n!}{k!(n-k)!} p^k q^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n,$$

then whatever the value of α

$$P\left\{\frac{k-np}{\sqrt{(npq)}}<\alpha\right\}\to\frac{1}{\sqrt{(2\pi)}}\int_{-\infty}^{\alpha}e^{-\frac{1}{2}x^2}dx \quad \text{as} \quad n\to\infty.$$

We may now pass by similar analysis to two theorems each of which may be considered as a special case of an important theorem of Liapounoff. It is convenient to discuss these two further special cases in detail before passing to a simplified form of the generalized theorem which we shall prove in the next chapter. In each of these special cases we shall assume a theorem used implicitly in the last theorem, and it may be well therefore to state it explicitly.

THEOREM. If R_n is bounded, that is, if there exists a number which exceeds $|R_n|$ whatever n, then

$$\left(1+\frac{R_n}{n^{1+\delta}}\right)^n\to 1 \quad \text{as} \quad n\to\infty,$$

if $\delta>0$. Write $$u=\left(1+\frac{R_n}{n^{1+\delta}}\right)^n.$$

Then $$\log u=n\log\left(1+\frac{R_n}{n^{1+\delta}}\right)=n\left[\frac{R_n}{n^{1+\delta}}-\frac{1}{2}\frac{R_n^2}{n^{2(1+\delta)}}+\dots\right].$$

Each term of this series tends separately to zero as $n\to\infty$, provided $\delta>0$, since R_n is bounded. It follows therefore that $\log u\to 0$ as $n\to\infty$ and that $u\to 1$.

Laplace's theorem concerning the limiting distribution of a binomial variable was proved in an earlier chapter by straightforward analysis. The next two theorems could also be proved without reference to the characteristic function, but there is no doubt, as in Laplace's theorem, that considerably heavy algebra would be involved. Using the characteristic function limit theorem the proofs are comparatively simple. Let us consider an extreme case and show that under certain conditions the distribution of the sum of n standardized variables, each following Poisson's limit to the binomial law, tends to normality as n tends to infinity.

THEOREM. $x_1, x_2, \dots, x_n$ are random independent variables each having a probability law

$$p(x_k)=\frac{e^{-\lambda_k}\lambda_k^{x_k}}{x_k!} \quad \text{for } k=1,2,\dots,n \quad \text{and} \quad x_k=0,1,2,\dots,+\infty.$$

If $x = x_1 + x_2 + \ldots + x_n$, then, under certain conditions,

$$P\left\{\frac{x - \sum_{k=1}^{n} \lambda_k}{\left(\sum_{k=1}^{n} \lambda_k\right)^{\frac{1}{2}}} < \alpha\right\} \to \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\alpha} e^{-\frac{1}{2}t^2} dt \quad \text{as} \quad n \to \infty.$$

We have previously shown that if

$$p(x_k) = \frac{e^{-\lambda_k} \lambda_k^{x_k}}{x_k!} \quad \text{for } x_k = 0, 1, 2, \ldots, +\infty,$$

then $$\phi_{x_k}(t) = \exp\left[\lambda_k(e^{it} - 1)\right].$$

If $$x = x_1 + x_2 + \ldots + x_n,$$

then $$\phi_x(t) = \exp\left[(e^{it} - 1) \sum_{k=1}^{n} \lambda_k\right],$$

from which it may be deduced that the mean and variance of x are each equal to $\sum_{k=1}^{n} \lambda_k$. It follows that

$$Y = \frac{x - \sum_{k=1}^{n} \lambda_k}{\left(\sum_{k=1}^{n} \lambda_k\right)^{\frac{1}{2}}}$$

is a standardized Poisson variable. Write for convenience

$$\sum_{k=1}^{n} \lambda_k = \sigma_\lambda^2.$$

From first principles

$$\phi_Y(t) = \exp\left[\sigma_\lambda^2 e^{it/\sigma_\lambda} - \sigma_\lambda^2 - it\sigma_\lambda\right].$$

It is necessary here to distinguish between two cases.

Case I. As $n \to \infty$ it may happen that σ_λ^2 tends to a finite limit, i.e.

$$\sum_{k=1}^{n} \lambda_k = \sigma_\lambda^2 \to L < +\infty \quad \text{as} \quad n \to \infty.$$

If this is so, then

$$\lim_{n \to \infty} \phi_Y(t) = \exp\left[Le^{it/L^{\frac{1}{2}}} - L - itL^{\frac{1}{2}}\right],$$

which is recognized as being the characteristic function of a Poisson distribution with mean L. Hence the theorem cannot be true if $\sum_{k=1}^{n} \lambda_k$ has a finite limit as $n \to \infty$.

Case II. Assume that $\sum_{k=1}^{n} \lambda_k = \sigma_\lambda^2 \to +\infty$ as $n \to \infty$ and consider the logarithm of the characteristic function of Y.

$$\log \phi_Y(t) = -\frac{t^2}{2} + \frac{(it)^3}{3!\sigma_\lambda} e^{\theta it/\sigma_\lambda}, \quad \text{where } 0 < \theta < 1.$$

Hence, as $n \to \infty$, $\lim_{n \to \infty} \phi_Y(t) = e^{-\frac{1}{2}t^2}$

and the theorem follows. Accordingly the standardized sum of n independent Poisson variables will tend to be normally distributed as increases without limit provided that the sum of the means of the n variables tends to infinity with n.

THEOREM. The standardized mean of n random variables, each independently and rectangularly distributed, tends to be normally distributed as $n \to \infty$.

$x_1, x_2, \ldots, x_n$ are random independent variables which are rectangularly distributed. Assume therefore that

$$p(x_k) = \frac{1}{2a} \quad \text{for } -a \leqslant x_k \leqslant +a \text{ and zero elsewhere,}$$

$$\text{for } k = 1, 2, \ldots, n.$$

We begin by finding $\mathscr{E}(\bar{x})$ and $\sigma_{\bar{x}}$.

$$\mathscr{E}(\bar{x}) = \frac{1}{n} \sum_{j=1}^{n} \int_{-a}^{+a} x_j \frac{1}{2a} dx_j = 0,$$

$$\sigma_{\bar{x}}^2 = \mathscr{E} \frac{1}{n^2} \left(\sum_{j=1}^{n} x_j \right)^2 = \frac{1}{n^2} \sum_{j=1}^{n} \int_{-a}^{+a} \frac{x_j^2 dx_j}{2a} = \frac{a^2}{3n}.$$

If therefore Y is the standardized mean of the x's, then

$$Y = \frac{\bar{x} - \mathscr{E}(\bar{x})}{\sigma_{\bar{x}}} = \frac{\bar{x}}{a/\sqrt{3n}} = \frac{\sum_{j=1}^{n} x_j}{a\sqrt{\frac{1}{3}n}}.$$

For any x,

$$\phi_x(t) = \int_{-a}^{+a} e^{itx} \frac{1}{2a} dx = \frac{e^{ita} - e^{-ita}}{2ait} = \frac{\sin at}{at}.$$

Hence $$\phi_Y(t) = \left(\phi_{x_j} \left(\frac{t}{\sigma_{\Sigma x}} \right) \right)^n = \left(\frac{\sin t\sqrt{(3/n)}}{t\sqrt{(3/n)}} \right)^n.$$

The numerator may be expanded in a sine series and we have, for $0<\theta<1$,

$$\phi_Y(t) = \left(1-\frac{1}{3!}\left(t\sqrt{\frac{3}{n}}\right)^2+\frac{1}{5!}\left(t\sqrt{\frac{3}{n}}\right)^4\cos\left(\theta.t\sqrt{\frac{3}{n}}\right)\right)^n$$

$$= \left(1-\frac{t^2}{2n}\right)^n\left(1+\frac{9t^4}{5!\,n^2}\,\frac{\cos\left(\theta t\sqrt{(3/n)}\right)}{(1-(t^2/2n))}\right)^n.$$

By a previous theorem the right-hand bracket may be shown to tend to unity as n tends to infinity and

$$\left(1-\frac{t^2}{2n}\right)^n = \left(1-\frac{t^2}{2n}\right)^{\frac{2n}{t^2}\cdot\frac{t^2}{2}} \to e^{-\frac{1}{2}t^2} \quad \text{as} \quad n\to\infty.$$

We have therefore that the standardized mean of n variables, each of which is independently and rectangularly distributed, tends to be normally distributed as n is increased without limit.

REFERENCES AND READING

The standard text (in English) on characteristic functions in probability is H. Cramér, *Random Variables and Probability Distributions*, Cambridge Tracts in Mathematics, no. 36. This text has been out of print for some time and is not easily available. In any case a certain degree of mathematical knowledge is necessary in its reading.

In M. G. Kendall, *The Advanced Theory of Statistics*, the student will find the characteristic function used in a variety of ways. There is no elementary text which can be recommended as it is very difficult to develop characteristic function theory without making considerable use of the theory of functions.

CHAPTER XVI

CHARACTERISTIC FUNCTIONS: MOMENTS AND CUMULANTS. LIAPOUNOFF'S THEOREM

Before proceeding to discuss a generalization of the two theorems proved at the end of the last chapter, whereby it may be shown that under certain conditions the mean of the sum of n random variables of whatever distribution will tend to be normally distributed as n tends to infinity, it will be necessary to discuss certain properties of the characteristic function as a moment generating function as these properties will be needed for the proof of the theorem. We shall consider first of all the relation between the characteristic function and the moments and cumulants of a probability distribution.

$\phi_x(t)$, the characteristic function of the random variable, x, is defined as

$$\phi_x(t) = \int_{-\infty}^{+\infty} e^{itx} p(x)\,dx \quad \text{or} \quad \phi_x(t) = \sum_x e^{itx} p(x),$$

according as the variable is continuous or discontinuous. We now assume that in the neighbourhood of $t = 0$, $\phi_x(t)$ is differentiable with respect to t as often as desired, and we write for both continuous and discontinuous variables,

$$\phi_x(t) = \phi_x(0) + t\phi_x'(0) + \frac{t^2}{2!}\phi_x''(0) + \ldots.$$

We shall confine ourselves to discussing the case where the variable is continuous in that the situation is a little more complicated than for the discontinuous case, but the discussion for the discontinuous variable can be exactly paralleled by the reader substituting a summation for an integral sign. We shall consider the terms on the right-hand side of the expansion in turn.

$\phi_x(0)$. $\quad \phi_x(0) = \int_{-\infty}^{+\infty} p(x)\,dx = 1$ by definition.

$\phi_x'(0)$. $\quad \phi_x'(t) = \lim_{\delta t \to 0} \dfrac{\phi_x(t+\delta t) - \phi_x(t)}{\delta t}$

$$= \lim_{\delta t \to 0} \frac{1}{\delta t} \int_{-\infty}^{+\infty} p(x)\,(e^{i(t+\delta t)x} - e^{itx})\,dx.$$

Expand $e^{i(t+\delta t)x}$ in the following way:

$$e^{i(t+\delta t)x} = e^{itx} + \delta t(ix)\,e^{itx} + \frac{(\delta t)^2}{2!}(ix)^2 e^{i(t+\theta\delta t)x}, \text{ where } 0<\theta<1$$

and substitute into $\phi'_x(t)$.

$$\phi'_x(t) = \lim_{\delta t \to 0} \int_{-\infty}^{+\infty} p(x)\left[ix\,e^{itx} + \frac{\delta t}{2!}(ix)^2 e^{i(t+\theta\delta t)x}\right]dx$$

$$= i\int_{-\infty}^{+\infty} e^{itx}x\,.p(x)\,dx + \lim_{\delta t \to 0}\frac{\delta t}{2} i^2 \int_{-\infty}^{+\infty} e^{i(t+\theta\delta t)x}x^2\,.p(x)\,dx.$$

The first integral on the right-hand side will be finite provided

$$\int_{-\infty}^{+\infty} |\,x\,|\,p(x)\,dx$$

is finite, and the second provided the first two moments of x are finite. Let δt tend to zero and we have

$$\phi'_x(t) = \frac{\partial\phi(t)}{\partial t} = i\int_{-\infty}^{+\infty} e^{itx}x\,.p(x)\,dx$$

and at the poi t $t = 0$

$$\phi'_x(0) = i\int_{-\infty}^{+\infty} x\,.p(x)\,dx = im'_1.*$$

In a similar way it may be shown that

$$\phi''_x(t) = \frac{\partial^2\phi(t)}{\partial t^2} = i^2\int_{-\infty}^{+\infty} e^{itx}x^2\,.p(x)\,dx$$

and therefore that

$$\phi''_x(0) = i^2\int_{-\infty}^{+\infty} x^2\,.p(x)\,dx = i^2 m'_2.$$

A similar reasoning will give values for the higher derivatives. We have then

$$\phi_x(t) = 1 + it\,.m'_1 + \frac{(it)^2}{2!}m'_2 + \frac{(it)^3}{3!}m'_3 + \ldots,$$

that is, the numerical coefficient of

$$\frac{(it)^r}{r!} \quad r = 1, 2, \ldots$$

in the expansion of $\phi_x(t)$ in powers of t is the rth moment of the random variable x about an arbitrary origin.

* m is used instead of μ for ease of writing.

A simple extension of this theory gives the relationship between moments about an arbitrary origin and moments about the mean; for,

$$\phi_x(t) = \int_{-\infty}^{+\infty} e^{itx} p(x)\,dx = e^{itm_1'} \int_{-\infty}^{+\infty} e^{it(x-m_1')} dx = e^{itm_1'} \phi_{x-m_1'}(t)$$

or
$$\phi_x(t)\, e^{-itm_1'} = \phi_{x-m_1'}(t)$$

as proved previously. Expand each side in powers of t

$$\left(1 + it.m_1' + \frac{(it)^2}{2!} m_2' + \ldots\right)\left(1 - itm_1' + \frac{(it)^2}{2!} m_1'^2 - \ldots\right)$$
$$= \left(1 + \frac{(it)^2}{2!} m_2 + \frac{(it)^3}{3!} m_3 + \ldots\right)$$

and equate the coefficients of equal powers of t. We have that

$$m_2 = m_2' - m_1'^2,$$
$$m_3 = m_3' - 3m_2' m_1' + 2m_1'^3,$$
$$m_4 = m_4' - 4m_3' m_1' + 6m_2' m_1'^2 - 3m_1'^4$$

and so on, relations which are very familiar to the statistics student.

Example. Find the moments of the binomial variable, k, whose elementary probability law is

$$p(k) = \frac{n!}{k!(n-k)!} p^k q^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n.$$

We have already found the moments in two different ways. This third method is an improvement in labour on the first and is as quick to carry out as the second; thus

$$\phi_k(t) = (q + pe^{it})^n, \quad \phi_k'(0) = inp, \quad \phi_k''(0) = i^2(np - np^2 + n^2p^2)$$

and so on.

Example. Find the moments of the normal variate x whose integral probability law is

$$P\{\alpha < x < \beta\} = \frac{1}{\sigma\sqrt{2\pi}} \int_\alpha^\beta \exp\left[-\frac{x^2}{2\sigma^2}\right] dx \quad \text{for } -\infty < x < +\infty.$$

The characteristic function is

$$\phi_x(t) = \exp\left[-\frac{t^2\sigma^2}{2}\right],$$

whence

$$\phi'_x(0) = 0, \quad \phi''_x(0) = i^2\sigma^2, \quad \phi'''_x(0) = 0, \quad \phi^{\text{iv}}_x(0) = i^4 3\sigma^4, \text{ etc.}$$

Example. Find the moments of the variable, χ^2, whose integral probability law is

$$P\{\alpha < \chi^2 < \beta\} = \frac{1}{2^{\frac{1}{2}n}\Gamma(\frac{1}{2}n)} \int_\alpha^\beta (\chi^2)^{\frac{1}{2}n-1} e^{-\frac{1}{2}\chi^2} d(\chi^2) \quad \text{for } 0 < \chi^2 < +\infty.$$

In a previous chapter it was found for χ^2 distributed in this way that the characteristic function is

$$\phi_{\chi^2}(t) = (1 - 2it)^{-\frac{1}{2}n}.$$

Differentiating with respect to t and putting $t = 0$ we have

$$\phi'_{\chi^2}(0) = in, \quad \phi''_{\chi^2}(0) = i^2 n(n+2),$$

and generally

$$\frac{\partial^r \phi_{\chi^2}(0)}{\partial t^r} = i^r n(n+2)\dots(n+2(r-1)).$$

It follows that

$$m'_1 = n, \quad m_2 = 2n, \quad m_3 = 8n, \quad m_4 = 12n(n+4).$$

We may now carry the theory a stage further. Assume that there are n random independent variables $x_1, x_2, \dots, x_n$. It follows from definition that

$$\phi_{\sum_{j=1}^n x_j}(t) = \phi_{x_1}(t)\phi_{x_2}(t)\dots\phi_{x_n}(t).$$

Hence if we define another function, sometimes spoken of as the cumulative function, as

$$\psi_x(t) = \log \phi_x(t)$$

it will follow that

$$\psi_{\sum_{j=1}^n x_j}(t) = \psi_{x_1}(t) + \psi_{x_2}(t) + \dots + \psi_{x_n}(t).$$

Thus, whatever the distributions of the random variables, provided that they are independent, their cumulative functions will be additive. We may use the definition of the cumulative function to define the cumulants of a distribution. These cumulants are the same as the semi-invariants of Thiele. Consider $\psi_x(t)$ and assume, as for $\phi_x(t)$, that it is differentiable as often as desired in the neighbourhood of $t = 0$, i.e. assume that we may write

$$\psi_x(t) = \psi_x(0) + t\psi'_x(0) + \frac{t^2}{2!}\psi''_x(0) + \dots.$$

Following closely the previous work it may be shown, by differentiating the logarithm of the characteristic function the required number of times, that

$$\psi_x(t) = m_1'(it) + m_2\frac{(it)^2}{2!} + m_3\frac{(it)^3}{3!} + (m_4 - 3m_2^2)\frac{(it)^4}{4!} + \ldots.$$

Writing formally

$$\psi_x(t) = \kappa_1(it) + \kappa_2\frac{(it)^2}{2!} + \kappa_3\frac{(it)^3}{3!} + \kappa_4\frac{(it)^4}{4!} + \ldots,$$

where $\kappa_1, \kappa_2, \ldots$ are called the cumulants of the variable x, it will be seen that

$$\kappa_1 = m_1', \quad \kappa_3 = m_3, \qquad \kappa_5 = m_5 - 10m_3m_2,$$

$$\kappa_2 = m_2, \quad \kappa_4 = m_4 - 3m_2^2, \quad \kappa_6 = m_6 - 15m_4m_2 - 10m_3^2 + 30m_2^3.$$

These may be written the other way round to give

$$m_4 = \kappa_4 + 3\kappa_2^2, \; m_5 = \kappa_5 + 10\kappa_3\kappa_2, \; m_6 = \kappa_6 + 15\kappa_4\kappa_2 + 10\kappa_3^2 + 15\kappa_2^3.$$

It is also sometimes useful to be able to express moments about an arbitrary origin in terms of these cumulants, viz.

$$m_1' = \kappa_1,$$

$$m_2' = \kappa_2 + \kappa_1^2,$$

$$m_3' = \kappa_3 + 3\kappa_2\kappa_1 + \kappa_1^3,$$

$$m_4' = \kappa_4 + 4\kappa_3\kappa_1 + 3\kappa_2^2 + 6\kappa_2\kappa_1^2 + \kappa_1^4,$$

$$m_5' = \kappa_5 + 3\kappa_4\kappa_1 + 10\kappa_3\kappa_2 + 10\kappa_3\kappa_1^2 + 15\kappa_2^2\kappa_1 + 10\kappa_2\kappa_1^3 + \kappa_1^5,$$

$$m_6' = \kappa_6 + 6\kappa_5\kappa_1 + 15\kappa_4\kappa_2 + 15\kappa_4\kappa_1^2 + 10\kappa_3^2 + 60\kappa_3\kappa_2\kappa_1 + 20\kappa_3\kappa_1^3 + 15\kappa_2^3 + 45\kappa_2^2\kappa_1^2 + 15\kappa_2\kappa_1^4 + \kappa_1^6.$$

These may be checked by substituting moments about the mean for moments about an arbitrary origin. Now if

$$\psi_{x_1}(t) = \log\phi_{x_1}(t) = \kappa_{11}(it) + \kappa_{21}\frac{(it)^2}{2!} + \kappa_{31}\frac{(it)^3}{3!} + \ldots,$$

$$\psi_{x_2}(t) = \log\phi_{x_2}(t) = \kappa_{12}(it) + \kappa_{22}\frac{(it)^2}{2!} + \kappa_{32}\frac{(it)^3}{3!} + \ldots,$$

then the cumulants of the distribution of x_1+x_2 will be given by the addition of the cumulants of the separate distributions, always provided x_1 and x_2 are independent.

$$\begin{aligned}\psi_{x_1+x_2}(t) &= \log \phi_{x_1}(t) + \log \phi_{x_2}(t) \\ &= it(\kappa_{11}+\kappa_{21}) + \frac{(it)^2}{2!}(\kappa_{21}+\kappa_{22}) + \frac{(it)^3}{3!}(\kappa_{31}+\kappa_{32}) + \dots \\ &= it\bar{\kappa}_1 + \frac{(it)^2}{2!}\bar{\kappa}_2 + \frac{(it)^3}{3!}\bar{\kappa}_3 + \dots .\end{aligned}$$

It is this additive property of the cumulants which makes them of great use.

Example. Find Sheppard's corrections for moments calculated from grouped data.

One of the first things studied by the reader in statistics is the grouping of observations and the correction of moments calculated from these grouped observations for the effect of grouping. If X_E be the true value of a variate, X_G the central value of the group extending from $X_G-\frac{1}{2}h$ to $X_G+\frac{1}{2}h$, and x the error introduced by grouping, then assuming independence

$$X_G - X_E = x \quad \text{or} \quad X_E = X_G - x,$$

$$\phi_{X_E+x}(t) = \phi_{X_E}(t)\phi_x(t),$$

and

$$\psi_{X_E+x}(t) = \psi_{X_E}(t) + \psi_x(t).$$

Hence, if $\kappa_1(G), \kappa_2(G), \dots, \kappa_r(G), \dots$ are the cumulants of X_G, $\kappa_1(E), \kappa_2(E), \dots \kappa_r(E), \dots$ are the cumulants of X_E and $\kappa_1(x), \kappa_2(x), \dots, \kappa_r(x), \dots$ are the cumulants of x, we shall have

$$\kappa_r(G) = \kappa_r(E) + \kappa_r(x) \quad \text{and} \quad \kappa_r(E) = \kappa_r(G) - \kappa_r(x).$$

The cumulants of the grouped observations will be calculated from the data. It remains to consider the cumulants of the grouping error, x, and in order to do this it will be necessary to make assumptions about its distribution. There are many which may be made but we shall only take the simplest, i.e. that x is equally likely to take any value between $-\frac{1}{2}h$ and $+\frac{1}{2}h$. This is equivalent to saying that it is assumed that the integral probability law of x is

$$F(x) = P\{\alpha < x < \beta\} = \frac{1}{h}\int_\alpha^\beta dx \quad \text{for } -\tfrac{1}{2}h < x < +\tfrac{1}{2}h.$$

The characteristic function will be

$$\phi_x(t) = \frac{1}{h}\int_{-\frac{1}{2}h}^{+\frac{1}{2}h} e^{itx}dx = \frac{\sin \frac{1}{2}ht}{\frac{1}{2}ht}.$$

Taking the logarithm of $\phi_x(t)$ and expanding, as in previous examples, we have

$$\psi_x(t) = \frac{h^2}{12}\frac{(it)^2}{2!} - \frac{h^4}{120}\frac{(it)^4}{4!} + \frac{h^6}{252}\frac{(it)^6}{6!} - \ldots.$$

Hence $$\kappa_2(x) = m_2 = \frac{h^2}{12}, \quad \kappa_4(x) = m_4 - 3m_2^2 = -\frac{h^4}{120},$$

$$\kappa_6(x) = m_6 - 15m_4m_2 - 10m_3^2 + 30m_2^3 = \frac{h^6}{252}.$$

If we write $\mu_2(G)$, $\mu_3(G), \ldots$ for the moments of the grouped observations about the mean then from the relationship

$$\kappa_r(E) = \kappa_r(G) - \kappa_r(x)$$

it is found on substitution that

$$\kappa_2(E) = \kappa_2(G) - \kappa_2(x) = \mu_2(G) - h^2/12,$$
$$\kappa_3(E) = \kappa_3(G) - \kappa_3(x) = \mu_3(G),$$
$$\kappa_4(E) = \kappa_4(G) - \kappa_4(x) = \mu_4(G) - 3\mu_2^2(G) + h^4/120,$$

and the corrected moments of the distribution will be

$$\mu_2(E) = \mu_2(G) - h^2/12,$$
$$\mu_3(E) = \mu_3(G),$$
$$\mu_4(E) = \mu_4(G) - (\tfrac{1}{2}h^2)\,\mu_2(G) + \tfrac{7}{240}h^4.$$

The corrections for the higher even moments may be calculated similarly. There will be no corrections for the odd moments as might be expected from the assumption regarding the distribution of the grouping error.

Example. Find the cumulants of the binomial variate, k, whose elementary probability law is

$$p(k) = \frac{n!}{k!(n-k)!}p^kq^{n-k}.$$

It is known that $\phi_k(t) = (q+pe^{it})^n$

or, for k referred to np as origin,

$$\phi_x(t) = \phi_{k-np}(t) = (qe^{-ipt}+pe^{itq})^n.$$

Hence $\psi_x(t) = n \log (qe^{-ipt}+pe^{itq})$

and the cumulants of the binomial are obtained by successive differentiation.

$$\kappa_1 = 0, \quad \kappa_3 = -npq(p-q), \quad \kappa_5 = -npq(p-q)(1-12npq),$$

$$\kappa_2 = npq, \quad \kappa_4 = npq-6n^2p^2q^2, \quad \kappa_6 = npq-30n^2p^2q^2+120n^3p^3q^3.$$

Exercise. Find a recurrence formula for the binomial cumulants.

Exercise. Find the moments and cumulants of a random variable whose probability law is Neyman's contagious distribution.

Example. Find the cumulants of the continuous variable χ^2 whose integral probability law is

$$P\{\alpha < \chi^2 < \beta\} = \frac{1}{2^{n/2}\Gamma(\frac{1}{2}n)} \int_\alpha^\beta (\chi^2)^{\frac{1}{2}n-1} e^{-\frac{1}{2}\chi^2} d(\chi^2) \quad \text{for } 0 < \chi^2 < +\infty.$$

The characteristic function of the variable χ^2 has been shown to be

$$\phi_{\chi^2}(t) = (1-2it)^{-\frac{1}{2}n}.$$

Hence $\psi_{\chi^2}(t) = \log \phi_{\chi^2}(t) = -\frac{1}{2}n \log (1-2it).$

By successive differentiation (and putting t equal zero), or by expanding in powers of t, it may be shown that

$$\kappa_1 = n, \quad \kappa_2 = 2n, \quad \kappa_3 = 8n, \quad \kappa_4 = 48n, \quad \kappa_5 = 384n,$$

and generally that $\kappa_r = 2^{r-1}(r-1)!\, n.$

The moments of χ^2 are easily obtained from the relation between cumulants and moments. If there are s independent variables each distributed as χ^2 with integral probability law

$$P\{\alpha < \chi_j^2 < \beta\} = \frac{1}{2^{\frac{1}{2}n_j}\Gamma(\frac{1}{2}n_j)} \int_\alpha^\beta (\chi_j^2)^{\frac{1}{2}n_j-1} e^{-\frac{1}{2}\chi_j^2} d(\chi_j^2)$$

$$\text{for } 0 < \chi_j^2 < +\infty \quad \text{and} \quad j = 1, 2, \ldots, s,$$

then the cumulants of the distribution of their sum, i.e.

$$\chi^2 = \chi_1^2 + \chi_2^2 + \ldots + \chi_s^2$$

will be given by

$$\bar{\kappa}_r = \sum_{j=1}^{s} \kappa_r(j) = 2^{r-1}(r-1)! \sum_{j=1}^{s} n_j,$$

where r takes values $1, 2, 3, \ldots$ to give $\bar{\kappa}_1, \bar{\kappa}_2, \bar{\kappa}_3, \ldots$.

The relation between the characteristic function and the moments of a random variable will be useful to prove a simplified form of Liapounoff's theorem, specialized cases of which have been discussed in the preceding chapter.

Liapounoff's Theorem (simplified)

(1) If $x_1, x_2, \ldots, x_n$ are mutually independent random variables.

(2) If $x_1, x_2, \ldots, x_n$ each possesses the first three absolute moments,

$$\beta_{1k}, \quad \beta_{2k}, \quad \beta_{3k} \quad (k = 1, 2, \ldots, n),$$

where

$$\beta_{ak} = \int_{-\infty}^{+\infty} | x_k - \mathscr{E}(x_k) |^a p(x_k)\, dx_k$$

or

$$\beta_{ak} = \sum_{x_k} | x_k - \mathscr{E}(x_k) |^a p(x_k) \quad \text{for } a = 1, 2, 3,$$

according as the variables are continuous or discontinuous.

(3) If the second and third moments are each bounded, i.e. if there exist two pairs of numbers

$$m_2 \leqslant M_2 \quad \text{and} \quad m_3 \leqslant M_3$$

such that

$$0 \leqslant m_2 \leqslant \beta_{2k} \leqslant M_2 \quad \text{and} \quad m_3 \leqslant \beta_{3k} \leqslant M_3,$$

then the standardized sum of the x's tends to be normally distributed as n tends to infinity.

If we write

$$\sigma^2 = \sum_{k=1}^{n} \sigma_k^2,$$

then x, the standardized sum of the x's may be written

$$x = \frac{\sum_{k=1}^{n} x_k - \sum_{k=1}^{n} {}_1\mu'_k}{\sigma},$$

where ${}_1\mu'_k$ denotes the mean of the variable x_k, and σ_k its standard error (i.e. $\sqrt{{}_2\mu_k}$).

Let

$$\xi_k = x_k - {}_1\mu'_k.$$

Then $\mathscr{E}(\xi_k) = 0, \quad \mathscr{E}(\xi_k^2) = \sigma_k^2 = {}_2\mu_k, \quad \mathscr{E}(\xi_k^3) = {}_3\mu_k,$

and clearly, provided the moments of x_k exist, the moments of ξ_k will exist.

The characteristic function of x will be

$$\phi_x(t) = \mathscr{E}\left(\exp\left[\frac{it}{\sigma}\left(\sum_{k=1}^{n} x_k - \sum_{k=1}^{n} {}_1\mu_k'\right)\right]\right)$$

$$= \prod_{k=1}^{n} \mathscr{E} \exp\left[\frac{it\xi_k}{\sigma}\right] = \prod_{k=1}^{n} \phi_{\xi_k}\left(\frac{t}{\sigma}\right).$$

Expand $\phi_{\xi_k}(t)$ in powers of t. The mean value of ξ_k is zero and we have therefore

$$\phi_{\xi_k}(t) = 1 + \frac{(it)^2}{2!}\sigma_k^2 + \frac{t^3}{3!}\frac{\partial^3\phi(0t)}{\partial t^3}, \quad \text{where } 0 < \theta < 1.$$

Generally

$$\frac{\partial^s \phi_\xi(t)}{\partial t^s} = i^s \int_{-\infty}^{+\infty} \xi^s e^{it\xi} p(\xi)\, d\xi \quad \text{or} \quad \frac{\partial^s \phi_\xi(t)}{\partial t^s} = i^s \sum_{\xi} \xi^s e^{it\xi} p(\xi),$$

and hence

$$\phi_{\xi_k}(t) = 1 + \frac{(it)^2}{2!}\sigma_k^2 + \frac{(it)^3}{3!}\int_{-\infty}^{+\infty} \xi_k^3 e^{it\theta\xi_k} p(\xi_k)\, d\xi_k$$

for the continuous variable. For the discontinuous variable the summation sign will replace the integral sign. It is clear that

$$\left|\int_{-\infty}^{+\infty} \xi_k^3 e^{it\theta\xi_k} p(\xi_k)\, d\xi_k\right| \leqslant \int_{-\infty}^{+\infty} |\xi_k|^3 p(\xi_k)\, d\xi_k.$$

The right-hand side of this inequality is finite; for whatever k there exists two numbers m_3 and M_3 such that

$$m_3 \leqslant \int_{-\infty}^{+\infty} |\xi_k|^3 p(\xi_k)\, d\xi_k \leqslant M_3.$$

We may now proceed to a consideration of $\phi_x(t)$. Write

$$R_k = \int_{-\infty}^{+\infty} \xi_k^3 e^{it\theta\xi_k} p(\xi_k)\, d\xi_k \quad \text{or} \quad \sum_{\xi_k} \xi_k^3 e^{it\theta\xi_k} p(\xi_k)$$

as required. Then

$$\phi_x(t) = \prod_{k=1}^{n} \left(1 + \frac{(it)^2}{2!}\frac{\sigma_k^2}{\sigma^2} + \frac{(it)^3}{3!}\frac{R_k}{\sigma^3}\right)$$

or, if we take logarithms,

$$\psi_x(t) = \sum_{k=1}^{n} \log\left(1 - \frac{t^2}{2!}\frac{\sigma_k^2}{\sigma^2} - \frac{it^3}{3!}\frac{R_k}{\sigma^3}\right).$$

Expand in the usual way

$$\psi_x(t) = \sum_{k=1}^{n} \left[-\left(\frac{t^2}{2!}\frac{\sigma_k^2}{\sigma^2} + \frac{it^3}{3!}\frac{R_k}{\sigma^3}\right) - \frac{1}{2}\left(\frac{t^2}{2!}\frac{\sigma_k^2}{\sigma^2} + \frac{it^3}{3!}\frac{R_k}{\sigma^3}\right)^2 \left(1\Big/\left(1-\epsilon\left(\frac{t^2}{2!}\frac{\sigma_k^2}{\sigma^2} + \frac{it^3}{3!}\frac{R_k}{\sigma^3}\right)\right)\right)^2\right],$$

where $0 < \epsilon < 1$. Writing

$$Q_k = \left(1\Big/\left(1-\epsilon\left(\frac{t^2}{2!}\frac{\sigma_k^2}{\sigma^2} + \frac{it^3}{3!}\frac{R_k}{\sigma^3}\right)\right)\right)^2,$$

$\psi_x(t)$ may be written as

$$\psi_x(t) = -\frac{t^2}{2} - \frac{it^3}{6}\sum_{k=1}^{n}\frac{R_k}{\sigma^3} - \frac{t^4}{8}\sum_{k=1}^{n}\frac{\sigma_k^4 Q_k}{\sigma^4} - \frac{it^5}{12}\sum_{k=1}^{n}\frac{\sigma_k^2}{\sigma^5}R_k Q_k + \frac{t^6}{72}\sum_{k=1}^{n}\frac{R_k^2 Q_k}{\sigma^6}.$$

We have now to consider the behaviour of each of these terms as n is increased without limit.

(1) $\sum_{k=1}^{n}\frac{R_k}{\sigma^3}$. It has been shown that

$$|R_k| \leqslant M_3$$

and therefore

$$\left|\sum_{k=1}^{n} R_k\right| \leqslant \sum_{k=1}^{n}|R_k| \leqslant nM_3.$$

Similarly

$$\sigma^2 = \sum_{k=1}^{n}\sigma_k^2 \geqslant nm_2.$$

Hence

$$\frac{\left|\sum_{k=1}^{n} R_k\right|}{\sigma^3} \leqslant \frac{nM_3}{n^{\frac{3}{2}}m_2^{\frac{3}{2}}}.$$

M_3 and m_2 are fixed numbers. It follows that as n tends to infinity the right-hand side of this inequality tends to zero, and therefore that

$$\lim_{n\to\infty}\frac{\sum_{k=1}^{n} R_k}{\sigma^3} = 0.$$

(2) Q_k. Q_k was defined as

$$Q_k = \left(1\Big/\left(1-\epsilon\left(\frac{t^2}{2!}\frac{\sigma_k^2}{\sigma^2} + \frac{it^3}{3!}\frac{R_k}{\sigma^3}\right)\right)\right)^2,$$

where $0 < \epsilon < 1$.

We have investigated the behaviour of the term involving R_k. We must now find

$$\lim_{n\to\infty} \frac{\sigma_k^2}{\sigma^2}.$$

From definition

$$0 < \frac{\sigma_k^2}{\sigma^2} < \frac{M_2}{nm_2}$$

and hence the required limit as n tends to infinity will be zero. It follows that

$$\lim_{n\to\infty} Q_k = 1.$$

It may therefore be shown that as n increases without limit each of the terms except the first in the expansion of $\psi_x(t)$ tends separately to zero, or that

$$\psi_x(t) \to -\tfrac{1}{2}t^2 \quad \text{as} \quad n\to\infty$$

and

$$\phi_x(t) \to e^{-\frac{1}{2}t^2} \quad \text{as} \quad n\to\infty.$$

Hence under the given restrictions of the theorem

$$P\left\{\alpha \leqslant \frac{\sum_{k=1}^{n} x_k - \sum_{k=1}^{n} {}_1\mu'_k}{\left(\sum_{k=1}^{n} \sigma_k^2\right)^{\frac{1}{2}}} \leqslant \beta\right\} \to \frac{1}{\sqrt{(2\pi)}} \int_\alpha^\beta e^{-\frac{1}{2}x^2}\,dx \quad \text{as} \quad n\to\infty.$$

This is an extremely powerful result. We have shown that, subject to the random variables being independent and possessing the first three absolute moments, their standardized sum will tend to be normally distributed as the number of variables is increased no matter what the distributions of the variables. Moreover, although we do not discuss it here, all the conditions of the theorem need not necessarily be satisfied; under certain conditions the variables need not be independent and it is sufficient to assume the absolute moments of order $2+\delta$ exist, where δ is some number greater than zero.

The generalized theorem of Liapounoff may be proved in several different ways. Perhaps one of the simplest methods of proof is by using Liapounoff's inequality for moments. We shall state both the moment inequality and the generalized theorem but will refer the reader for proof to any of the treatises on the calculus of probability.

Liapounoff's Inequality for Moments

The absolute moment of order k of a random variable, x, is defined as

$$\beta_k = \int_{-\infty}^{+\infty} |x - \mathscr{E}(x)|^k p(x)\,dx \quad \text{or} \quad \beta_k = \sum_x |x_k - \mathscr{E}(x)|^k p(x),$$

according as the variable is continuous or discontinuous.

If a, b and c are real numbers such that

$$a \geqslant b \geqslant c \geqslant 0,$$

then

$$\beta_b^{a-c} \leqslant \beta_c^{a-b} \beta_a^{b-c}.$$

The proof follows directly from repeated applications of Cauchy's inequality.

Liapounoff's Theorem

If (1) $x_1, x_2, \ldots, x_n$ are random independent variables with zero means,

(2) $x_1, x_2, \ldots, x_n$ each possess absolute moments of order $2+\delta$, where $\delta > 0$,

(3) σ is the standard deviation of $\sum_{k=1}^{n} x_k$,

(4) the ratio $\dfrac{{}_1\beta_{2+\delta} + {}_2\beta_{2+\delta} + \ldots + {}_n\beta_{2+\delta}}{\sigma^{2+\delta}}$ tends to zero as n tends to infinity,

then the standardized sum of the x's tends to be normally distributed as n tends to infinity.

A large number of distributions may be found in statistics which will satisfy the conditions of the theorem. If it is known that observations have been independently and randomly drawn from a population, the moments of which satisfy the theorem, then as the size of the sample is increased the standardized mean of the sample will tend to be normally distributed. The theorem has therefore a wide field of application in statistical theory, possibly wider than any other single theorem.

REFERENCES AND READING

The relation between the characteristic function and the moments and cumulants of a variable is discussed in many texts. The development as given by Thiele is set out in Arne Fisher, *Mathematical Theory of Probability.*

A more modern treatment may be found in E. C. Cornish and R. A. Fisher, *Moments and Cumulants in the Specification of Distributions,* and again in M. G. Kendall, *The Advanced Theory of Statistics,* but these references are not by any means exhaustive.

The generalized theorem of Liapounoff is set out clearly in J. V. Uspensky, *Introduction to Mathematical Probability.*

CHAPTER XVII

CHARACTERISTIC FUNCTIONS. CONVERSE THEOREMS

In the previous two chapters we have been concerned with the derivation of the characteristic function of a variable from its probability law and the proof of various theorems by means of the characteristic function. In the proof of the theorems it has been usual to derive a limiting characteristic function which is then recognized as being the characteristic function of a given probability law. In most of the cases where the elementary theory of this treatise is applicable this procedure is adequate, but it cannot have escaped attention that there may be occasions when the characteristic function cannot be recognized as belonging to any known probability law. The probability law of any random variable, x, may be calculated, if its characteristic function is known, by means of known theorems. We shall state and prove the theorem when the variable is discontinuous, and state the theorem without proof when the variable is continuous.

THEOREM. x is a discontinuous random variable which may take only zero or positive integral values. If $p_x(k) = P\{x = k\}$ is the elementary probability law of x, and $\phi_x(t)$ its characteristic function, then

$$p_x(k) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \phi_x(t)\, e^{-itk}\, dt.$$

By definition $$\phi_x(t) = \mathscr{E}(e^{itx}) = \sum_{x=0}^{\infty} e^{itx} p(x).$$

We shall consider $$\frac{1}{2\pi}\int_{-\pi}^{+\pi} \phi_x(t)\, e^{-itk}\, dt$$

and prove that it is equal to $p_x(k)$.

$$\frac{1}{2\pi}\int_{-\pi}^{+\pi} \phi_x(t)\, e^{-itk} dt = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \sum_{x=0}^{\infty} p(x)\, e^{it(x-k)}\, dt.$$

Since the series is uniformly convergent with respect to t the

summation and integral signs may be transposed and we have

$$\frac{1}{2\pi}\int_{-\pi}^{+\pi}\sum_{x=0}^{\infty}p(x)\,e^{it(x-k)}dt=\frac{1}{2\pi}\sum_{x=0}^{\infty}p(x)\int_{-\pi}^{+\pi}e^{it(x-k)}dt$$

$$=\frac{1}{2\pi}\left(\sum_{x=0}^{k-1}p(x)\int_{-\pi}^{+\pi}e^{it(x-k)}dt\right.$$

$$\left.+p_x(k)\int_{-\pi}^{+\pi}dt+\sum_{x=k+1}^{\infty}p(x)\int_{-\pi}^{+\pi}e^{it(x-k)}dt\right).$$

The first and third integrals vanish and we have

$$\frac{1}{2\pi}\int_{-\pi}^{+\pi}\phi_x(t)\,e^{-itk}dt=p_x(k).$$

THEOREM. x is an absolutely continuous random variable the probability law of which is $p(x)$. If the characteristic function corresponding to $p(x)$ is $\phi_x(t)$, then

$$p(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\phi_x(t)\,e^{-itx}dt.$$

Example. Given

$$\phi_x(t)=\exp\left[-\lambda(1-e^{it})\right]$$

and that x may only take zero and positive integer values $0, 1, 2, \ldots$, find the probability law of x.

$$p_x(k)=\frac{1}{2\pi}\int_{-\pi}^{+\pi}e^{-\lambda(1-e^{it})-itk}dt=\frac{e^{-\lambda}}{2\pi}\int_{-\pi}^{+\pi}\left(e^{-itk}\sum_{r=0}^{\infty}\frac{\lambda^r e^{irt}}{r!}\right)dt.$$

The series is uniformly convergent with respect to t and we may therefore write

$$p_x(k)=\frac{e^{-\lambda}}{2\pi}\sum_{r=0}^{\infty}\frac{\lambda^r}{r!}\int_{-\pi}^{+\pi}e^{it(r-k)}dt$$

$$=\frac{e^{-\lambda}}{2\pi}\left[\sum_{r=0}^{k-1}\frac{\lambda^r}{r!}\int_{-\pi}^{+\pi}e^{it(r-k)}dt+\frac{\lambda^k}{k!}\int_{-\pi}^{+\pi}dt+\sum_{r=k+1}^{\infty}\frac{\lambda^r}{r!}\int_{-\pi}^{+\pi}e^{it(r-k)}dt\right],$$

giving

$$p_x(k)=\frac{e^{-\lambda}\lambda^k}{k!},$$

which is recognized to be the elementary probability law of Poisson's binomial limit.

Example. Given

$$\phi_x(t) = \exp\left[-\tfrac{1}{2}t^2\sigma^2 + it\xi\right]$$

and that x is a continuous random variable which may take any values between $-\infty$ and $+\infty$, find the elementary probability law of x.

$$p(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\phi_x(t)\,e^{-itx}\,dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\exp\left[-\tfrac{1}{2}t^2\sigma^2 - it(x-\xi)\right]dt.$$

Complete the square in the exponent.

$$p(x) = \frac{1}{2\pi}\exp\left[-\frac{1}{2\sigma^2}(x-\xi)^2\right]\int_{-\infty}^{+\infty}\exp\left[-\frac{\sigma^2}{2}\left(t+\frac{i}{\sigma^2}(x-\xi)\right)^2\right]dt$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(x-\xi)^2\right].$$

Thus the elementary probability law of x is a normal distribution as will already have been recognized from the form of the characteristic function.

Example. Given that x is a continuous variable which may take any values between $-\infty$ and $+\infty$ and that

$$\phi_x(t) = e^{-a|t|},$$

where $a > 0$, find the elementary probability law of x.

$$p(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\phi_x(t)\,e^{-itx}\,dt = \frac{1}{2\pi}\int_{-\infty}^{0}e^{-t(ix-a)}\,dt + \frac{1}{2\pi}\int_{0}^{+\infty}e^{-t(ix+a)}\,dt$$

$$= \frac{1}{2\pi}\left[\frac{1}{a-ix} + \frac{1}{a+ix}\right] = \frac{a}{\pi(a^2+x^2)}.$$

The result of this last example should be remembered. It is comparatively easy to obtain the probability law from a knowledge of the characteristic function but it is not so simple to proceed in the reverse direction without a knowledge of contour integration. For the reader who is not familiar with this kind of integration it is legitimate to memorize that a certain characteristic function comes from a certain probability law, or vice versa, and to prove the connexion by whichever process is simpler.

Example. If $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$, where it is known that the x's are independent continuous random variables, and that for any $j = 1, 2, \ldots, n$

$$p(x_j) = \frac{1}{\pi(1+(x_j-a)^2)},$$

use the characteristic function to deduce the probability law of $\bar{x}$.

From the evidence of the preceding example we may guess the characteristic function of x_j to be

$$\phi_{x_j}(t) = e^{-|t|+ait}$$

and prove that it is so by direct integration.

$$p(x_j) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \phi_{x_j}(t)e^{-itx_j}\,dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-it(x_j-a)-|t|}\,dt.$$

Dividing the integral into two parts as before and integrating separately we have

$$p(x_j) = \frac{1}{2\pi}\left[\frac{1}{1-i(x_j-a)} + \frac{1}{1+i(x_j-a)}\right] = \frac{1}{\pi(1+(x_j-a)^2)}.$$

Since to every probability law there is a uniquely defined characteristic function, $\phi_{x_j}(t)$ as defined above will be the characteristic function of the variable x_j, following the given probability law $p(x_j)$.

It has been demonstrated earlier that

$$\phi_{\bar{x}}(t) = (\phi_{x_j}(t/n))^n.$$

Hence the characteristic function of $\bar{x}$ will be

$$\phi_{\bar{x}}(t) = \left(\exp\left[-\left|\frac{t}{n}\right| + ai\frac{t}{n}\right]\right)^n = e^{-|t|+ait}.$$

It follows that

$$p(\bar{x}) = \frac{1}{\pi(1+(\bar{x}-a)^2)}.$$

Exercise. Given that x_j is a discontinuous random variable which may take only positive values and that

$$\phi_{x_j}(t) = e^{ita}[pe^{ith}+q]^n,$$

where a and h are constants, find $p(x_j)$ and $p(\bar{x})$, where

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j.$$

Example. x is a discontinuous random variable which may take only zero or positive integer values. If

$$\phi_x(t) = \exp[-a(1-e^{-b(1-e^{it})})],$$

where $a>0$ and $b>0$, calculate the probability law of x.

$$P\{x=k\} = p_x(k) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \exp[-itk - a(1-e^{-b(1-e^{it})})]\,dt.$$

If, as before, we replace the exponential by a series uniformly convergent in t, then

$$p_x(k) = \frac{e^{-a}}{2\pi}\sum_{r=0}^{\infty}\frac{a^r e^{-rb}}{r!}\sum_{l=0}^{\infty}\frac{(rb)^l}{l!}\int_{-\pi}^{+\pi} e^{it(l-k)}\,dt.$$

When $l=k$ the integral is equal to 2π and when $l \neq k$ the integral is zero. Hence

$$p_x(k) = \frac{e^{-a}b^k}{k!}\sum_{r=0}^{\infty}\frac{a^r r^k e^{-rb}}{r!}.$$

This is the probability law of Neyman's contagious distribution discussed earlier.

REFERENCES AND READING

The proof of the converse theorem for the continuous random variable will be found in H. Cramér, *Random Variables and Probability Distributions*, Cambridge Tracts in Mathematics, no. 36.

It is difficult to find an elementary treatment. It is suggested that the reader desiring further problems should take the examples of the previous chapters and solve their converse.

INDEX