STATISTICS

with applications to

HIGHWAY TRAFFIC ANALYSES

BRUCE DOUGLAS GREENSHIELDS, C.E., Ph.D.
Professor of Civil Engineering
The George Washington University

FRANK MARK WEIDA, Ph.D.
Professor of Statistics
The George Washington University

THE ENO FOUNDATION FOR HIGHWAY TRAFFIC CONTROL
SAUGATUCK · 1952 · CONNECTICUT

Eno Foundation Publications
are provided through an endowment by the late William P. Eno

 Published by the Eno Foundation at Saugatuck, Connecticut, October, 1952. Copies of this book are not to be sold.

FOREWORD

Realizing the need for a publication to encourage further scientific approach to the solution of many traffic problems, the Eno Foundation is pleased to present this methodical discussion of some statistical theories and their application in the analysis of traffic data.

The Foundation was fortunate in acquiring the services of Dr. Bruce D. Greenshields, Professor and Executive Officer, Civil Engineering Department, and Dr. Frank M. Weida, Executive Officer, Department of Statistics, The George Washington University, as co-authors. By knowledge and experience they are eminently qualified. They have been guided by a practical insight and have shown an unusual and necessary discernment of the subject.

In some quarters, thinking on traffic as a national problem has reached a degree of desperation. This is due partly to confusion. It is hoped this study will provide some clarification by emphasizing the importance of an analytical basis for initiating logical improvements. Such procedure should tend to create better understanding and much-needed uniform basic methods.

It has been a privilege for the Eno Foundation to provide the preparation and publication of this monograph. Publication has resulted from considerable time and effort by both authors and the Foundation Staff.

THE ENO FOUNDATION

PREFACE

The engineer, and particularly the traffic engineer working in a comparatively new field, faces constantly the need for new, more precise information. To obtain this information, he collects and analyzes data. The theory and procedures to be followed in such analyses have long been known to the statistician, but not always to the engineer.

Mathematics he learns for his engineering is of the classical type—algebra, trigonometry, calculus — in which exact answers are obtained. In statistics no answer is exact for there is always a range of variability within which the true answer lies. Variance, the measure of this variability, may in some cases be so small that the result for practical purposes may be considered exact. But usually it is not. In traffic behavior, a phase of human behavior, it is well to employ the "mathematics of human welfare."

Traffic research carried on at various times over a period of years by one of the writers has served to confirm the fact that traffic behavior tends to follow definite statistical patterns. The difficulty of solving the problems encountered in analyzing the data collected during that research pointed to the need for someone to gather together and explain the statistical methods most pertinent to traffic analyses.

In response to this need, this monograph is written. Desired information, it was felt, could be assembled, developed, and presented most effectively, by a traffic engineer and a statistician working together. The one would know the viewpoint of the engineer and the limitation of his statistical training and vocabulary. The other would provide that knowledge and skill in his own field that can be obtained only after years of work and study.

The authors, despite the work involved, have enjoyed what seemed to them a very worth while undertaking. This monograph is not in any sense the last word on the subject. It is merely an introduction, which they hope will assist the engineer in determining the type and amount of data he needs to obtain sufficiently

accurate answers to his problems and save him time and effort. They trust that if it is a new tool to him it will be to his liking.

In the first four chapters the authors have attempted to explain this mathematical tool, and in the last one they have attempted to show how to use it.

The authors wish to thank the Eno Foundation and staff for its kindly criticism, good counsel, encouragement and sponsorship. They are indebted to Professor Herman Betz of the Department of Mathematics at the University of Missouri for his careful review of the manuscript.

Washington D. C.
June 1, 1952

BRUCE D. GREENSHIELDS
FRANK M. WEIDA

ACKNOWLEDGEMENTS

Professor RONALD A. FISHER, Cambridge, Dr. FRANK YATES, Rothamstead, and Messrs. OLIVER AND BOYD LTD., Edinburgh, for permission to reprint Appendix Tables II and IV from their book, "*Statistical Tables for Biological, Agricultural, and Medical Research.*"

GEORGE W. SNEDECOR and the IOWA STATE COLLEGE PRESS, Ames, Iowa for permission to reprint Appendix Table V from their book "*Statistical Methods,*" 4th edition.

BUREAU OF PUBLIC ROADS, Washington, D.C. for charts used from "*Highway Capacity Manual.*"

TABLE OF CONTENTS

Page

APPENDIX

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

THE NATURE AND UTILITY OF STATISTICS

I. 1. *General Remarks.* The rapid movement of traffic on our streets and highways in ever changing patterns is one of the most familiar and beneficial phenomena of our daily lives and at the same time one of the most confusing and vexing. The annoyances and even danger experienced in driving over congested streets and highways, the lack of places to park and, in general, the inadequacies of our highway system are widely recognized. There is clearly a need for increased knowledge of traffic behavior in order that traffic regulation and planning may be made more scientific. The method by which scientific knowledge is increased is to observe what happens and then by inductive reasoning to establish general laws pertaining to these happenings. It is the purpose of this book to develop a scientific system known as *Statistical Methods* and show how to use these methods for analyzing and solving traffic problems.

Mathematical probability, which is the basis of all statistical theory, had its beginning in ancient times. Certain mathematical patterns developed as pastimes by the Greeks and others were first found to coincide with chance happenings such as occur in card games and later found to coincide with actual happenings. It was not until the Seventeenth Century that one of the first practical uses was made of probability, when life expectancy tables were published for use in computing life insurance premiums and benefits. Among the early important contributors to the theory of probability we find the names of DeMoivre, La Place, Gauss, Pascal, Fermat and Bernoulli.

The methods of statistics have long been employed by the chemist, the sociologist, the physicist, the biologist, the bacteriologist, the physiologist, the economist, the meteoroligist, the business man, the psychologist, and many others. In the *biological sciences*, the whole theory of evolution and heredity rests in reality on a statistical basis. Likewise, the behavior of the body mechanism itself lends itself to statistical analysis. Statistical theory is the

basis of various aspects of *theoretical physics* and *chemistry* as demonstrated by Gibbs, Bohr, Einstein, Fermi, Dirac and others. In the *social sciences*, statistics is used in the measurement of the sizes of the population, the birth, marriage, mortality and morbidity rates, and in determining the distribution of the population by trade or income, wages, prices, production, foreign trade, and transportation. In *manufacturing*, statistics facilitates efficient management, economic control of the quality of manufactured products, and the evaluation of laws of behavior to determine control or lack of control. Statistics is the basis of corrective legislation. But in spite of this wide-spread use, it is only within the last few years that the traffic engineer has come to realize that statistics is his most useful tool[1]. The traffic engineer should fully realize the importance of the statistical approach to the solution of his problems. If there has been some failure on his part to do so, it no doubt is due to its omission from his engineering training in which he has been taught to assume that the values with which he is dealing are exact and always the same. Each individual piece of material of a given kind and size is assumed to behave the same as any other piece of the same kind and size. Statistics deals with measurements which at best are approximate values which are usually not the same when repeated. In traffic engineering, the individuals are human and it can not be assumed that they will always behave in precisely the same manner.

The automobile does not become a complete mechanism until the driver is behind the wheel. It is the driver who sees the curve ahead and turns the steering wheel accordingly, who sees the obstruction and applies the brakes. It is the emotional and physical characteristics of the driver that must be measured and evaluated. To this end, the traffic engineer must use the special type of mathematics that applies to the problem he is considering.

In this attempt to make statistics more readily available to the traffic engineer and others, an effort will be made not only to explain statistical methods, but to show by example how they may be used in the solution of traffic problems. An understanding of the calculus is desirable but not essential for use of the methods involved. In using statistics it must be kept in mind that it is the

handmaiden of reality and not reality itself. In all cases it must be demonstrated that the statistical law of behavior to be used agrees with actual behavior.

As the statistical methods are developed, it will be found that they constitute a unified structure. This will become apparent as the development is followed step by step. The first step will be to explain statistical terms through the derivation and explanation of the mathematical and *statistical* probability formulae which form the basis of statistics. The use of these formulas will become clear through their application to the solution of typical problems.

I. 2. *Definition and Nature of Statistics. Statistics is the fundamental and most important part of inductive logic.* It is both an art and a science, and it deals with the collection, the tabulation, the analysis and interpretation of quantitative and qualitative measurements. It is concerned with the classifying and determining of actual attributes as well as the making of estimates and the testing of various hypotheses by which probable, or expected, values are obtained. It is one of the means of carrying on scientific research in order to ascertain the *laws of behavior* of things – be they animate or inanimate. Statistics is the technique of the *Scientific Method.*

I. 3. *Statistics and Mathematics.* Statistics is a branch of applied mathematics. It differs from so-called pure mathematics in that the values in statistics are approximations or estimates, but not mere guesses. The rules and methods of operation are those of pure mathematics for it is the tool of statistical analysis.

An "exact" value in pure mathematics may be thought of as one of the possible values a variable may assume. There are but two possibilities in pure mathematics, namely: the variable has a certain value or it does not have that value. In the first case, the probability is 1, meaning that it is certain that the variable has that value, while in the second case the probability is zero, meaning that it is certain that the variable does not have that value.

The variable in statistics, called *stochastic variable* or *variate*, is much more general than the variable in pure mathematics. The stochastic variable is one, to each of the many possible values of

which, there is attached a probability, p, that it attains said value. As will be shown in Chapter III, this probability may have any value between zero and one. This fact is expressed mathematically as $0 \leq p \leq 1$.

The stochastic or random variable may be *discrete* or *continuous*. It is called discrete if it can take on only certain isolated values in an interval and it is called continuous if it can take on *any* value in an interval. It is to be noted that the probability that a continuous stochastic variable has a specific value is always zero.

I. 4. *Two General Types of Problems.* Statistics deals with problems that fall into two general categories.

1. The first of these categories of problems has to do with characterizing a given set of numerical measurements or estimates of some attribute or set of attributes applying to an individual or a given group of individuals. This entails the finding of a mathematical model that fits the pattern of the variation in measurements or the variation in the things being measured. The engineer is familiar with the fact that a distance may be measured several times with a different result each time, and he knows that the mathematical pattern called "*The Principle of Least Squares*" is used in characterizing such measurements. In studying some attribute such as the ability of students, it is found that there are just as many brighter than "average" as there are less bright and this pattern is called "*normal*" and there is a mathematical equation for such a *normal curve*. Other laws of behavior (distributions) are found to follow other mathematical patterns, such as Poisson's "*random*" curves (distributions), the Pearson system of distribution and others.

Fortunately, these mathematical patterns are all of the same basic nature. It will be one of our tasks to describe and explain this phase of statistical mathematics.

2. The second category of problems has to do with characterizing an attribute or attributes belonging to all individuals of the group one is investigating, such as all white pine lumber or all the people living in Ponca City, all people with red hair, or all aluminum alloys of a given specification. These well defined classes of items

are called *populations* or "*universes*". This second class of problems involves the selection of *random* samples from the population, the statistical study of these samples, and the drawing of inferences from them.

The problems just mentioned indicate that (1) the data must be summarized as will be discussed in Chapter II; (2) they must be thoroughly analyzed by obtaining mathematical patterns of the laws of their behavior as will be discussed in Chapter III; and (3) it must be possible to draw inferences from the samples in regard to the reliability and significance of pertinent summary values obtained from the samples for the purpose of characterizing the "universe" as will be discussed in Chapter IV.

I. 5. *Types of Sampling.* One may classify random sampling in two ways: (1) Sampling by attributes; and (2) Sampling by variables, either discrete or continuous. In sampling by attributes, one determines the number of times (the frequency) the event happened as specified and the number of times the event did not happen as specified. In sampling by variables, we measure such things as the weight or length of an object, the duration of an event or the intensity of a force. We may also measure a group of individuals in order to characterize them in regard to multiple categories such as weights, heights, temperatures, etc., to be considered jointly. The basis of all such characterizations is counting. Hence we must determine the frequency of the occurrence of a characteristic or event among n possible occurrences or non-occurrences or among n trials.

I. 6. *The Variables to be Measured and Interpreted.* The statistical or scientific method applies not only to the analysis and interpretation of data but to the whole procedure of first recognizing the need for increased knowledge about a particular problem; second, the gathering of data about the problem; third, studying the significance of the data; and finally, presenting the results of the investigation in a report. In carrying out this statistical procedure there are certain precautions that must be observed.

The recognition of the need for more information about a particular problem usually comes from those who have to deal with it.

A research project conducted in Ohio in 1939[4] will serve to illustrate the steps in conducting an investigation to obtain certain specific information. This study had to do with center-line markings of roadways. The fact that different states had, and still have, different systems of markings, causing confusion to motorists, pointed to the obvious need of determining the best type.

The first question to be answered was: Is the problem solvable by statistical methods? If so, what method or methods are applicable, what variables need to be measured, how much data are needed, and how best to obtain the needed data?

In the problem of center-line marking, one is interested in the qualities that make a good center-line marking. Some such qualities are visibility, interpretability and durability. But what about other things? Is a broken line just as satisfactory for a center-line as a solid line? The broken line is cheaper because it requires less paint. What kind of a line or lines should be used to mark a "no-passing" zone? Such questions, of course, can only be answered after the study is made. Hence it was necessary to make a *provisional conjecture* as to what types of center-line marking should be tested.

I. 7. *Means of Measuring the Variable, and Precautions to be taken.* Having decided provisionally on what types of center-lines to test, the next step was to devise a means of measurement. Should it be done by noting the behavior response pattern of drivers to different types of markings? Should a speed check be made? Should drivers be questioned? Should some other methods be used? What is the probable cost and efficiency of the different possible methods? What type of equipment is necessary to make the recordings?

It has been found by experience that it is sometimes necessary to design and construct special equipment or apparatus to record field data. It is recalled that in 1932[2] it was only after considerable thought that the rather simple expedient of time-motion pictures was used to record the speed and spacing of vehicles. A mechanical device, provided it is first checked for mechanical defects, is always more reliable than human judgment. The picture method possessed one other feature that is not often attained. It

gave complete information on all that happened within the field of view. The pertinent information could then be selected at leisure and if a wrong conjecture was made, other information already in hand could be studied.

It was decided in the 1939 project to take speed recordings with the Eno-scope, a device using mirrors so arranged that the time at which a vehicle passes two successive positions on the roadway can be recorded by means of a stop watch. These positions must be a considerable distance apart, usually 88 or 176 feet, so that the human variation in snapping the watch will not cause an appreciable error. Another source of error that is not so readily apparent is the inability of the observer to take a random sample without taking the proper precautions to obtain one. It would seem that if the observer simply recorded the speed of as many vehicles as possible it would result in an unbiased sample, but such is not the case. Vehicles tend to bunch into queues behind the slower drivers. Depending upon the alertness of the observer, he may be unconsciously selecting slow or fast vehicles. He must arbitrarily select some convenient numbered vehicle such as every third one.

This device is not infallible. Suppose, for instance, that an origin-destination survey is being conducted to determine the travel routes of people living in different sections of a city, and that it has been decided to interview every tenth house starting from an arbitrary point. But would we be correct in assuming that every tenth house constitutes a *good* random sample? It could be that every tenth house is a corner house and hence may be a shop of some kind. In this case, some special procedure must be used, such as writing the numbers on cards and after shuffling, picking every tenth card.

I. 8. *The Size of the Sample.* The size of the sample is the quantity of data needed to meet certain considerations. One of the considerations is cost, another is time. These depend upon the decision as to (1) the maximum error that will be tolerated and (2) the degree of certainty demanded that this allowable or maximum error will not be exceeded. This definitely determines the size of the sample or the amount of data to be collected. The method of gathering the data

is largely dependent upon the structure and character of the "universe" from which the sample is taken.

In the Ohio study of 1939, it was desired among other things to get the opinions of drivers about center-lines. Did they prefer a yellow line, a white one, a broken line, or a solid line? The obvious procedure was, of course, to stop each motorist and ask his opinion. But how many? Would the majority of 30 or 40 people agreeing on one combination as being the best be sufficient? At first one might possibly say yes, but on second thought he would realize that all opinions might not be unbiased. Perhaps the drivers from Pennsylvania had grown accustomed to a certain combination and would prefer that, or the drivers from Ohio might prefer a different system. This possible tendency to biased opinions meant that a larger sample should be taken and also that along with the opinions, the residence of the driver should be ascertained.

Sometimes opinions are unconsciously biased. This fact also was brought out in the Ohio study. It was decided to try road signs worded to warn drivers that they were entering a "no-passing" zone. It was doubted that a large percentage of the motorists would see the signs, but surprisingly enough, over 98 percent of them stated they had seen the signs. This was so unexpected that it was questionable, and a way of checking these answers was sought.

The means of checking was revealed through consideration of the purpose of the sign. Signs aside from those whose shape conveys a message, must be read. A sign much larger than the "no-passing" sign was prominently displayed to warn the drivers that they were entering a "test-zone". This might have been guessed from the fact that they had seen 3 or 4 different types of marking within a mile or so, but, over one-third when questioned said they did not know they were in a "test-zone". The conclusion reached was that at least one-third and probably more did not see the "no-passing" signs in spite of the fact that 98 percent said they had.

I. 9. *The Validity and Reliability of Measurement.* It is not only opinion measurements that must be checked for validity. In a study of brake-reaction-time made in Ohio in 1934[3], it was decided to determine whether the facts warranted the assumption that those

with quick reaction-time were safer drivers. It was perhaps perfectly logical to assume that a quick reaction will enable a driver to avoid accidents, but the study showed no relationship of accidents to brake-reaction-time. If this were true, and other investigations have shown that it is, then we deduce that an individual with a slow reaction-time employs a larger margin of safety and so compensates for his shortcoming. In other words, brake-reaction-time is not a valid measurement to determine whether a driver is a safe driver or not since it does not in fact measure what it was supposed to measure.

A measurement is reliable if there is consistency in obtaining it. In other words, consistency in measurements increases our confidence in the reliability of the conclusion we wish to draw from the set of measurements.

I. 10. *Cost of the Project.* After the amount of data needed to obtain results accurate to the degree desired has been estimated, the apparatus needed has been decided and the procedure outlined, it is possible to estimate the *minimum* cost. This cost will depend to a large extent on the amount of personnel needed and the time required to complete the study. The cost of development research is easier to estimate than that of basic or fundamental research. In the former we know much more about the expected results. Development research follows the fundamental. It is often used to verify results that have been suggested by more basic studies. In any case, however, it is necessary to estimate the cost. The skill of the researcher is rightly or wrongly measured by his ability to estimate correctly this cost and effort required to carry on an investigation to the point where definite results, whether positive or negative, are obtained and reported.

I. 11. *The Report.* A preconceived idea or system of thinking must not be allowed to influence the reporting of results. A negative result is just as important as a positive one. Too often an investigation is conducted to prove a point and this attempt to adhere to an established opinion may have undue influence in selecting the attribute to measure.

The results of a scientific investigation should be presented with the same care that was used in conducting the survey. All too often, information is brought to light only to lose its value through poor presentation. Knowledge is useful only as it becomes known. Fortunately there has been developed a recognized style of engineering reports and several good books on the subject are available.[5] It should be emphasized that the writing of the report should be considered a part of any scientific investigation, and a most important part.

I.12. *Purpose of the Book.* Having indicated the general procedure, and noted some of the precautions that need to be taken, we shall now attempt to discuss the necessary theory and outline the techniques for the solution of traffic problems. Finally we shall attempt the solution or partial solution of some of the more typical problems.

Chapter II presents the method of summarizing data and obtaining summary numbers that are useful for the analysis, characterization and interpretation of one or more sets of measurements.

Chapter III presents the theory and basis of the various mathematical patterns (laws of behavior) that are the underlying principles upon which the analysis and interpretation of the results depend.

Chapter IV shows the use of summary methods of Chapter II and the basic theory of Chapter III to solve problems by statistical methods and to ascertain the reliability, validity, significance, and meaning of the solution.

Chapter V outlines the solution or partial solution of some typical as well as some of the more unusual traffic problems.

REFERENCES, CHAPTER I

[1] Kinzer, John P. "*Application of the Theory of Probability to Problems of Highway Traffic*," Proceedings, Institute of Traffic Engineers, 1934, pages 118–123.

Adams, W. F., "*Road Traffic Considered as a Random Series*," Institution of Civil Engineers Journal, November 1936, pages 121–130.

Greenshields, Bruce D., "*Initial Traffic Interference,*" Presented for discussion at the 16th Annual Meeting of the Highway Research Board, November 19, 1936, Washington, D. C., 9 pages mimeo and the comments by W. F. Adams

[2] Greenshields, Bruce D., "*The Photographic Method of Studying Traffic Behavior,*" Proceedings, Highway Research Board, Washington, D.C., 1933 pages 384–399.

Ibid., Schapiro, Donald; and Ericksen, Elroy L.; "*Traffic Performance at Urban Street Intersections,*" Yale Bureau of Highway Traffic, New Haven, Connecticut, 1947, pages 73–118.

[3] Ibid., "*Reaction Time in Automobile Driving,*" Journal of Applied Psychology, Vol. XIX, No. 3, June 1936, pages 353–358.

[4] Report of Highway Research Board Project Committee on "*Markings for No-Passing Zones,*" November 1939.

[5] Nelson, J. Raleigh, "*Writing The Technical Report,*" McGraw-Hill Book Co., 1947.

CHAPTER II

SUMMARIZING OF DATA

II. 1. *Objective.* After the data have been collected, it is not only convenient but necessary that they be condensed in order to be analyzed and interpreted by means of *summary numbers* which serve to characterize the data. Some summary numbers are averages and included among them are the mean, the median, the mode, and the standard deviation.

This chapter shows how to summarize data both analytically and graphically. The procedures will be made clear by examples.

II. 2. *Frequency Distribution.* A frequency distribution constitutes the first step in classifying and condensing data. It is an arrangement in which the data consisting of separate values or measurements of a variable are combined into groups called *classes* covering a limited range of values, such as 1 to 5 miles, 5 to 10 miles, etc. The number of values in each class is called the *class frequency.* Once the observations have been combined into groups, the individual items lose their identity and the midpoint of the class group becomes a unit quantity with a broader meaning. This requires that the grouping be done in such a way that it will accurately represent the items from which it is computed. The methods to be followed will become clear with an examination of the construction of a frequency table.

II. 3. *Class Interval and Class Mark.* A *class interval* sets *boundaries* or *limits* to a class of a frequency distribution. In Table II. 1., the *lower* bounds of the classes are 15, 20, ...; the *upper* bounds are 19, 24, 29, ...; the *lower boundaries* or limits are 14.5, 19.5 ...; the upper *limits* or *boundaries* are 19.5, 24.5, ... The class interval is 5. By the laws of approximate numbers, the data have been rounded off to the nearest whole number so that the speeds are correct to the nearest mile per hour.

Table II. 1

SPEED IN MILES PER HOUR OF FREE MOVING VEHICLES ON SEPTEMBER 16, 1939, IN OAKLAWN, ILLINOIS ON U.S.H. 12 and 20 AT A POINT ONE MILE EAST OF HARLEM AVENUE

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Speed in m.p.h.	*Number of Vehicles*	*Smoothed Frequency*	*PerCent of Vehicles*	*Relative Frequency*	*Cumulative Frequency*	*Cumulative Per Cent Frequency*
s	f	f_c	100 f/n	f/n	f_c	100 f_c/n
70–74	0	0	0	0		
65–69	0	0.7	0	0		
60–64	2	5.7	0.67	0.0067	300	100.00
55–59	15	10.3	5.00	0.0500	298	99.33
50–54	14	19.3	4.67	0.0467	283	94.33
45–49	29	39.0	9.67	0.0967	269	89.67
40–44	74	54.3	24.67	0.2467	240	80.00
35–39	60	65.7	20.00	0.2000	166	55.33
30–34	63	50.7	21.00	0.2100	106	35.33
25–29	29	32.7	9.67	0.0967	43	14.33
20–24	6	14.3	2.00	0.0200	14	4.67
15–19	8	4.7	2.67	0.0267	8	2.67
10–14	0	2.7	0	0	0	.00
	300 = n	300.1 = n	100.02	1.0002		

Data furnished by Public Roads Administration, Washington, D. C.

Note: This illustration is of a continuous stochastic variable which may take any value. An illustration of a discontinuous variable is the numbers of vehicles that pass over a highway in any time interval. There is no such thing as a part of a vehicle. An illustration of a discontinuous stochastic variable where only even integers are possible is the distribution of rows of kernels on ears of corn.

A *class mark* is the mid-value of the class interval. In Table II. 1., column (1), the class marks are 17, 22, 27, ...

The exact values of a discontinuous variable are usually taken equal to the class marks. For many purposes, all the values of a continuous variable that fall within a given class interval are grouped at the class mark as a convenient approximation.

The number of values that the variable has within a certain class interval is called a *class frequency*. In Table II. 1. the frequency 63 in column (2) corresponds to the class 30–34 in column (1).

Two conditions which serve as a guide in the choice of the size of a class interval are: (a) the desire to be able to treat all the values assigned to any one class, without appreciable error, as if they were equal to the mid-value or class mark of the class interval: (b) for convenience and brevity, it is desirable to make the class interval as large as possible, but always subject to the first condition. These two conditions will in general be fulfilled if the interval is so chosen that the number of classes lies between ten and thirty. This does not mean, however, that the minimum may not be less than ten classes nor the maximum more than thirty classes;

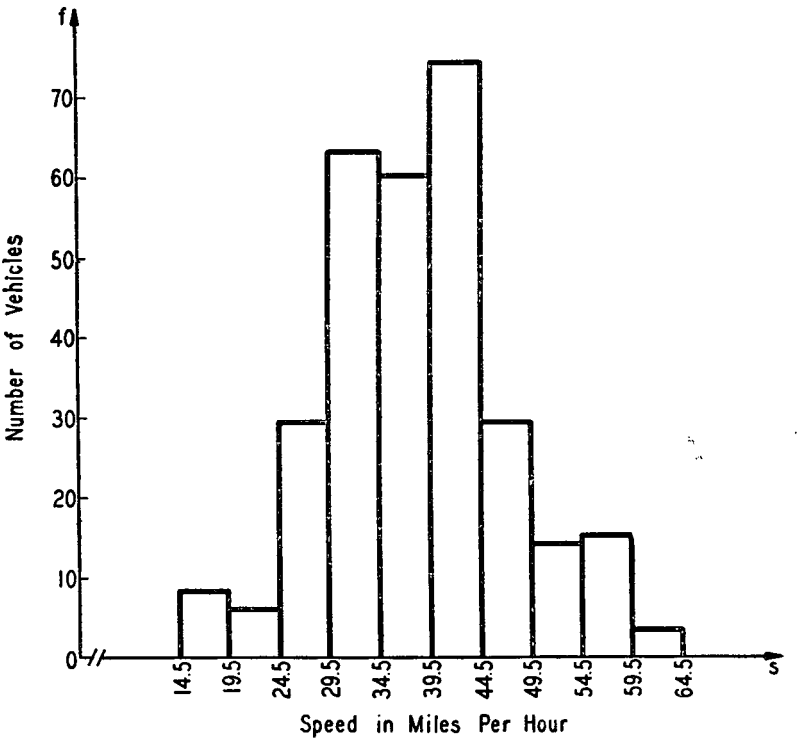


FIGURE II. 1

FREQUENCY RECTANGLES OF OBSERVED VEHICLE SPEEDS

it merely means that in most cases it is possible to form the classification with the number of intervals lying between ten and thirty.

Another convenient means of classification is the *graphical summary method.* There are five types of graphs that have been found useful: namely, the *Frequency Rectangles*, the *Histogram*, the *Frequency Polygon*, the *Smoothed Frequency Polygon*, and the *Frequency Curve.* We shall now discuss these in the order named.

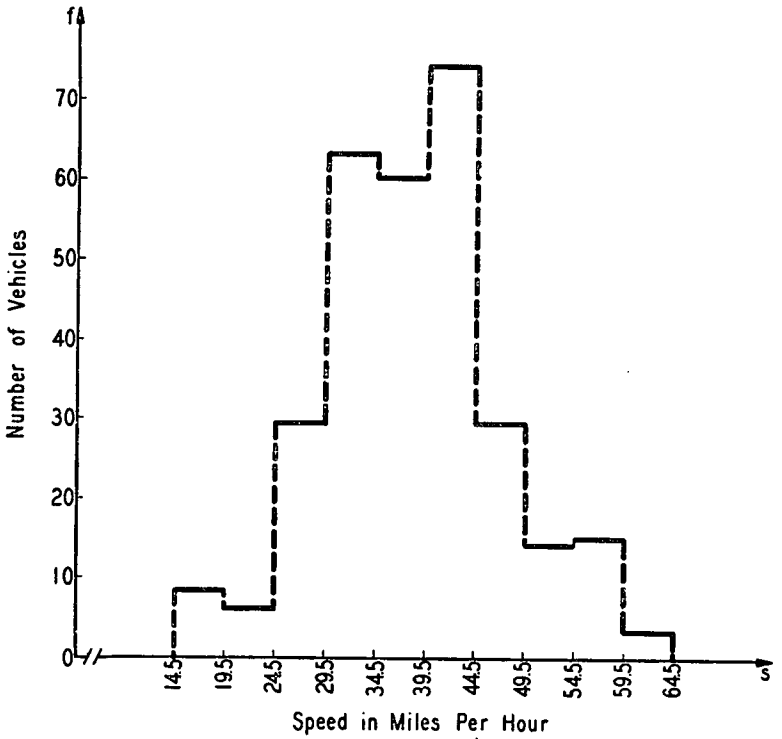


FIGURE II. 2
HISTOGRAM OF OBSERVED VEHICLE SPEEDS

II. 4. *Frequency Rectangles.* Using the frequency distribution as given by columns (1) and (2) in Table II. 1., the rectangles, shown in

Figure II. 1 may be drawn. The class intervals are the bases and the altitudes (ordinates) are equal to the frequencies of the classes.

Unit area is defined as that of a rectangle whose base is a class interval and whose altitude is a unit of frequency. This gives a one to one correspondence between area and frequency. In other

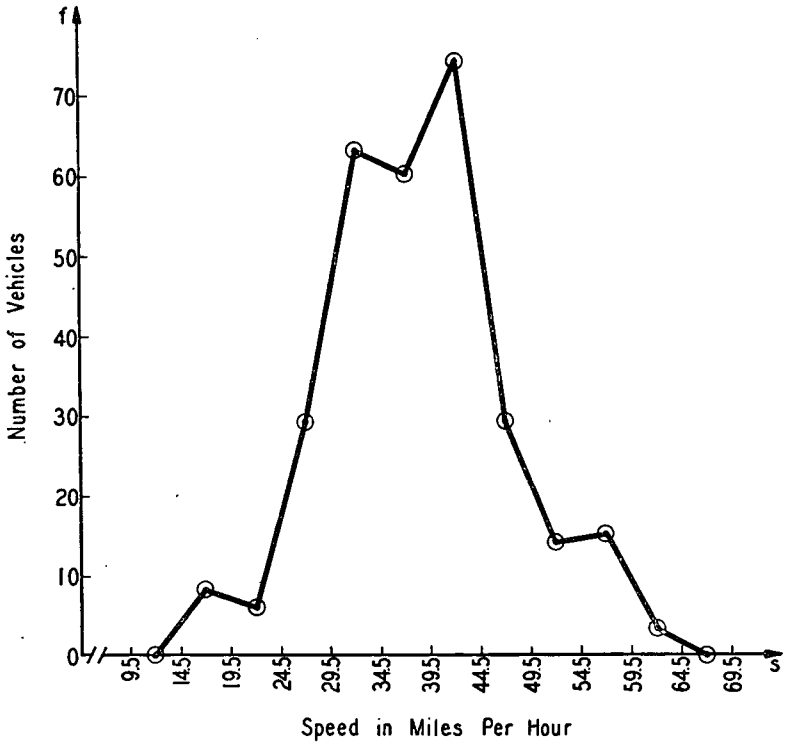


FIGURE II. 3

FREQUENCY POLYGON OF OBSERVED VEHICLE SPEEDS

words, since the base is equal to one (class interval), the height is the frequency.

II. 5. *Histogram.* A histogram is the system of upper bases of the frequency rectangles. It is illustrated in Figure II. 2. for the frequency distribution given by columns (1) and (2) of Table II. 1.

II. 6. *Frequency Polygon.* A frequency polygon is formed by selecting a convenient horizontal scale for the variable being measured and a vertical scale for the class frequency and then plotting the points so that the class marks are the abscissas and the class frequencies are the ordinates. This method is shown in Figure II. 3. for the distribution given in Table II. 1.

II. 7. *Smoothed Frequency Polygon.* The smoothed frequency polygon is a means of *graduation* sometimes called a method of *moving averages.* It is useful in obtaining an approximation to the probable frequency curve or theoretical law of behavior of the attribute that is being measured.

One method of obtaining moving averages is illustrated in Columns (1), (2), (3), in Table II. 1., in which the smoothed value for an interval is obtained by summing the frequencies in that interval and the two adjacent intervals and dividing by three. Hence, the smoothed value for the interval 15–19 is equal to the sum of the frequencies 0, 8, and 6, divided by 3. For the interval 20–24, we add the frequencies 8, 6, and 29, and divide the sum by 3. We proceed likewise for the remaining intervals. The smoothed frequency polygon for the distribution given in columns (1) and (3) of Table II. 1. is shown in Figure II. 4. By comparing Figure II. 4 with Figure II. 3., it is seen that the smoothed frequency polygon has removed the irregularities found in Figure II. 3. and is closer, in appearance, to a frequency curve. See definition of frequency curve, Article II. 8.

The number of classes over which an average is taken does not need to be three. The decision as to the number of classes that should be taken depends upon the total frequency, the total number of classes in the distribution, the size of the class interval, the equality or inequality of the classes, and the experimental error, the discussion of which is beyond the scope of this book. The process of smoothing tends to correct for *sampling errors*, *grouping errors*, and *experimental errors.*

An important point to note is that the total area within the rectangles, the histogram, the frequency polygon, the smoothed frequency polygon and within the frequency curve is equal to the

total frequency n. This total frequency in terms of probability is thought of as *one* and in terms of per cent as 100 per cent. The height of the frequency rectangles is then expressed as a fraction or a per cent.

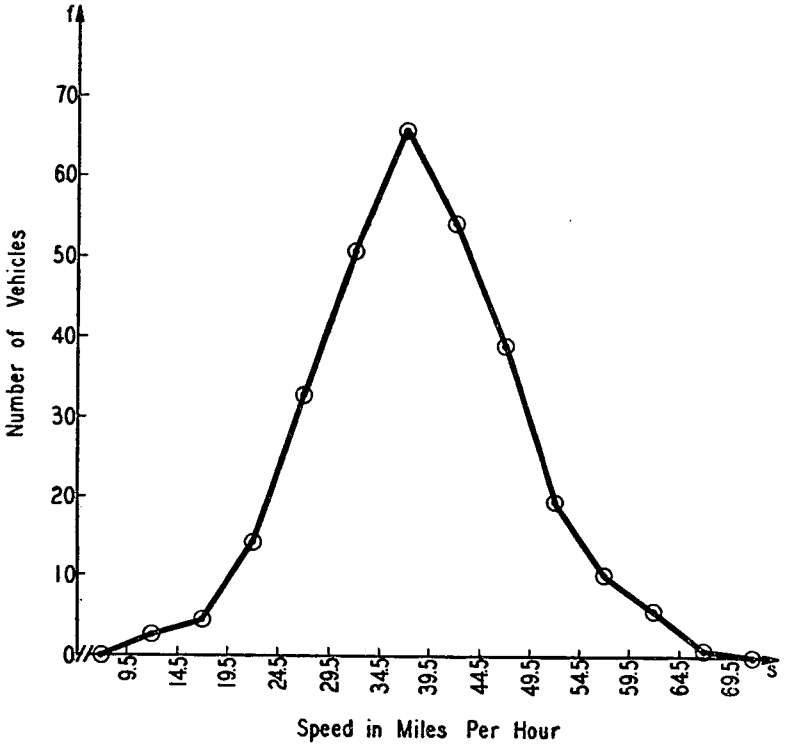


FIGURE II. 4

SMOOTHED FREQUENCY POLYGON OF OBSERVED VEHICLE SPEEDS

II. 8. *Frequency Curve.* A smooth curve superimposed upon the frequency polygon or smoothed frequency polygon so that the area under it is equal to the total frequency is known as a *frequency curve.* The *frequency curve* is an estimate of the limit that would be approached by a frequency polygon or a smoothed frequency polygon if we indefinitely decreased the size of the class intervals

and at the same time indefinitely increased the frequency n. An illustration of a frequency curve for the distribution given in Table II. 1. is given in Figure II. 5. where the points of the smoothed frequency polygon have been used.

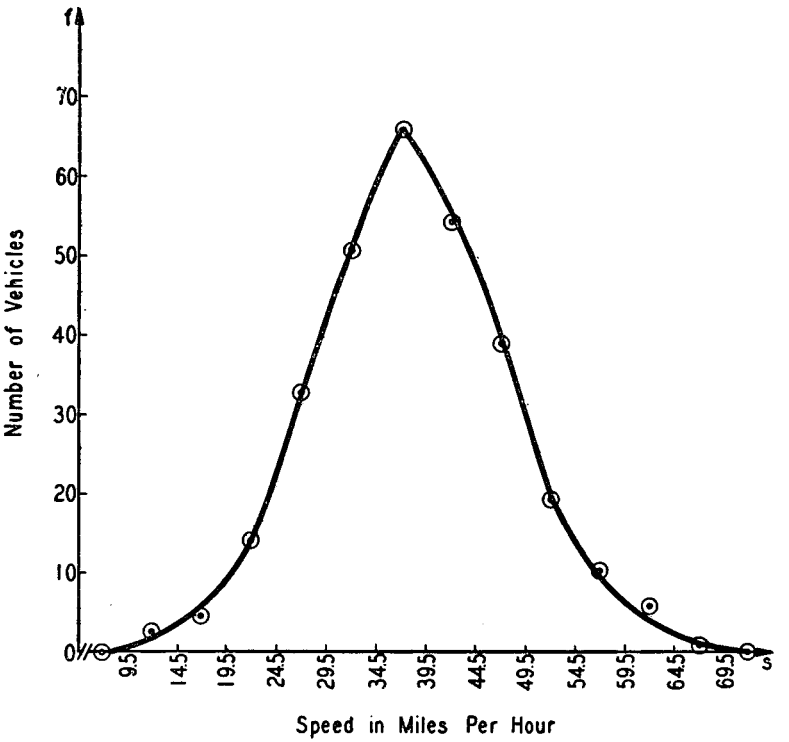


FIGURE II. 5

FREQUENCY CURVE OF OBSERVED VEHICLE SPEEDS

II. 9. *Cumulative Frequencies.* Another type of distribution can be secured by the use of cumulative frequencies. These values are shown in column (6), Table II. 1., and are obtained by successive adding of the frequencies, beginning with the lowest interval. To illustrate: starting with 8, add 6 to 8 and get 14; then 29 + 14 which equals 43, and so on until 298 plus 2 equals 300 for the last cumulative frequency which, of course, is the total number of cases.

The cumulative frequency distribution in the example given shows how many vehicles had a speed below (or above) a given speed. From columns (1) and (6) in Table II. 1., we find that 8 vehicles had a speed less than 19.5 miles per hour, 14 had a speed less than 24.5 miles per hour; 43 had a speed less than 29.5 miles per hour and so on. In some cases the cumulative frequencies expressed as per cents of the total frequencies are more meaningful. These per cents are given in column (7), Table II. 1. According to column (7), 2.67 per cent of the vehicles have a speed less than 19.5 miles per hour, 4.67 per cent of the vehicles have a speed less than 24.5 miles per hour and so on.

To obtain the graph of the cumulative frequencies or the cumulative per cent frequencies, the points are plotted with cumulative values as ordinates and the *upper limits* of the corresponding classes as abscissas.

The points then are connected with straight line segments (polygon) or with a smooth curve. In either case the resulting graph is called an *ogive*. The curve may be interpreted as portraying a law of growth. If the cumulation is in the opposite direction, we would obtain a law of negative growth. In the case given, 2 vehicles (0.67 per cent) have a speed greater than 59.5 miles per hour; 17 vehicles (5.67 per cent) have a speed greater than 54.5 miles per hour and so on. The ogive for both the absolute and percentage scale is shown in Figure II. 6.

The class frequencies may also be expressed as per cents or relative frequencies. These values are shown in columns (4) and (5) of Table II. 1. In the former case, the total area has been made 100 units of area and in the latter case the total area has been made *the unit* of area.

If $Y = f(X)$ is the equation of the frequency curve, then

$$\int_{X_1}^{X_2} Y dX$$

is the number of observations having a value between X_1 and X_2.

If A is the lower limit of possible values of the variable and B is the upper limit, then the total area N, namely, the total frequency is

$$\int_A^B YdX = N.$$

In terms of relative frequency or statistical probability, we have

$$\int_A^B YdX = 1$$

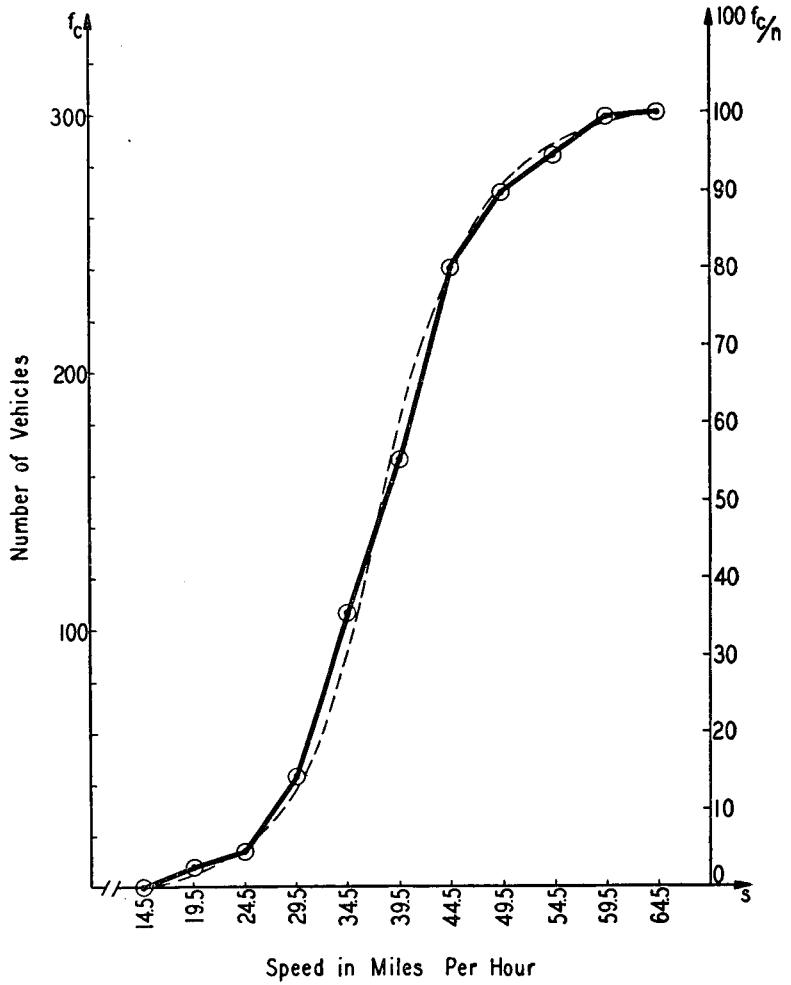


FIGURE II. 6

CUMULATIVE FREQUENCY CURVE OF OBSERVED VEHICLE SPEEDS

where the whole area under the frequency curve is taken as the unit of area.

In the latter case, Y is called the *probability density* and YdX is called the *probability element.*

For the cumulative frequency distribution, in the theoretical case in terms of probability, the expression

$$F(X) = \int_A^X YdX$$

is known as the Distribution Function of Probability where $F(A) = 0$ and $F(B) = 1$ and $A \leq X \leq B$.

Frequency distributions are characterized by summary numbers which often are those functions of the measurements known as *averages.* These averages show the location of central tendencies (if any) and serve as bases for evaluating differences between values (dispersion) as well as skewness and flatness of the distribution. They are also instrumental in isolating extreme or unusual values.

II. 10. *Average. An average is a function of the entire group of values such that if all the values were equal to one another it would equal each one of the group of equal values.*

In general, the values or measurements are unequal, some being larger and some being smaller than the average.

Of the many averages, those which are of most use and interest to the statistician are first, the *common averages* including the *arithmetic mean,* the *median,* the *mode,* the *geometric mean,* and the *harmonic mean*; and second, the *averages of differences* including the *mean (average) deviation,* the *centra harmonic mean,* the *standard deviation,* and the *moments.*

II. 11. *Arithmetic Mean. Graphically,* the arithmetic mean is the abscissa of the centroid of the total area under the frequency curve or frequency polygon.

It is the point at which if the whole area is considered to be concentrated, the first moment of the total area will equal the sum of the first moments of the components of area into which the total area is divided.

From Figure II. 7., if $f_1, f_2, \ldots f_k$ are component areas and if $X_1, X_2, \ldots X_k$ are their corresponding distances from the Y-axis and

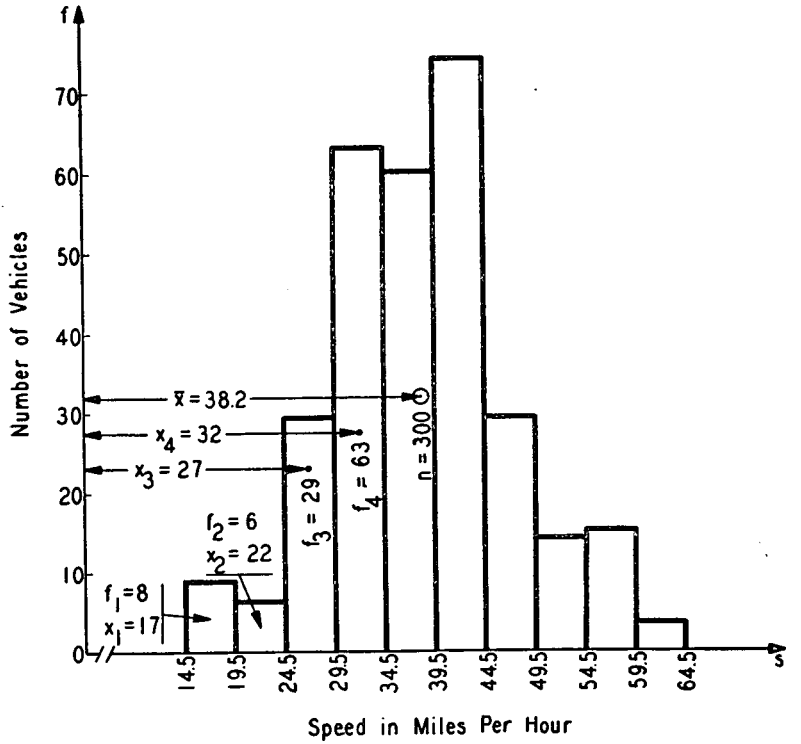


FIGURE II. 7

ARITHMETIC MEAN OF OBSERVED VEHICLE SPEEDS

if $n = f_1 + f_2 + \ldots + f_k$, is the total area and $\overline{X}$ is its distance from the Y-axis, then

$$n\,\overline{X} = f_1 X_1 + f_2 X_2 + \ldots + f_k X_k$$

whence

$$\overline{X} = \frac{f_1 X_1 + f_2 X_2 + \ldots + f_k X_k}{n} = \frac{\sum_1^k f_i X_i.}{n} \quad \text{II. 11. 1.}$$

Algebraically: The arithmetic mean is the sum of all the values of the variable divided by the number of values. If $\overline{X}$ is the arithmetic mean and $X_1, X_2, \ldots, X_n$ represent the values of the variable X, then

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_1^n X_1}{n}. \qquad \text{II. 11. 2.}$$

To illustrate: Let the values of the variable X be 10, 13, 17, and 18. The arithmetic mean of these values is

$$\overline{X} = \frac{X_1 + X_2 + X_3 + X_4}{4} = \frac{\sum_1^4 X_1}{4} = \frac{10 + 13 + 17 + 18}{4} = 14.5$$

When certain values of the variable occur more than once, the same notation may be used, namely:

$$\overline{X} = \frac{X_1 + X_1 + X_1 + X_2 + X_2 + X_3 + \ldots + X_k}{n} \qquad \text{II. 11. 3.}$$

But another symbolic representation is more convenient. Let f_1 be the frequency or number of times the variable X has the value X_1. The sum of the values X_1 is $f_1 X_1$. Let n be the sum of the f_1 where, say, there are k different values of X_1 and hence of the f_1. This symbolic representation gives

$$\overline{X} = \frac{\sum_1^k f_1 X_1}{\sum_1^k f_1} = \frac{\sum_1^k f_1 X_1}{n} \qquad \text{II. 11. 4.}$$

If in II. 11. 4., each $f_1 = 1$ and $k = n$, the expression for $\overline{X}$ is the same as that given in II. 11. 2.

If the class intervals are unequal in size, the computational process may be simplified by making a *simple translation*. Let

$$x'_1 = X_1 - X_0 \qquad \text{II. 11. 5.}$$

where X_0 may be any convenient value whatsoever. In practice it is best to use for X_0 the midpoint of the middle class if there are an odd number of classes, if there are an even number of classes, use

the midpoint of a class as near the middle of the distribution as possible.

Substituting the value of X_1 as given in II. 11. 5. in equation II. 11. 4., we have

$$\bar{X} = \frac{\sum_1^k f_1 X_1}{n} = \frac{\sum_1^k f_1 (x'_1 + X_0)}{n} = \frac{\sum_1^k f_1 x'_1}{n} + \frac{X_0 \sum_1^k f_1}{n}$$

Since $\sum_1^k f_1 = n$ and $\sum_1^k f_1 / n = 1$;

$$\bar{X} = X_0 + \frac{\sum_1^k f_1 x'_1}{n} \qquad \text{II. 11. 6.}$$

In the special case when all class intervals are *equal*, we may use the linear transformation (*translation and change of unit*)

$$x_1 = \frac{X_1 - X_0}{c} \qquad \text{II. 11. 7.}$$

where c is the size of the class interval.

Using the value of X_1 from II. 11. 7. in II. 11. 2.,

$$\bar{X} = \frac{\sum_1^k f_1 (cx_1 + X_0)}{n}$$

$$= \frac{X_0 \sum_1^k f_1}{n} + \frac{c \sum_1^k f_1 x_1}{n}.$$

This when simplified becomes

$$\bar{X} = X_0 + c \left(\frac{\sum_1^k f_1 x_1}{n} \right) \qquad \text{II. 11. 8.}$$

To illustrate II. 11. 8., we may use the frequency distribution given in table II. 1.

Table II. 2

SPEED IN MILES PER HOUR OF FREE MOVING VEHICLES ON SEPTEMBER 16, 1939, IN OAKLAWN, ILLINOIS ON U.S.H. 12 and 20 AT A POINT ONE MILE EAST OF HARLEM AVENUE

Speed in miles per hour	*Number of Vehicles*	$X - X_0 = S - S_0$	$\frac{S - S_0}{c}$	
$X = S$	f	s'	s	fs
70–74	0	30	6	0
65–69	0	25	5	0
60–64	2	20	4	8
55–59	15	15	3	45
50–54	14	10	2	28
45–49	29	5	1	29
40–44	74	0	0	0
35–39	60	–5	–1	–60
30–34	63	–10	–2	–126
25–29	29	–15	–3	–87
20–24	6	–20	–4	–24
15–19	8	–25	–5	–40
	300			–227

Substituting in II. 11. 8. the necessary values from Table II. 2., we find

$$\overline{X} = X_0 + c\left(\frac{\sum_1^k f_1 x_1}{n}\right)$$

becomes

$$\overline{X} = 42 + 5\left(\frac{-227}{300}\right) = 38.2. \qquad \text{II. 11. 9.}$$

This result is approximate in that in addition to its possessing a sampling error and an experimental error, it possesses a grouping error. These errors will be discussed later.

This arithmetic mean speed of 38.2 miles per hour is the estimate of the *probable* or *expected* speed of a vehicle at the highway point observed. What we wish to know about the mean speed is first, whether or not it is reliable and second, the range of speeds above

or below it. Is 38.2 miles per hour characteristic for all vehicles and if so, to what extent? We are able, with measures of dispersion, to find the answers to these questions. After doing this, we must look for a rational explanation of the agreement between the statistically obtained values and the actual facts; we must also determine what these facts mean. Were different types of vehicles observed or was the variety of speeds due to drivers with different desires or different abilities in driving, or to some other cause? This will be discussed and illustrated in Chapter IV.

II. 12. *Measure of Central Tendency.* A measure of *central tendency* is sometimes thought of as a characterizing or *descriptive value*, a *norm* or a *typical value*. It is always an average. But an average in itself is not necessarily a measure of central tendency. For this to be true, the average must agree fairly closely with all of the values from which it is obtained.

II. 13. *Mathematical Expectation or Expected Value of a Variable.* The expected value of a particular value X_1 of the variable X is the product of X_1 and the probability, p_1 that X takes the value X_1. If $E(X_1)$ denotes the expected value of X_1, then

$$E(X_1) = p_1 X_1 \qquad \text{II. 13. 1.}$$

Since the expected value of a sum is the sum of the expected values, it follows that the expected value E (X) of a variable X which may assume a set of values X_1 ($i = 1, 2, \ldots, n$) with corresponding probabilities p_1 ($i = 1, 2, \ldots, n$) is

$$E(X) = \sum_{1}^{n} p_1 X_1 \qquad \text{II. 13. 2.}$$

II. 14. *Deviation from Arithmetic Mean.* An important characterizing property of the arithmetic mean is that the algebraic sum of the deviations of the values from the arithmetic mean is equal to zero. This property is true for no other average.

To illustrate: Let it be required to find the mean weight of four men, who weigh respectively 128, 140, 150, and 190 pounds. Their arithmetic mean weight is

$$\bar{X} = \frac{128 + 140 + 150 + 190}{4} = 152 \text{ lbs.}$$

The differences between the individual weights of these four men and their arithmetic mean weight are:

Weights	Algebraic Differences
X	$X - \overline{X}$
190	38
150	— 2
140	— 12
128	— 24
	Sum = 0

The above demonstration may be stated in the form of a Theorem: *The sum of the algebraic differences between the values of a variable X and their arithmetic mean* $\overline{X}$ *is equal to zero.*

Let X_i ($i = 1, 2, \ldots, k$) be the values of the variable X, let f_i ($i = 1, 2, \ldots, k$) be the corresponding frequencies and let $\overline{X}$ be the arithmetic mean. Then

$$\sum_1^k{}_i f_i (X_i - \overline{X}) = \sum_1^k{}_i f_i X_i - \overline{X} \sum_1^k{}_i f_i.$$

But

$$\sum_1^k{}_i f_i = n \text{ and } \sum_1^k{}_i f_i X_i = n\overline{X},$$

Hence

$$\sum_1^k{}_i f_i (X_i - \overline{X}) = n\overline{X} - n\overline{X} = 0.$$

This Theorem may be expressed in terms of mathematical expectation as follows: *The expected value* $E\{X - E(X)\}$ *of the deviations of a variable from its expected value* $E(X)$ *is zero,* that is:

$$E\{X - E(X)\} = 0 \qquad \text{II. 14. 1.}$$

Another characteristic of the arithmetic mean is its *additive property.* The meaning of this property may be made clear by finding the mean of two sets of given values. Let the first set be 115, 128, 140 and the second be 150, 190.

The arithmetic mean of the first set is $\frac{115 + 128 + 140}{3} = 127\ 2/3$

and of the second set is $\frac{150 + 190}{2} = 170$. The arithmetic mean of the composite of the two sets is $\frac{115+128+140+150+190}{5} = 144\frac{3}{5}$.

But the weighted arithmetic mean of the two arithmetic means is

$$\frac{3\,(127\frac{2}{3}) + 2\,(170)}{3 + 2} = 144\frac{3}{5}.$$

This illustrates a theorem: *The arithmetic mean of the sum of two variables is the weighted arithmetic mean of their arithmetic means. Symbolically*: If $\overline{X}_1$ is the arithmetic mean of the first set having n_1 values and $\overline{X}_2$ is the arithmetic mean of the second set having n_2 values and if $\overline{X}_{\bar{x}_1 + \bar{x}_2}$ is the weighted arithmetic mean of the two arithmetic means, then

$$\overline{X}_{\bar{x}_1 + \bar{x}_2} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{n_1 + n_2} = \overline{X}, \qquad \text{II. 14. 2.}$$

where $\overline{X}$ is the arithmetic mean of the $n_1 + n_2$ values. This may be generalized to any number of variables.

In terms of *expected values* the theorem is stated as follows: *The expected value of the sum of two variables is the sum of their expected values*, that is:

$$E\,(X_1 + X_2) = E\,(X_1) + E\,(X_2). \qquad \text{II. 14. 3.}$$

To illustrate another theorem, reconsider the set of values 115, 128, 140. If we multiply each value by 2, we have the values 230, 256, 280. The arithmetic mean of 115, 128, 140 each multiplied by 2 is

$$\frac{230 + 256 + 280}{3} = 2\left\{\frac{115 + 128 + 140}{3}\right\} = 2\,(127\tfrac{2}{3})$$

The theorem is: *The arithmetic mean of a constant times a variable is equal to the constant times the arithmetic mean of the variable.*

In terms of expected values the theorem is: *The expected value of a constant times a variable is equal to the product of the constant by the expected value of the variable*, that is:

$$E\,(cX) = cE\,(X) \qquad \text{II. 14. 4.}$$

Let us reconsider the arithmetic mean, namely:

$$\overline{X} = \frac{\sum_1^k f_1 X_1}{n} = \frac{f_1}{n} X_1 + \frac{f_2}{n} X_2 + \ldots\ldots + \frac{f_k}{n} X_k$$

where $\sum_1^k \frac{f_1}{n} = 1$.

It is important to note that the coefficients of the X_1, namely, the f_i/n, are the relative frequencies of occurrence of these values.

But from the definition of statistical probability (see Chapter III), the limiting values of the f_i/n, as n becomes large beyond all bounds, are the p_1, where p_1 is the probability of occurrence of a value X_1 of X among a set of mutually exclusive values X_1.
Symbolically:

$$E(X) = \lim_{n \to \infty} \overline{X} = \lim_{n \to \infty} \frac{\Sigma f_1 X_1}{n} = \Sigma p_1 X_1 \qquad \text{II. 14. 5.}$$

where $p_1 X_1$ is the expected value of a particular value X_1 of X and $\Sigma_1 p_1 X_1$ is the sum of the expected values of the different particular values X_1 of X. But the sum of expected values is the expected value of the sum, and is called the *mathematical expectation*. It is also known as the *probable* or *expected value* of the variable.

It also follows from II. 14. 5. that the arithmetic mean $\overline{X}$ of a sample is an approximation to the *probable* or *expected value*, namely, the *true* or *universe value*.

The arithmetic mean is most important in estimating and predicting. The arithmetic mean $\overline{X}$ of a sample is the *unbiased* estimator (a value whose expected value is the true value) of the true mean of the population—the latter being E (X).

To illustrate: Suppose we have a considerable number of observations of the speeds in miles per hour of vehicles passing a given point. These may vary, say, from 19 miles per hour up to 70 miles per hour. Suppose we wish to answer the question: At what speed in miles per hour will a vehicle pass this point? The answer definitely is the *expected value* if we have the "universe", or the *arithmetic mean* if we have a random sample of the observed speeds. The arithmetic mean is the only one of the averages for a set of measure-

ments that is an expected value. Furthermore, no quantity is of any real value for predicting purposes unless it is a *probable* or *expected* value or unless as determined from a sample it is an *optimum* or *unbiased* estimator. An optimum estimator is one that is consistent, efficient, and sufficient.

Another important theorem concerned with expected values is: *The expected value of the product of two mutually independent variables is the product of their expected values. To illustrate*:

Toss three pennies and throw three dice. The number of heads occurring with the corresponding probabilities is shown in Table II.3. Likewise, the number of one spots occurring with the corresponding probabilities is shown in Table II.3.

Table II.3

Pennies		*Dice*	
No. of Heads X	*Probability* p_1	*No. of One Spots* Y	*Probability* p_2
0	$1/8$	0	$125/216$
1	$3/8$	1	$75/216$
2	$3/8$	2	$15/216$
3	$1/8$	3	$1/216$

Table II.4

EXPECTED VALUES

Pennies		*Dice*	
X	$p_1 X$	Y	$p_2 Y$
0	0	0	0
1	$3/8$	1	$75/216$
2	$6/8$	2	$30/216$
3	$3/8$	3	$3/216$
E (X)	$3/2$	E (Y)	$1/2$

In Table II.4 is shown the expected number of times for the different possibilities for number of heads occurring as well as the expected number of heads. Also, there is shown the expected number of times for the different possibilities for number of one spots occurring as well as the expected number of one spots.

Table II.5 lists for the compound event the expected number of times for the different possibilities for number of heads and one spots occurring as well as the expected number of heads and one spots.

Table II.5
EXPECTED VALUES

Dice and Pennies

Heads X	*One Spot* Y	*Compound Probability* $p_1 p_2$	X Y $p_1 p_2$
0	0	$^{125}/_{1728}$	0
0	1	$^{75}/_{1728}$	0
0	2	$^{15}/_{1728}$	0
0	3	$^{1}/_{1728}$	0
1	0	$^{375}/_{1728}$	0
1	1	$^{225}/_{1728}$	$^{225}/_{1728}$
1	2	$^{45}/_{1728}$	$^{90}/_{1728}$
1	3	$^{3}/_{1728}$	$^{9}/_{1728}$
2	0	$^{375}/_{1728}$	0
2	1	$^{225}/_{1728}$	$^{450}/_{1728}$
2	2	$^{45}/_{1728}$	$^{180}/_{1728}$
2	3	$^{3}/_{1728}$	$^{18}/_{1728}$
3	0	$^{125}/_{1728}$	0
3	1	$^{75}/_{1728}$	$^{225}/_{1728}$
3	2	$^{15}/_{1728}$	$^{90}/_{1728}$
3	3	$^{1}/_{1728}$	$^{9}/_{1728}$

$E(XY) = {}^{1296}/_{1728} = {}^{3}/_{4}$

From the above tables, it is seen that $[E(X) = \frac{3}{2}]$ $[E(Y) = \frac{1}{2}]$ $= [E(XY) = \frac{3}{4}]$ which symbolically is,

$$E(XY) = E(X)\,E(Y). \qquad \text{II.14.6.}$$

In the case of two samples of data: *The arithmetic mean of the product of two mutually independent variables is the product of their arithmetic means.*

This theorem may be generalized to any number of mutually independent variables.

II. 15. *The Deviations from Any Arbitrary Value.* The arithmetic mean of all the deviations from any arbitrary number, added to that number is the arithmetic mean of the values. This theorem may be explained by considering the weights of five persons who weigh respectively 135, 175, 180, 185, 190. Suppose we select $X_0 = 180$ as the arbitrary number, then

X	f	$x'' = X - X_0$
135	1	− 45
175	1	− 5
180	1	0
185	1	5
190	1	10
	n = 5	− 35

and $\overline{X} = 180 - \frac{35}{5} = 173$.

This is a much shorter method than adding all the items and dividing by their number.

Symbolically the theorem may be expressed as

$\overline{X} = X_0 + \Sigma x''/n$

where

X_0 = any arbitrary value but usually a *guessed mean* meaning that it is as near the actual mean as can be estimated.

x'' = deviation of each value from X_0, the estimated mean.

n = number of cases (individual values).

II. 16. *Mean Values in General. A Mean Value* in general may be thought of as the centroid of a frequency diagram. Let $y = f(x)$ be continuous in the x—interval (a, b).

Divide (a, b) into n equal parts, of length Δx and let y_i ($i = 1, 2, \ldots, n$) be the value taken by y in the ith part. The arithmetic mean of the numbers $y_1, y_2, \ldots, y_n$, that is

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_i + \ldots + y_n}{n} \qquad \text{II. 16. 1.}$$

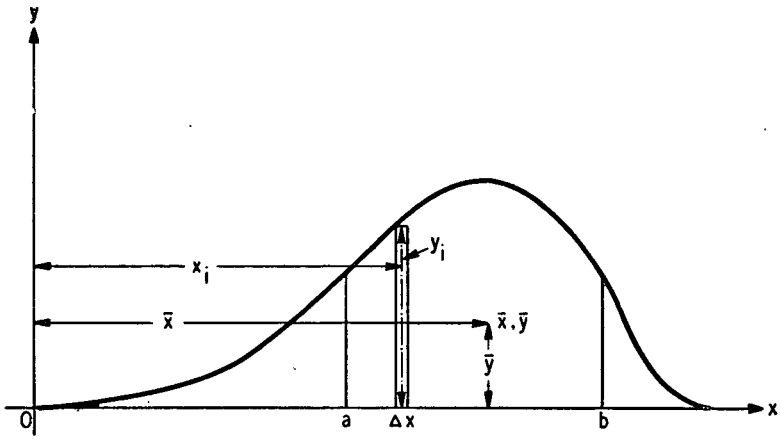


FIGURE II. 8

GRAPHICAL REPRESENTATION OF THE MEAN VALUE

will approach a definite limit as n tends to infinity. If the numerator and denominator of II. 16. 1. are multiplied by Δx, its form is changed to

$$\frac{y_1\Delta x + y_2\Delta x + \ldots + y_i\Delta x + \ldots + y_n\Delta x}{n\Delta x} \qquad \text{II. 16. 2.}$$

But $n\Delta x = b - a$ and the area A under the curve between the limits a and b is

$$A = \underset{\substack{\Delta x \to 0 \\ n \to \infty}}{\text{Limit}} (y_1\Delta x + y_2\Delta x + \ldots + y_i\Delta x + \ldots + y_n\Delta x)$$

$$= \int_a^b d\,A = \int_a^b y\,d\,x.$$

Hence, the mean value $\bar{y}$ of y is

$$\bar{y} = \underset{n \to \infty}{\text{Limit}} \frac{\sum_1^n y_i \Delta x}{n\Delta x} = \frac{\int_a^b y\,d\,x}{b - a} \qquad \text{II. 16. 3.}$$

Likewise, the mean value $\bar{X}$ of X is found by taking first moments about the y-axis, namely:

$$A\,\bar{X} = \int_a x\,d\,A, \quad \text{whence}$$

$$\bar{X} = \frac{\int_a^b x\,y\,dx}{\int_a^b y\,dx}. \qquad \text{II. 16. 4.}$$

II. 16. 2. may be interpreted as the average weight of $n\Delta x$ objects having various weights where Δx objects have a weight of y_1, Δx have a weight of y_2, ...

II. 16. 3. may also be obtained by the use of *moments* as illustrated in Figure II. 8. Here $y_1\Delta x$ objects have, say, a distance x_1. The moment of $y_1\Delta x$ about the y-axis is $x_1y_1\Delta x$. The moment of the whole, if $\bar{x}$ is its distance, is $\bar{x}\,(b - a)$ and also $\sum_1^n x_1\,y_1\,\Delta x$.

$$\text{Hence:} \quad \bar{x}\,(b - a) = \lim_{\Delta x \to 0} \sum_1^n x_1\,y_1\,\Delta x = \int_a^b x\,y\,dx,$$

$$\text{whence:} \quad \bar{x} = \lim_{\Delta x \to 0} \frac{\sum x_1\,y_1\,\Delta x}{b - a} = \frac{\int_a^b x\,y\,dx}{b - a}.$$

The notion of mean is readily extended to functions of two or more variables. To see this generalization, the reader is referred to any book on Calculus or Mechanics.

II. 17. *The Mode.* The mode or modal value of a variable is that value of a variable which occurs most frequently, if such a value exists. It is the *most probable value*, or in other words, the value for which the frequency is a maximum. The expression *most probable value* when it refers to the number of successes in n trials is used in the general theory of probability to designate the number to which there corresponds a larger probability of occurences than to any other number. The point at which the frequency is most dense is the abscissa of the maximum point of the frequency curve and can be determined accurately only from the equation of the curve.

For a given grouping the class mark of the maximal class frequency is called the *empirical mode.*

An approximation to the mode may be obtained by passing a parabola through the midpoints of the upper bases of the modal class and the two adjacent classes. Figure II. 9. shows three such points h, i, j.

The general equation of a parabola with its axis parallel to the y-axis is

$$y = \alpha + \beta x + \gamma x^2. \qquad \text{II. 17. 1.}$$

In Figure II. 9., take the origin at the point 0, namely, at the lower limit of the modal class. Let c equal the class interval and $\Delta_1 = OG$ and $\Delta_2 = ED$. When $x = -c/2$, $y = 0$; $x = c/2$, $y = \Delta_1$; $x = 3c/2$, $y = \Delta_1 - \Delta_2$. Substitute these values for x and y in II. 17. 1. and

$$\begin{aligned} 0 &= \alpha - \beta\,(c/2) + \gamma\,(\tfrac{c^2}{4}) \\ \Delta_1 &= \alpha + \beta\,(c/2) + \gamma\,(\tfrac{c^2}{4}) \\ \Delta_1 - \Delta_2 &= \alpha + \beta\,(3c/2) + \gamma\,(9\tfrac{c^2}{4}) \end{aligned} \qquad \text{II. 17. 2.}$$

Solving these equations for α, β, γ,

$$\alpha = \frac{5\Delta_1 + \Delta_2}{8}; \qquad \beta = \frac{\Delta_1}{c}; \qquad \gamma = -\frac{\Delta_1 + \Delta_2}{2c^2} \qquad \text{II. 17. 3.}$$

The maximum point on the curve $y = \alpha + \beta x + \gamma x^2$ is found by setting

$$\begin{aligned} dy/dx &= \beta + 2\gamma x = 0 \\ d^2y/dx^2 &= 2\gamma < 0 \end{aligned} \qquad \text{II. 17. 4.}$$

From II. 17. 4.,

$$\begin{aligned} x &= -\beta/2\gamma \\ \gamma &< 0 \end{aligned} \qquad \text{II. 17. 5.}$$

Substituting the values for β and γ from II. 17. 3. in II. 17. 5.,

$$x = \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) c \qquad \text{II. 17. 6.}$$

The quantity found for x in II. 17. 6. when added to the lower limit of the modal class is the approximate value of the mode, namely

$$\text{Mode} = l_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) c \qquad \text{II. 17. 7.}$$

where

l_1 = lower limit of the class with maximum frequency.
$\Delta_1 = f_0 - f_1$ (See Figure II. 9.)
$\Delta_2 = f_0 - f_r$ (See Figure II. 9.)

In Table II. 1., $f_0 = 74$, $f_1 = 60$, $f_r = 29$. Substituting these values in II. 17. 7., we obtain

$$\text{Mode} = 39.5 + \left(\frac{14}{14 + 45}\right) 5 = 40.7. \qquad \text{II. 17. 8.}$$

The graphical counterpart of the solution just given for finding the mode is as follows. Consider the distribution given in Table II. 1.

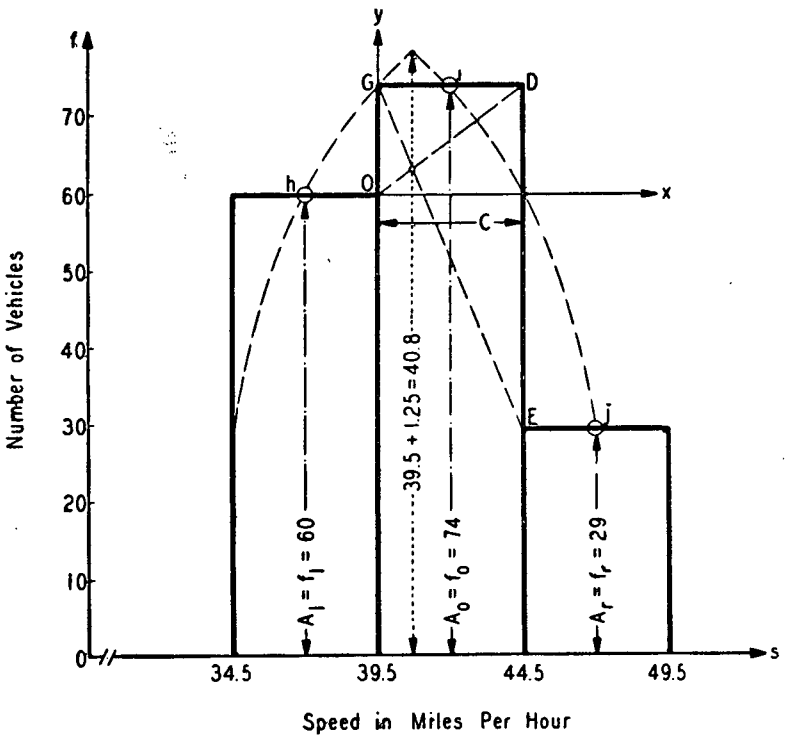


FIGURE II. 9

GRAPHICAL SOLUTION FOR FINDING THE MODAL VALUE OF A SET OF OBSERVATIONS

From this table select the modal class and the class adjacent to it on either side of it and for these three classes plot on graph paper these three frequency rectangles as illustrated in Figure II. 9.

Connect the points G and E with a straight line and the points O and D with a straight line. Then from the point of intersection of these two lines drop a perpendicular to the horizontal axis. The number read on the horizontal scale at the point where this perpendicular cuts the horizontal scale is the graphical solution of the mode. In this case it is 40. 8. Comparing the value of the mode found graphically with the value of the mode just found arithmetically, it is seen that the difference is 0.1, which is negligible.

It is not difficult to show, that the abscissa of the point of intersection of the lines joining OD and GE is

$$x = \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) c$$

which proves that the graphical solution given is theoretically the same as the analytical.

It is obvious that for most practical purposes since graphically the value of the mode can be obtained with slight error the graphical solution of the mode will suffice. This result means that the *most probable speed* of a vehicle at the point observed is 40.7 miles per hour. In other words, more vehicles pass this point at a speed of 40.7 miles per hour than at any other speed.

II. 18. *Median.* The median of a variable is a number which is such that half of the measurements have a value less than it and the other half have a value greater than it. It is thus the abscissa of the point the vertical through which divides the total area under the frequency curve or frequency rectangles into two equal parts. To compute the median of a sample set of n values of the variable, compute the abscissa of a point, the vertical through which divides the total area of the frequency rectangles into two equal parts.

Illustration:

From columns (1) and (6) in Table II. 1., and from Figure II. 10., it is seen that the sum of the frequencies (sum of the areas) of the classes up to X = 34.5 is 106 and the sum of the frequencies (sum of the areas) of the classes up to X = 39.5 is 166. But one-half the total frequency is 150 which is between 106 and 166. Hence the

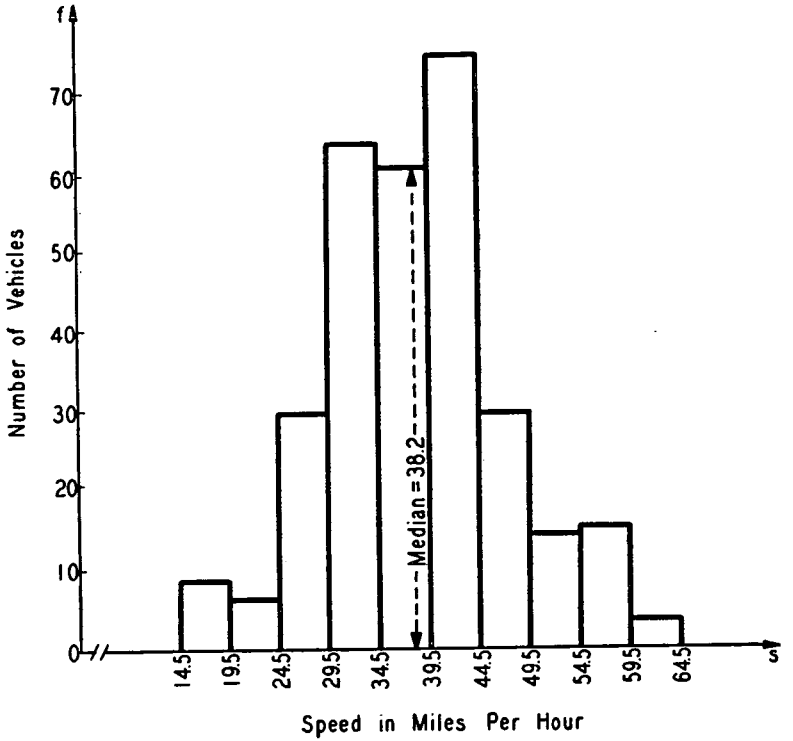


FIGURE II. 10

MEDIAN VALUE OF OBSERVED VEHICLE SPEEDS

median value, by definition, lies between X = 34.5 and X = 39.5 at a point which is the same proportion of the distance from X = 34.5 to X = 39.5 as 150 is from 106 to 166.

Symbolically it is seen that

$$\text{Median} = l_1 + \left(\frac{n/2 - f_{cl_1}}{f_m}\right) c \qquad \text{II. 18. 1.}$$

where

l_1 = lower bound of class in which median value falls.

n = total frequency.

f_{cl_1} = cumulative frequency to lower limit of class in which median value lies.

f_m = frequency of class in which median lies.
c = length of class interval.
Hence for the given distribution

$$\text{Median} = 34.5 + \left(\frac{150 - 106}{60}\right) 5 = 38.2 \qquad \text{II. 18. 2.}$$

II. 19. *Quantiles:* Quantiles are location and division numbers. They, like the median, divide the distribution into sections. There are many quantiles, but we shall mention and briefly discuss only those frequently used. There are the *quartiles* (*quarters*), *quintiles* (*fifths*), *deciles* (*tenths*), and *percentiles* (*hundredths*). The method of finding them is similar to that of finding the median.

A quantile value (or percentile) is a number such that the specified quantile (percentage) proportion of cases have a measure less than it and the remainder have a measure greater than it. Symbolically,

$$\text{Quantile} = l_1 + \left(\frac{k\,n - f_{cl_1}}{f_q}\right) c \qquad \text{II. 19. 1.}$$

where

l_1 = lower bound of class in which quantile value falls.
k = proportion of cases below specified quantile value.
n = total frequency.
f_{cl_1} = cumulative frequency to lower limit of class in which quantile value lies.
f_q = frequency of class in which the specified quantile value lies.

To illustrate: It is desired to find the lower quartile Q_1 or the 25th percentile and the upper quartile Q_3 or the 75th percentile.

In the former case, $k = \frac{1}{4}$, and from columns (1) and (6) of Table II. 1., it is seen that $f_{cl_1} = 43$ and $f_q = 63$ and $l_1 = 29.5$. Hence II. 19. 1. becomes

$$Q_1 = 29.5 + \left(\frac{\frac{1}{4}(300) - 43}{63}\right) 5 = 32.0 \qquad \text{II.19.2.}$$

In the latter case, $k = \frac{3}{4}$, it is seen that $f_{cl_1} = 166$ and $f_q = 74$ and $l_1 = 39.5$. Here II.19.1. becomes

$$Q_3 = 39.5 + \left(\frac{\frac{3}{4}(300) - 166}{74}\right) 5 = 43.5. \qquad \text{II.19.3.}$$

These two values mean that 25 per cent of the vehicles at the observed point had a speed less than 32.0 miles per hour and 25 per cent of the vehicles had a speed greater than 43.5 miles per hour.

If it is desired to know the 4th decile, then $k = 0.4$ in II.19.1. and if it is desired to know the thirty-second percentile, then $k = 0.32$. In other words the 4th decile means a speed such that 0.4 of the vehicles have a lower speed and 0.6 a higher speed and the thirty-second percentile means a speed such that 32 per cent have a lower speed and 68 per cent a greater speed.

Having found the values of the arithmetic mean, the median and the mode, what are the differences in their values and meanings? It can be proved that the median value always lies between the arithmetic mean and the mode such that either

$$\bar{X} \leq \text{Median} \leq \text{Mode} \text{ or}$$
$$\text{Mode} \leq \text{Median} \leq \bar{X} \qquad \text{II.19.4.}$$

For the distribution of Table II.1., it was found that $\bar{X} = 38.2.$, the Median $= 38.2.$, the Mode $= 40.7$ miles per hour. The apparent equality of the median and arithmetic mean in this sample is due primarily to grouping and sampling errors and to some extent due to experimental error. The modal value of 40.7 reveals that a greater proportion of the vehicles at the point observed travel at a speed greater than the probable or expected speed of 38.2 miles per hour. This observed tendency is important and can and must be explained from a subjective study. The other results show that 25 per cent of vehicles travelled with a speed less than 32.0 miles per hour and 25 per cent with a speed greater than 43.5 miles per hour and 50 per cent with a speed of from 32.0 to 43.5 miles per hour. The lower 25 per cent had a range in speed of $32.0 - 14.5 = 17.5$ miles per hour, the middle 50 per cent had a range of $43.5 - 32.0 = 11.5$ miles per hour, and the upper 25 per cent had a range in speed of $74.5 - 43.5 = 31.0$ miles per hour. Similarly, the second 25 per cent had a range in speed of $38.2 - 32.0 = 6.2$ miles per hour and the third 25 per cent a range of $43.5 - 38.2 = 5.3$ miles per hour. These results indicate rather plainly a lack of stability and uniformity in speeds due to drivers, type of vehicles, and topography at point observed.

II. 20. *Geometric Mean.* The geometric mean of a set of n positive measurements is the nth root of their product. If X_i ($i = 1, 2, \ldots . n$) are the n values for a variable X, the geometric mean,

$$\text{G.M.} = (\prod_{i}^{n} X_i)^{\frac{1}{n}} = (X_1 \cdot X_2 \ldots . \cdot X_n)^{\frac{1}{n}} \qquad \text{II.20.1.}$$

where Π is the symbol for the product.

For a frequency distribution,

$$\text{G.M.} = (X_1^{f_1} \cdot X_2^{f_2} \cdot \ldots . \; X_i^{f_i} \cdot \ldots . \cdot X_k^{f_k})^{\frac{1}{n}} \qquad \text{II. 20. 2.}$$

where $\sum_{1}^{k} f_i = n$. It is significant that the

$$\log. \text{G.M.} = \frac{f_1 \log X_1 + f_2 \log X_2 + \ldots . + f_k \log X_k}{n}$$

$$= \frac{\sum_{1}^{k} f_i \log X_i}{n} \qquad \text{II.20.3.}$$

This means that the logarithm of the geometric mean is the arithmetic mean of the logarithms of the measurements. Recalling the relationship between relative frequency and probability, it is evident that as the number of measurements is indefinitely increased the logarithm of the geometric mean becomes the probable or expected value of the logarithm of the variable X.

For analyzing a frequency distribution, the geometric mean has no immediate value. The geometric mean is the average of a set of rates and is the only average which is the average of a set of rates or the average of a set of things that behave like rates. Two examples will illustrate this property:

(1) A city had a population in 1900 of 100,000 and in 1910 of 120,000. What is the average annual rate of increase in population? This problem is analagous to a problem in compound interest where the amount, principal, and time are known and the rate of interest is to be found. Hence

$$P_n = P_0 (1 + r)^n \qquad \text{II.20.4.}$$

where

P_n = the population at the end of n years.
P_0 = the population at the beginning of the period.
n = number of time intervals.

Substitute the above values in II.20.4., then

$$120{,}000 = 100{,}000\,(1 + r)^{10}$$

Solving for r, it is found that

$$r = .0184 = 1.84\% \text{ change per annum.}$$

(2) Given the information shown in tabular form:

Community	*Native Born Inhabitants*	*Foreign Born Inhabitants*	*Ratio of Foreign Born to Native Born*	*Ratio of Native Born to Foreign Born*
A	a = 9000	c = 4500	c/a = 50%	a/c = 200%
B	b = 2000	d = 4000	d/b = 200%	b/d = 50%

It may be shown that the arithmetic mean is not *the* average rate of increase.

The arithmetic mean of the ratios of Foreign Born to Native born is

$$\frac{50\% + 200\%}{2} = 125\% = \frac{c/a + d/b}{2} = \frac{cb + ad}{2\,ab}.$$

The arithmetic mean of the ratios of Native born to Foreign born is

$$\frac{200\% + 50\%}{2} = 125\% = \frac{a/c + b/d}{2} = \frac{ad + bc}{2\,cd}$$

Since the product of these two results is not unity or 100%, they are illogical and the arithmetic mean is not the proper average to use.

The geometric mean of the ratios of Foreign born to Native born is

$$\text{G.M.} = \sqrt{.50 \cdot 2.00} = 1.00 = 100\% = \sqrt{c/a \cdot d/b} = \sqrt{cd/ab}.$$

The geometric mean of the ratios of Native born to Foreign born is

$$\text{G.M.} = \sqrt{2.00 \cdot .50} = 1.00 = 100\% = \sqrt{a/c \cdot b/d} = \sqrt{ab/cd}$$

The product of these two results is unity or 100%.

Now $\dfrac{c+d}{a+b}=\dfrac{4500+4000}{9000+2000}=\dfrac{8500}{11000}=.7727=77.27\%$ and

$$\frac{a+b}{c+d}=\frac{9000+2000}{4500+4000}=\frac{11000}{8500}=1.2941=129.41\%.$$

But $\dfrac{c+d}{a+b}\cdot\dfrac{a+b}{c+d}=1$ and .7727 times 1.2941 = 1.

Since the product of the ratios must be unity, it is seen that the geometric mean is *the* average rate.

II. 21. *Harmonic Mean.* The harmonic mean of a set of measures is the reciprocal of the arithmetic mean of the reciprocals of the measures.

Symbolically, if H M. is the harmonic mean,

$$\text{H.M.}=\frac{n}{f_1/x_1+f_2/x_2+\ldots.+f_k/x_k} \qquad \text{II. 21. 1.}$$

To illustrate: Suppose we have a vehicle that travels 25 miles per hour for 20 miles, then 30 miles per hour for 10 miles, then 50 miles per hour for 50 miles, then 40 miles per hour for 10 miles and finally, 12 miles per hour for 10 miles. What is the average speed of this vehicle for the 100 miles travelled ? It is the harmonic mean, namely,

$$\text{H.M.}=\frac{100}{20\,(1/25)+10\,(1/30)+50\,(1/50)+10\,(1/40)+10\,(1/12)}$$
$$=31.1 \text{ miles per hour.}$$

This average speed may be found by an arithmetic mean method if weights are properly chosen. If $\bar{X}'$ is the symbol for the average speed for an arithmetic mean method,

$$\bar{X}'=\frac{25\{(.04)\,(20)\}+30\{(.033)\,(10)\}+50\{(.02)(50)\}+40\{(.025)(10)\}+12\{(.083)\,(10)\}}{3.21}$$
$$=\frac{25\,(.8)+30\,(.333)+50\,(1)+40\,(.25)+12\,(.833)}{3.21}$$
$$=\frac{20.000+9.999+50.000+10.000+9.996}{3.21}=31.1 \text{ miles per hour}$$

where 0.8, 0.333, 1, 0.25, and 0.833 are the weights.

The latter method, while it solves the problem, is not as direct and simple as the harmonic mean. Of all the averages, the harmonic mean is the only one that is *the average time rate* or *the average* of things that behave like time rates.

II. 22. *Root Mean Square.* The *root mean square* R.M.S., σ, often called the *standard deviation* in statistics is similar to the *radius of gyration k* in mechanics. The radius of gyration of the area under a frequency curve about the ordinate through the center of gravity of that area is, in fact, equal to σ.

The physical meaning of radius of gyration is that it is a distance such that if all the mass of a body (or area) were concentrated at a point that distance from an axis of rotation it would have the same rotational effect as the actual distributed mass (area). It is also the root mean square of the radial distances of a set of n equal particles from an axis. In the same way, σ, the standard deviation of a frequency distribution (area) thought of as a set of n equal particles of area is the square root of the arithmetic mean of the squares of the radial distances of the several particles from the centroidal axis, that is, it is the R.M.S. as well as k with respect to the centroidal axis.

It is believed that a review of the significance of second moments and the radius of gyration k in mechanics will help to understand the corresponding terms in statistics.

Let A be any area and YY an axis through the centroid O as shown in Figure II. 11.

Let dA represent an element of area and let x be its distance from the centroidal axis YY.

The moment of inertia I_Y is by definition the sum of all the $x^2\,dA$, that is,

$$I_Y = \int_A x^2\,dA \qquad \text{II.22.1.}$$

and the radius of gyration,

$$k^2 = \frac{I_Y}{A} \qquad \text{II. 22. 2.}$$

If the moment of inertia of an area with respect to a centroidal

axis is known, the moment of inertia with respect to a parallel axis may be found as follows:

In Figure II.11., let Y'Y' be any axis parallel to YY and at a distance d from YY.

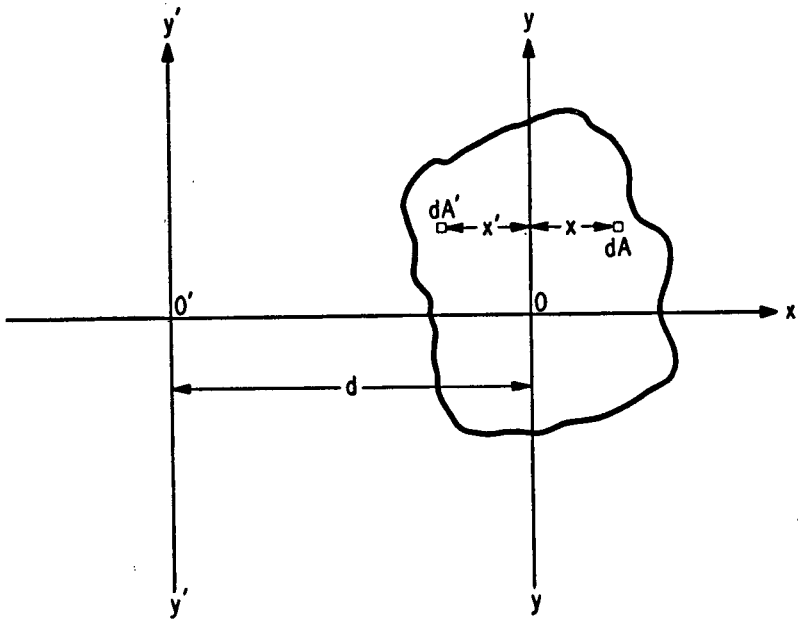


FIGURE II. 11
MOMENT OF INERTIA
OF AN AREA WITH RESPECT TO A PARALLEL AXIS

The moment of inertia of the element dA about Y'Y' is equal to $(x + d)^2\, dA$ and I_Y' for the total area is

$$\begin{aligned} I_Y' &= \int_A (x + d)^2\, dA \\ &= \int_A x^2\, dA + 2d \int_A x\, dA + d^2 \int_A dA \\ &= I_Y + Ad^2 \end{aligned} \qquad \text{II.22.3.}$$

since $$\int_A xdA = A\bar{x} = 0.$$

The fact that $\int_A xdA = 0$ may be comprehended if it is remembered that for every element dA on the right, there is an element (dA)′ at a distance x′ to the left, such that x′ (dA)′ = xdA. In other words, we may think of the area as being balanced about the centroidal axis.

The frequency diagram in statistics may be treated in the same manner as an area is treated in mechanics. The notation is slightly different and so is the point of view and interpretation as is shown in Figure II.12. Otherwise, the procedure is the same.

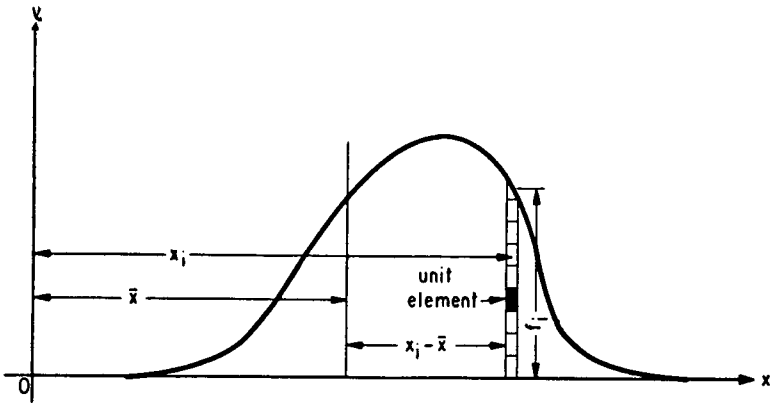


FIGURE II. 12
FREQUENCY DIAGRAM

Using the notation shown in Figure II.12.

$$\sigma^2 = k^2 = (1/n) \sum_1^n (x_1 - \bar{x})^2. \qquad \text{II.22.4.}$$

This may be written in the form

$$2\,\sigma^2 = 2\,k^2 = (1/n^2) \sum_{1j}^n (x_1 - x_j)^2 \qquad \text{II.22.5.}$$

We thus see that the standard deviation is (1) the square root of the arithmetic mean of the squares of the differences between the measurements and their arithmetic mean and (2) proportional

to the square root of an average of the square of the differences between the measurements taken two at a time where the constant of proportionality is $(1/\sqrt{2})$.

In the continuous case, we may write

$$E^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-y)^2\,dF(x)\,dF(y)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}dF(x)\,dF(y)\{x^2 - 2xy + y^2\}$$

$$= \int_{-\infty}^{\infty}x^2\,dF(x)\int_{-\infty}^{\infty}dF(y) - 2\int_{-\infty}^{\infty}x\,dF(x)\int_{-\infty}^{\infty}y\,dF(y)$$

$$+ \int_{-\infty}^{\infty}dF(x)\int_{-\infty}^{\infty}y^2\,dF(y)$$

$$= 2\mu_2' - 2\mu_1'^2 = 2\mu_2. \qquad \text{II. 22. 6.}$$

The square of the standard deviation is the *variance*. It is also the second moment about the mean. Variance is half the mean square of all possible variate differences without reference to deviations from a central value.

The arithmetic mean of the squares of the differences between the measurements and their arithmetic mean is equal to the arithmethic mean of the squares of the measurements minus the square of the arithmetic mean of the measurements.

Expressed mathematically, it is,

$$\frac{\Sigma(X-\bar{X})^2}{n} = \frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2 \qquad \text{II.22.7.}$$

which, if the measurements are 3, 5, 6, 9, 12 becomes

$$\frac{(3-7)^2 + (5-7)^2 + (6-7)^2 + (9-7)^2 + (12-7)^2}{5}$$

$$= \frac{3^2 + 5^2 + 6^2 + 9^2 + 12^2}{5} - \left(\frac{3+5+6+9+12}{5}\right)^2,$$

where 7 is the arithmetic mean of the measurements. This, upon simplification becomes $10 = 59 - 49 = 10$ which demonstrates II.22.4.

Also

$$\{(3-3)^2+(3-5)^2+(3-6)^2+(3-9)^2+(3-12)^2+(5-3)^2+(5-5)^2+(5-6)^2+(5-9)^2+(5-12)^2+(6-3)^2+(6-5)^2+(6-6)^2+(6-9)^2+(6-12)^2+(9-3)^2+(9-5)^2+(9-6)^2+(9-9)^2+(9-12)^2+(12-3)^2+(12-5)^2+(12-6)^2+(12-9)^2+(12-12)^2\} \div (5)(5) = \frac{500}{25} = 20 = 2(10).$$

Hence $2\sigma^2 = \sum_{ij}(x_i - x_j)^2$ becomes $2(10) = 20$

which demonstrates II.22.5.

In case we have k values of X_i and each value occurs several times, or in case we have a frequency distribution where X_i is the class mark of the ith class and f_i is the frequency of the ith class, it is convenient to write

$$\frac{\sum_1^k f_i (X_i - \bar{X})^2}{n} = \frac{\sum_1^k f_i X_i^2}{n} - \frac{\sum_1^k f_i X_i^2}{n} \qquad \text{II.22.8}$$

Considering the limit definition of probability, namely,

$$\lim_{n\to\infty} f_i/n = p_i,$$

we have

$$E[(X - E(X)^2] = E(X^2) - [E(X)]^2 \qquad \text{II.22.9.}$$

which in words is the theorem: *The expected value of the square of the deviation of the variable from the expected value is equal to the expected value of the square of the variable minus the square of the expected value of the variable.*

In the special case when the class intervals are all *equal*, we may use the value of X_i from II.11.7. in II. 22.8 and then

$$\sigma^2 = \frac{\sum_1^k f_i (X_i - \bar{X})^2}{n} = c^2 \left\{ \frac{\sum_1^n f_i x_i^2}{n} - \left(\frac{\sum_1^n f_i x_i}{n} \right)^2 \right\} \qquad \text{II.22.10.}$$

To illustrate, consider the distribution given in columns (1) and (2) of Table II.1. and the tabulation as shown in Table II.6.

Making use of formula II.22.10., namely,

$$\sigma = c \sqrt{\frac{\Sigma f s^2}{n} - \left(\frac{\Sigma f s}{n}\right)^2}$$

where now $X = S$ and $x = s$,

Table II.6.

SPEED IN MILES PER HOUR OF FREE MOVING VEHICLES ON SEPTEMBER 11, 1939, IN OAKLAWN, ILLINOIS ON U.S.H. 12 AND 20 AT A POINT ONE MILE EAST OF HARLEM AVE.

Speed in miles per hour	*Number of Vehicles*			
S	f	s	fs	fs^2
70–74	0	6	0	0
65–69	0	5	0	0
60–64	2	4	8	32
55–59	15	3	45	135
50–54	14	2	28	56
45–49	29	1	29	29
40–44	74	0	0	0
35–39	60	— 1	— 60	60
30–34	63	— 2	— 126	252
25–29	29	— 3	— 87	261
20–24	6	— 4	— 24	96
15–19	8	— 5	— 40	200
	300		— 227	1121

Substitute the indicated values from Table II.6. in II.22.10, then

$$\sigma = 5\sqrt{\frac{1121}{300} - \left(\frac{-227}{300}\right)^2}$$

$$= 5\sqrt{3.7367 - 0.5726} = 5\,(1.779)$$

$$= 8.9 \text{ miles per hour.}$$

This means that we would expect the speed of a random vehicle to be somewhere between $38.2 - 8.9$ and $38.2 + 8.9$ miles per hour, namely, between 29.3 and 47.1 miles per hour.

From an examination of the distribution of speeds, we find that approximately 71 per cent of the vehicles had a speed between 29.3 and 47.1 miles per hour. Hence this relative frequency tells us that we are approximately 71 per cent certain that a random vehicle will pass the intersection with a speed between 29.3 and 47.1 miles per hour.

If on the other hand, we use the expected speed of 38.2 miles per hour as our estimate, it is 71 per cent certain that we will be in error by at most $\sigma/\overline{X} = 8.9/38.2 = 23.3$ per cent. On the other hand, it is 29 per cent certain that the error is at least 23.3 per cent.

This indicates that there is marked variability in speeds and there does not appear to be a typical speed at all for this point on the highway.

II. 23. *Centra Harmonic Mean.* The centra harmonic mean is a measure of relative dispersion. It is the arithmetic mean of the squares of the measures from an arbitrary origin divided by the arithmetic mean of the measures. Symbolically if C.H.M. is the centra harmonic mean, then

$$\text{C.H.M.} = \sum_1^n x_1^2 \Big/ \sum_1^n x_1. \qquad \text{II.23.1.}$$

The centra harmonic mean per se is of very little use today. However, a quantity similar to it, namely the *coefficient of variability* is useful as a measure of relative dispersion or a measure of per cent of error. If C.V. is the symbol for coefficient of variability, then, by definition

$$\text{C.V.} = \left\{\frac{\sum_1^n (X_1 - \overline{X})^2}{n}\right\}^{\frac{1}{2}} \div \frac{\sum_1^n X_1}{n} = \frac{\sigma}{\overline{X}} \qquad \text{II.23.2.}$$

In II.22. the C.V. was interpreted for the distribution given in Table II.1.

II. 24. *Mean or Average Deviation.* The mean or average deviation from an average is the A.M. of the deviations treating them all as positive. The deviations may be taken from any average, but the mean deviation is least when the median is the origin.

In case of a normal distribution with origin at the arithmetic mean or median, the mean deviation is the abscissa of the centroid of area under the right hand half of the frequency curve and its value is $0.7978\,\sigma = 0.8\,\sigma$ approximately. Assume the frequency for each class concentrated at the center of class as shown in

Figure II.13. Let the distances of these centers from the center of the class containing the median be d_1, d_2, ...

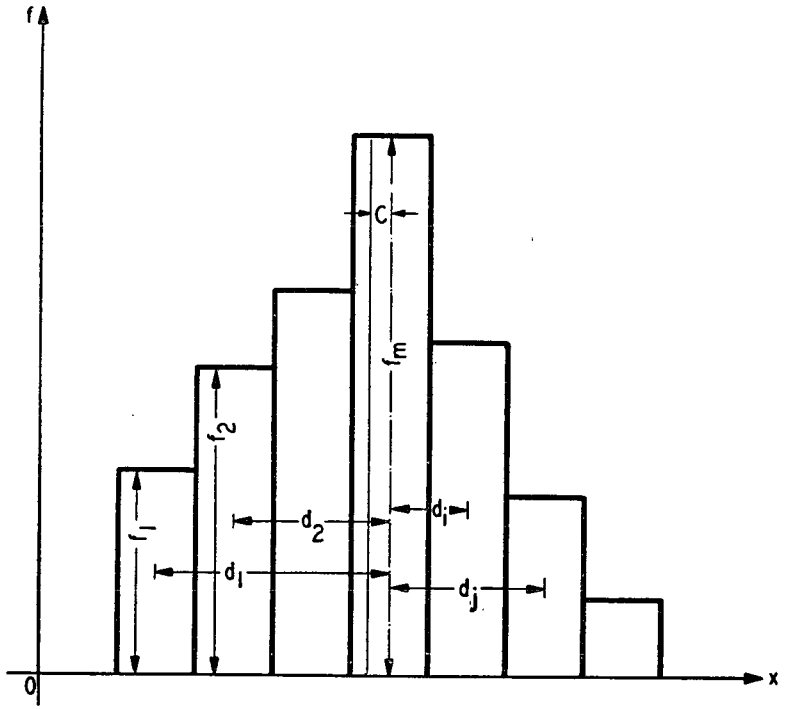


FIGURE II.13

MEAN OR AVERAGE DEVIATION OF A SET OF OBSERVATIONS

and let the corresponding class frequencies be f_1, f_2, ... so that the sum of moments about the median is $f_1d_1 + f_2d_2 + \ldots + f_nd_n$. Ignore the class containing the median for the present. All the products whose deviations lie below (to the left of) the median have deviations too short by an amount C and those above (to the right) are too long by an amount C. Next consider the sum of the deviations below the median class and above the median class. If N_a is the number of observations above and N_b the number below the median class, then we have as a first correction

$$(N_b - N_a)\, C. \qquad \text{II.24.1.}$$

If N_m is number of observations in the median class and if we assume these N_m observations uniformly distributed over the interval, then $(.5 + C)\,N_m$ cases are below and $(.5 - C)\,N_m$ are above the median. With a uniform distribution, the sum of these deviations below the median is

$$\frac{(.5 + C)^2\,N_m}{2} \text{ and above the median } \frac{(.5 - C)^2\,N_m}{2}$$

Hence the sum of all the deviations of the N_m values is

$$\frac{(.5 + C)^2\,N_m}{2} + \frac{(.5 - C)^2\,N_m}{2} = (.25 + C^2)\,N_m. \qquad \text{II.24.2.}$$

which is the second correction.

Let us now find the mean deviation from the median for the distribution given in Table II.1.

Table II.7.

SPEED IN MILES PER HOUR OF FREE MOVING VEHICLES ON SEPTEMBER 16, 1939, IN OAKLAWN, ILLINOIS, ON U.S.H. 12 AND 20 AT A POINT ONE MILE EAST OF HARLEM AVE.

X = S	f	x = s	f|s|*
70–74	0	7	0
65–69	0	6	0
60–64	2	5	10
55–59	15	4	60
50–54	14	3	42
45–49	29	2	58
40–44	74	1	74
35–39	60	0	0
30–34	63	− 1	63
25–29	29	− 2	58
20–24	6	− 3	18
15–19	8	− 4	32
	300 = n		415

* The symbol |s| means the numerical value of s which is always positive or zero.

Correction (1): $(N_b - N_a)\,C = (106 - 134)\,(1.2) =$	$-\ 33.6$
Correction (2): $(.25 + C^2)\,N_m = (.25 + 1.44)\,(60) =$	101.4
Sum of deviations for classes other than median class $=$	415.0
Sum of all deviations	482.8

$$\text{Mean Deviation} = \frac{482.8}{300} = 1.609 \text{ class intervals}$$

$$= 8.05 = 8.1 \text{ miles per hour.}$$

This means that the expected value of the difference between the speed of a vehicle and the median value of speeds is 8.1 miles per hour.

Given N values. Choose a certain number as origin such that x of the values will be greater than this number. Then $N - x$ will be less than the selected number. Let the deviations from the selected number (average) as origin be Δ. Displace the original origin by K units so that it is exceeded by only $x - 1$ values. Then $N - (x - 1)$ of the values will be less than the new number. By this change, the sum of the deviations in excess of the selected number is decreased by Kx, while the sum of the deviations less than the selected number is increased by $(N - x)\,K$. If Δ' is the new sum of deviations, then

$$\Delta' = \Delta + (N - x)\,K - Kx \text{ and}$$

$$\Delta' = \Delta + (N - 2x)\,K.$$

$$\text{If } x = N/2;\ \ \Delta' = \Delta.$$

$$\text{If } x > N/2;\ \ \Delta' < \Delta.$$

This proves that the sum of the numerical values of the deviations from the median is a minimum.

II. 25. *Moments and Mathematical Expectation of Powers of a Variable.*

The moments of a distribution are the expected values of the powers of the stochastic variable which has the given distribution. The term "moment" has been taken over by the statistician from mechanics. In mechanics, moment is a measure of a force with respect to its tendency to produce rotation. In statistics moments characterize the parameters of the distribution law which are the properties that describe for interpretation and meaning the law of behavior of the attribute that is being measured and studied.

The late Karl Pearson (*Biometrika*, Vol. 9, pp. 1–10) has shown that all the constants of a frequency distribution are expressible in terms of higher product moments. In the case of two variates, they are defined by

$$\nu_{q,\,q'} = \sum_{1}^{n}{}_{ij} \{p_{ij}\, x_1^{q}\, y_j^{q'}\} \qquad \text{II.25.1.}$$

for an arbitrary origin. If the origin is at the mean, namely, at P $(\bar{x}, \bar{y})$, then

$$\mu_{q,\,q'} = \sum_{1}^{n}{}_{ij} \{p_{1j}\,(x_1 - \bar{x})^{q}\,(y_j - \bar{y})^{q'}\} \qquad \text{II.25.2.}$$

In case of a single variable, the k th moment of a continuous variable x about an arbitrary origin denoted by ν_k is

$$\nu_k = \mathrm{E}\,(x^k) = \int_a^b x^k\, f\,(x)\, dx \qquad \text{II.25.3.}$$

and in the case of a discontinous variable x

$$\nu_k = \mathrm{E}\,(x^k) = \sum_{1}^{n}{}_{i} p_1\, x_1^k. \qquad \text{II.25.4.}$$

As has been seen, the first moment about an arbitrary origin is the probable or expected value and in case of a sample it is the *arithmetic mean* of the x values.

The k th moment of the variable x about an arbitrary point a is defined as

$$\mathrm{E}\,[(x - a)^k] = \int_a^b (x - a)^k\, f(x)\, dx \qquad \text{II.25.5.}$$

or

$$\mathrm{E}\,[(x - a)^k] = \sum_{1}^{n}{}_{i}\,(x_1 - a)^k\, p_1. \qquad \text{II.25.6.}$$

If a is the arithmetic mean $\bar{x}$ of x and if μ_k is the symbol for the k th moment about the mean, then

$$\mu_k = \mathrm{E}\,[(x - \bar{x})^k] = \mathrm{E}\,[(x - \nu_1)^k] = \int_a^b (x - \nu_1)^k\, f(x)\, dx \qquad \text{II.25.7.}$$

or

$$\mu_k = \mathrm{E}\,[(x - \nu_1)^k] = \sum_{1}^{n}{}_{i} p_1\,(x_1 - \nu_1)^k. \qquad \text{II.25.8.}$$

It is not hard to see that $\mu_2 = \sigma^2$.

It is easy to show that the moments about the mean can be expressed in terms of the moments about an arbitrary origin. These relations are:

$$\mu_r = \sum_1^k{}_i\, p_1 (x_1 - \nu_1)^r = \int_a^b (x - \nu_1)^r f(x)\, dx \qquad \text{II.25.9.}$$

Specifically:

$$\begin{aligned}
\mu_0 &= 1 \\
\mu_1 &= 0 \\
\mu_2 &= \nu_2 - \nu_1^2 \\
\mu_3 &= \nu_3 - 3\,\nu_1\,\nu_2 + 2\,\nu_1^3 \\
\mu_4 &= \nu_4 - 4\,\nu_1\,\nu_3 + 6\,\nu_1^2\,\nu_2 - 3\,\nu_1^4
\end{aligned} \qquad \text{II.25.10.}$$

. . . . . .

$\mu_r = \sum_0^r{}_i \binom{r}{i} (-\nu_1)^i \nu_{r-1}$, where $\binom{r}{i} = \dfrac{r!}{i!\,(r-i)!}$, namely the number of combinations of r things taken i at a time.

For a sample

$$\nu_r = \sum_1^k{}_i\, f_i\, X_i^r / n. \qquad \text{II.25.11.}$$

and

$$\mu_r = \sum_1^k{}_i\, f_i\, (X_1 - \bar{X})^r / n. \qquad \text{II.25.12.}$$

Now consider the translation $x' - X - X_0$, and if ν_r' = the rth moment of x', then

$$\nu_r = \sum_1^k{}_i\, f_i\, (X_1 - X_0)^r / n = \frac{\sum_1^k{}_i\, f_i\, (x_1')^r}{n} = \nu_r' \qquad \text{II.25.13.}$$

and similarly if $x = \dfrac{X - X_0}{c}$ and ν_r'' is the rth moment of x

$$\nu_r = \frac{\sum_1^k{}_i\, f_i\, (cx)^r}{n} = \frac{c^r \sum_1^k{}_i\, f_i\, x^r}{n} = c^r \nu_r'' \qquad \text{II.25.14.}$$

Hence:

$$\mu_r = \sum_0^r{}_i \binom{r}{i} (-\nu_1')^i \nu_{r-1}' \qquad \text{II.25.15.}$$

and

$$\mu_r = c^r \sum_1^r \binom{r}{i} (-\nu_1'')^i \nu_{r-i}''. \qquad \text{II.25.16.}$$

To illustrate: Consider the distribution of Table II.1. and find the first four moments about the mean using II.25.10 and II.25.16.

Table II.8.

SPEED IN MILES PER HOUR OF FREE MOVING VEHICLES ON SEPTEMBER 16, 1939 IN OAKLAWN, ILLINOIS, ON U.S.H. 12 AND 20 AT A POINT ONE MILE EAST OF HARLEM AVENUE

S	f	s	fs	fs^2	fs^3	fs^4
70–74	0	6	0	0	0	0
65–69	0	5	0	0	0	0
60–64	2	4	8	32	128	512
55–59	15	3	45	135	405	1215
50–54	14	2	28	56	112	224
45–49	29	1	29	29	29	29
40–44	74	0	0	0	0	0
35–39	60	−1	−60	60	−60	60
30–34	63	−2	−126	252	−504	1008
25–29	29	−3	−87	261	−783	2349
20–24	6	−4	−24	96	−384	1536
15–19	8	−5	−40	200	−1000	5000
	300 = n		−227	1121	−2057	11933

From Table II.8.

$$\nu_0'' = 1$$

$$\nu_1'' = \frac{-227}{300} = -0.75667$$

$$\nu_2'' = \frac{1121}{300} = 3.73667$$

$$\nu_3'' = \frac{-2057}{300} = -6.85667$$

$$\nu_4'' = \frac{11933}{300} = 39.77667$$

Hence from II.25.10 and II.25.16., it is found that

$$\mu_0 = 1.$$
$$\mu_1 = 0.$$
$$\mu_2 = c^2 (\nu_2'' - \nu_1''^2) = 25\,(3.73667 - .57255) = 79.1$$
$$\mu_3 = c^3 (\nu_3'' - 3\,\nu_1''\,\nu_2'' + 2\,\nu_1''^3) = 125\,[-6.85667 - 3\,(-0.75667)\,(3.73667) + 2\,(-0.75667)^3] = 311.5$$
$$\mu_4 = c^4 (\nu_4'' - 4\,\nu_1''\,\nu_3'' + 6\,\nu_1''^2\,\nu_2'' - 3\,\nu_1''^4) = 625\,[39.77667 - 4\,(-0.75667)\,(-6.85667) + 6\,(0.75667)^2\,(3.73667) - 3\,(-0.95667)^4] = 18342.1$$

It is also useful to find

$$\beta_1^2 = \frac{\mu_3^2}{\mu_2^3} = \frac{97032.25}{494913.67} = 0.196 \qquad \text{II.25.17.}$$

and

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18342.1}{6256.81} = 2.93. \qquad \text{II.25.18}$$

β_1 is an index of skewness and is useful to compare the intensity of the departure from symmetry of a distribution with another distribution. If the distribution is symmetrical, β_1^2, has the value zero.

β_2 is an index of kurtosis (flatness) and is sometimes used to determine whether a given distribution is more flat or less flat than a corresponding "normal" distribution.

β_1^2 and β_2^2 are useful for determining which curve of a set of curves is indicated by the data as a useful law of behavior. The theory attached to these concepts was developed by the late Karl Pearson and will be discussed briefly in Chapter III.

II.26. *Relation Between Means.* For positive numbers,

$$x_1 < x_2 < \ldots < x_k,$$
$$x_1 < \text{H.M.} < \text{G.M.} < \text{A.M.} < \text{R.M.S.} < \text{C.H.M.} < x_n.$$

II.27. *Desirable Properties of An Average.*

(a) An average should be precisely defined.

(b) An average should be based on all observations.

(c) An average should possess some simple and obvious properties to render its general nature comprehensible: it should not be too abstract in mathematical characterization.
(d) An average should be possible of easy and rapid calculation.
(e) It should be as little affected as may be possible by *fluctuations of sampling* or by *sampling errors*.
(f) The measure chosen should lend itself to algebraic treatment and its basis should be concordant with the basis of the problems to be analyzed.

These properties applied to the mean, median, and mode, geometric mean, and harmonic mean are:

I. *Arithmetic Mean.* The A.M. satisfies a, b, c, d, e, f. The arithmetic mean has the following properties.

(a) The sum of the deviations from the mean, taken with their proper signs is zero.
(b) The mean of a whole series can be readily expressed in terms of the means of its components.
(c) The mean of all the sums or differences of corresponding observations in two series (of equal numbers of observations) is equal to the sum or difference of the means of the two series.
(d) The sum of squares of the deviations from the arithmetic mean is a minimum.

II. *Median.* The median satisfies (b) and (c) but the definition does not necessarily lead in all cases to a determinate result. The median is easier to compute than the arithmetic mean. The arithmetic mean is superior to median in lending itself to algebraic treatment. No theorem for median exists similar to (b) for mean and likewise to (c). The median has the following advantages over the mean:

(a) It is very readily calculated: a factor to which, however, as already stated, too much weight ought not to be attached.
(b) It is readily obtained without necessity of measuring all objects to be observed.
(c) Sum of the deviations from Median, all > 0, is a minimum.

III. *Mode.* What we want to arrive at is the mid-value of the interval for which the frequency would be a maximum, if the intervals

could be made indefinitely small and at the same time their number be so increased that the class frequency would run smoothly. A smoothing process is necessary; viz. that of fitting an ideal frequency curve of given equation to actual figures.

IV. *Geometric Mean.* The geometric mean is used in averaging rates or ratios rather than quantities.

(a) If the ratios of the geometric average to the measures it exceeds or equals be multiplied together, the product will be equal to the product of the ratios of the geometric average to those measures which exceed it in value.

$$\text{If } x_1 < x_2 < x_3 < \ldots < x_k < \text{G.M.} < x_{k+1} < x_{k+2} < \ldots < x_n,$$

$$\text{then, } \frac{G}{x_1} \cdot \frac{G}{x_2} \cdot \ldots \cdot \frac{G}{x_k} = \frac{x_{k+1}}{G} \cdot \frac{x_{k+2}}{G} \cdot \ldots \cdot \frac{x_n}{G}. \qquad \text{II.27.1.}$$

(b) The geometric average of the ratios of corresponding observations in two series is equal to the ratio of their geometric averages.

(c) The geometric average of the series formed by combining n different series each with the same frequency is the geometric average of the geometric averages of the separate series.

V. *Harmonic Mean.* The harmonic average of a set of measurements must be used in the averaging of time rates.

Having shown the initial procedure necessary for a statistical analysis, namely, how to summarize data and how to obtain summary numbers for the purpose of characterizing the law of behavior of the observed facts, we shall now develop the necessary theory that is basic for the analysis and solution of traffic problems.

REFERENCES, CHAPTER II

[1] Yule, G. Udney, and Kendall, M. C., "*An Introduction to the Theory of Statistics*," C. Griffin & Co., London, 1937.

[2] Croxton, F. E., and Cowden, D. J., "*Applied General Statistics*," Prentiss-Hall Inc., New York, 1946.

[3] Rider, Paul, "*Statistical Methods*," John Wiley & Sons Inc., New York, 1939.

[4] Kendall, M. C., "*The Advanced Theory of Statistics*," Charles Griffin & Co., London, 1946, Vol. I.

CHAPTER III

STANDARD DISTRIBUTIONS AND THEIR MATHEMATICAL PATTERNS

III. 1. *Objective.* The purpose of this chapter is to explain the related problems of first ascertaining the nature of a universe of events and second finding a mathematical model or pattern that fits the universe. From experience and intuition, we know that a sample will tell us something about the entire series of events, and that the larger the sample the more accurately it reflects the characteristics of the parent universe. We reason that a mathematical model of the sample, if the sample is large, will also be a model of the universe. Obviously, this fitting of mathematical patterns will be much easier if we know something about the types of universes or distributions of events we may expect to find.

There are three of these theoretical distributions that constitute the basic patterns. They are, in the order of their discovery, the *Binomial* (James Bernoulli about 1700), the *Normal* (Demoivre about 1700, Laplace and Gauss about 1800), and the *Poisson* (B.D. Poisson about 1837). Other distribution patterns have been discussed by Gram (1879), Fechner (1897), Thiele (1900), Edgeworth (1904), Charlier (1905), Brun (1906), Romanowsky (1924), and others. These are in general either other approaches to, modifications, or generalizations of the three basic distributions. The most logical order to present these from the standpoint of clearness is also the historical order of appearance. But before considering the first of these, the Binomial distribution, we shall discuss the elements that make up a distribution.

III.2. *The Elements of a Distribution.* In order to define and to point out the interrelationships of the elements that make up a distribution, let us consider a trial like the throwing of a die. The result will be the happening or non-happening of a specific event such as the falling of the die with one spot on the top face.

An event, of course, can be the occurrence of any attribute or

characteristic as well as a happening. In traffic, for example, it could be the age of a driver, his seeing ability, the life of an automobile tire, the weight class of a truck, the volume of traffic, the speed of a vehicle, or any one of many other things.

The happening of a specific thing is called the *Event E*, and the non-happening is called the complementary event $\bar{E}$. If the die is thrown a limited number of times (number of trials), we get a sample distribution of E's and $\bar{E}$'s. If the number of trials is increased without limit, the observed sample distribution approaches the true or theoretical distribution of the *universe* or *total population* of the events.

There are thus two kinds of distributions: (a) the theoretical and (b) the experimental or sample distribution.

The Theoretical Distribution: In order to explain the theoretical distribution, let f_t be the number of ways in which the event E can take place, f_c the number of ways for the complementary event $\bar{E}$, and n the total number of trials or happenings and non-happenings.

The probability that the event E will occur is the ratio of the number of ways f_t in which E can happen to the total number of possible *and equally likely* happenings and non-happenings. Let p or P (E) be this probability, then symbolically

$$p = P(E) = f_t/n \qquad \text{III.2.1.}$$

Similarly, the total number of ways f_c in which the event $\bar{E}$ can happen divided by n is defined as the probability (a-priori, true, or theoretical) that the event $\bar{E}$ will occur. Let q or P $(\bar{E})$ be this probability, then symbolically

$$q = P(\bar{E}) = f_c/n = \frac{n - f_t}{n} = 1 - \frac{f_t}{n}. \qquad \text{III.2.2.}$$

In the case of a die, if E is the event of the die's falling with one-spot on the top face and $\bar{E}$ is the event of the die's falling some other way, then $f_t = 1$, $f_c = 5$, $n = 6$
and

$$p = P(E) = \tfrac{1}{6}; \; q = P(\bar{E}) = \tfrac{5}{6}; \quad \text{and} \quad p + q = \tfrac{1}{6} + \tfrac{5}{6} = 1.$$

Again if n is the *total* number of registered vehicles and f_t is the number of light trucks, then

$$p = P(E) = \frac{f_t}{n}$$

is the true probability that a vehicle is a truck.

In general, let a be the number of times the event E occurs, and let b be the number of times the event $\bar{E}$ occurs, these being the only possibilities. Then $p = a/(a + b)$ is the probability that the event happens as specified – event E, and $q = b/(a + b)$ is the probability that the event does not occur – event $\bar{E}$. It follows that $p + q = 1$, which simply demonstrates what we know intuitively that an event is certain to happen or not to happen. This also shows that both p and q are positive numbers. This is the *Fundamental additive property* in probability. This property is also referred to in the literature as the Rule of Complementation.

Let us now suppose that one tosses a penny twice and wishes to find the probability of getting two heads. One might reason falsely that there are three possibilities: two heads, two tails, or one head and one tail. One of these outcomes is two heads, therefore, one might reason that the probability is $\frac{1}{3}$, but this reasoning is false, for the events are not *equally likely*. The third event may occur in two ways for a head could appear on the first trial and the tail on the second, or the head could appear on the second and the tail on the first. There are really four equally likely outcomes or phases: HH, HT, TH, TT; and the correct probability is therefore $\frac{1}{4}$. The four events are independent and mutually exclusive. If two heads are up, that is the only possible combination, for if a penny is heads up, it obviously cannot at the same time be tails up. This mutual exclusiveness does not always exist. Suppose that one wishes to compute the probability of drawing a king or a heart from a deck of cards. The chances might be assumed to be $\frac{1}{17}$ since there are 4 kings and 13 hearts. But this is incorrect, for the drawing of a king does not exclude drawing of a heart. The king may also be a heart.

The Experimental Distribution: The experimental or sample distribution is obtained from a number of observations of events.

Let f_0 be the number of times the event E is observed to happen and n the total number of trials or observations. The ratio f_0/n is called *the relative frequency* of the event E and $\left(1 - \frac{f_0}{n}\right)$ is the relative frequency of the event $\bar{E}$.

The obtaining of the numerical values of the relative frequencies f_0/n is actually a very simple problem since it is essentially a problem of counting. The value of f_0/n in contrast to the true probability varies with the number of observations or trials n. One might count all the traffic violations that occurred at an intersection during the passing of 5000 vehicles and find that there were no violations. In this situation, the observed $f_0 = 0$, $n = 5000$ and $f_0/n = 0/5000$ equals zero. But if the violations occurring during the passing of 25000 vehicles were counted, it might be found that there were 4 violations, and now the observed $f_0 = 4$, $n = 25000$, and $f_0/n = 4/25000$. Actually, we need to know the probable or expected value of such observed relative frequencies, f_0/n. This is defined as the *true probability* p that the event E will occur and it is the limit that f_0/n approaches as the number of trials (observations) is indefinitely increased. Expressed symbolically, if $E\,(f_0/n)$ is the symbol for the probable or expected value of an observed relative frequency f_0/n, then

$$E\left(\frac{f_0}{n}\right) = \underset{n \to \infty}{\text{Limit}} \left(\frac{f_0}{n}\right) = p = p\,(E) \qquad \text{III.2.3.}$$

It should be noted that in actual cases n need not be infinite to give a *practical result.* It is, however, necessary that n is *not small.*

The discussion just given may be summarized with two definitions:

Definition 1. If an event E can happen in f_t cases out of a total of n possible cases which are all considered by mutual agreement to be equally likely, then the probability $p = p\,(E)$ that the event E will occur is defined to be (f_t/n). Symbolically, $p = P\,(E) = f_t/n$.

Definition 2. If a series of many observations or trials is made, and if the ratio of the number of times, f_0, the event E occurs, to the total number of observations, n, namely, f_0/n, approaches nearer and nearer to a definite number, $p = P\,(E)$, as larger and

larger sets of trials or observations are made, then the probability of E is defined to be p. Expressed symbolically,

$$\lim_{n \to \infty} \left(\frac{f_0}{n} \right) = p = P(E)$$

An important question yet to be answered is: How much in error is f_0/n from p for a given number of observations and how certain are we that this error is not exceeded? In other words, for a given degree of certainty, how large a sample of observations must be made to guarantee that a specified error will not be exceeded?

This question is answered by the fundamental theorems of Bernoulli[1] and Cantelli[2] and by the Bienayme – Tchebycheff criterion,[3] which will be stated without proof.

III. 3. *Bernoulli's Theorem.*[1] Bernoulli found that there is a definite number of observations that will give a certain assurance that a given error will not be exceeded. His finding is based upon a natural law which may be demonstrated by the tossing of a penny. If the penny is not defective, the probability p of getting a head is $\frac{1}{2}$. Let us now assume 4 heads have been obtained in 10 tosses. This relative frequency (f_0/n) or $\frac{4}{10}$ is in error from the true or theoretical probability p of $\frac{1}{2}$ by 0.1. Let us next assume that we have tossed the penny 100 times and obtained 51 heads. The relative frequency $\frac{51}{100}$ is now in error by only 0.01. With more tosses there would be a tendency toward a further decrease in error which would lead us to suspect that something may be known about the number of trials that are necessary in order to get from observations a probability that will differ from the theoretical probability p by less than an arbitrarily assigned positive quantity ε, known as the experimental error.

The next question to be answered is how certain are we that the error will not be more than ε. The measure of our confidence that ε is the maximum error is indicated by attaching a probability to ε. This probability is dependent upon the number of trials n.

The probability η that ε is not the maximum error is the complement of the probability that ε is the maximum error. This

probability, η, is the measure of our *lack of confidence* that ε is not exceeded and is called the *level of significance.* If η is the level of significance, then $1 - \eta$ is the measure of our confidence or ability to prove that ε is not exceeded. The number, *Eta,* is also sometimes called the *risk.* In common parlance, if we are 75 per cent certain of our result, we are 25 per cent uncertain, or in other words, the risk is 25 per cent.

If we wished to find the size of sample necessary to give us a 99 per cent guarantee that the relative frequency (f_0/n) obtained would differ from the theoretical probability p for the universe by not more that 0.03, ε would be 0.03 and η would be 0.01. The value of 0.01 for η would mean that 1 per cent of the time it would be impossible to explain the difference between the observed and the theoretical frequency other than that it just happened. In other words, it would mean that the odds are 99 to 1 in favor of finding at least one real reason for the existence of the difference other than that it was merely accidental.

Having examined the underlying theory of Bernoulli's theorem, we will now state it more rigorously: *For any arbitrarily given* $\varepsilon > 0$ *and* $0 < \eta < 1$ *there exists a number of trials* n_0 *dependent upon both* ε *and* η $\{$*symbolically* $n_0(\varepsilon, \eta)\}$ *such that for any single value of* $n > n_0(\varepsilon, \eta)$, *the probability that the observed relative frequency,* (f_0/n) *of an event E in a series of n independent trials with constant probability p will differ from this probability p by less than* ε, *will be greater than* $1 - \eta$.

Symbolically, this is written

$$P\{|f_0/n - p| < \varepsilon\} > 1 - \eta \text{ for } n > n_0. \qquad \text{III.3.1.}$$

The $n > n_0$ in Bernoulli's theorem is given by the following inequality:

$$n > n_0 = \frac{1 + \varepsilon}{\varepsilon^2} \log_e \frac{1}{\eta} + \frac{1}{\varepsilon} \qquad \text{III.3.2.}$$

Example 1. Given $\varepsilon = 0.01$ and $\eta = 0.01$. Substituting these given values in the inequality III.3.2., we get

$$n > n_0 = \frac{1.01}{(.01)^2} \log_e \frac{1}{0.01} + \frac{1}{0.01}, \text{ whence } n > n_0 = 46613.$$

In this example, $n_0 = 46613$. However, n is *any single* number greater than 46613.

Example 2. Given $\varepsilon = 0.01$ and $\eta = 0.05$. Substituting these given values in the inequality III.3.2., we find that

$$n > n_0 = \frac{1.01}{(.01)^2} \log_e \frac{1}{0.05} + \frac{1}{0.01}, \text{ whence } n > n_0 = 30357.$$

Hence $n_0 = 30357$ and n is *any single* number greater than 30357. A comparison of the results of the two examples shows that reducing the certainty from 99 per cent to 95 per cent reduced the size of the sample required from 46614 to 30358.

Increasing the allowable experimental error will also decrease the size of the sample required.

Example 3. Given $\varepsilon = 0.05$ and $\eta = 0.05$. Substituting these given values in III.3.2., it is found that

$$n > n_0 = \frac{1.05}{(.05)^2} \log_e \frac{1}{0.05} + \frac{1}{0.05}, \text{ whence } n > n_0 = 1278.$$

Under the conditions, n is *any single* number greater than 1278.

The result of Example 3 means that if a random set of 1279 observations is taken, we are 95 per cent certain that the true probability p for the occurrence of the event E will be between the values $f_0/n - 0.05$ and $f_0/n + 0.05$. This may be expressed symbolically as

$$P\{ | f_0/n - p | < 0.05 \} > 0.95$$

for any single $n > 1278$. There are similar interpretations for examples 1 and 2.

An examination of Bernoulli's theorem shows that the number of observations necessary for a given result is totally independent of the true probability p and hence is independent of the theoretical distribution law. In other words, without knowing anything about the nature of the law of behavior, it is possible to determine the sample size for a specified accuracy and certainty. If, however, we have some knowledge of the law of behavior which is the case in nearly all practical applications, the size of the sample will be much smaller than indicated in Examples 1, 2, 3, — sometimes even less than 100. This will be made more apparent in later discussions.

For the sake of clarity, let us summarize the various aspects of Bernoulli's theorem. This theorem is based upon the law that as n increases, the measure of uncertainty η decreases. It enables us to find for a fixed error ε and measure of uncertainty η the size of *a single n.* This being the case, it is now possible to learn how large n must be so that the sum of all the decreasing measures of risk (the η's) for *all* N's larger than n, is less than a selected η and an assigned error ε. It follows, of course, that if the sum of the risks in question is less than η, then any one of them is less than η.

More precisely: Instead of there being any single $n > n_0$, for a given ε and η there is a number of trials, N, which is such that the sum of the risks for *all* n's $>$ N, is at most η. The number N is found by Cantelli's theorem.

III. 4. *Cantelli's Theorem.*[2] *For a given* $\varepsilon < 1$, $\eta < 1$, *let* $n > N(\varepsilon, \eta)$ *be an integer satisfying the inequality:*

$$n > \frac{2}{\varepsilon^2} \log_e \frac{2}{\eta} + 2. \qquad \text{III.4.1.}$$

With the value of n given by the inequality, the probability that the observed relative frequency (f_0/n) *of an event E will differ from the actual theoretical probability p by less than* ε *in the nth and all the following trials is greater than* $1 - \eta$.

Thus Cantelli's theorem, as noted above gives the probability for *all* n's $>$ N (ε, η), namely for $n = N, N + 1, N + 2, \ldots$, that $|f_0/n - p| < \varepsilon$. The complementary probability is the probability that at *least* one of the inequalities $|f_0/n - p| < \varepsilon$ is true where n may be equal to either N, or N + 1, or N + 2, ... Since these different possibilities form a set of mutually exclusive events it follows that the probability that at least one of the events has occurred is the sum of the probabilities that that one and *all* the following events have occurred.

Now, if Q $(Q \leq \eta)$ is the probability of this complementary event then it is the probability that the experimental error is at most ε in the nth and any or *all* of the following trials.

If we know or specify any two of the quantities n, ε, η, the third may be found in terms of Bernoulli's theorem (III.3.2.) or Cantelli's theorem (III.4.1.).

Since the probability that the experimental error is at most ε in *any single* number of trials greater than a given number n_0 is more restricted than the probability that the experimental error is at most ε, in *all* the number of trials greater than N, we would expect, as is the case, that more trials are necessary for the less restricted situation covered by the Cantelli theorem than are necessary for the Bernoulli theorem.

It is important to note that in both Cantelli's and Bernoulli's theorems, the number of trials necessary is independent of the probability p that the event will happen as specified and hence is independent of the distribution law. In other words, the results are true as long as we are sure that the event will happen or will not happen, or speaking mathematically, so long as it is true that $p + q = 1$ where q is the probability that the event will not happen as specified.

If the value of p is known which is the same as saying that we know the distribution law, and n is also dependent on p then, in general, the number of trials found from theorems III.3.2. and III.4.1. is much too large. This fact will be demonstrated later.

Example 1. Letting $\varepsilon = 0.01$ and $\eta = 0.01$ as in example 1 above and substituting these given values in the inequality III.4.1.,

$$n > \frac{2}{\varepsilon^2}\log_e\frac{2}{\eta} + 2 = \frac{2}{(0.01)^2}\log_e\frac{2}{0.01} + 2, \text{ whence}$$

$$n > 152{,}021.$$

In this example, $N = n + 1 = 152{,}022$. Therefore in the 152,022nd trial and *all* the following trials (and hence in at least one) we are assured that the observed relative frequency (f_0/n) will differ from the theoretical probability p by at most 0.01 and that it is $(1 - \eta) = 0.99$ equals 99 per cent certain that this is true and only 1 per cent uncertain that this is true.

Example 2. Let $\varepsilon = 0.01$ and $\eta = 0.05$, then III.4.1. becomes

$$n > \frac{2}{(0.01)^2}\log_e\frac{2}{0.05} + 2, \text{ whence}$$

$$n > 119{,}832.$$

Example 3. Let, as in example 3 above, $\varepsilon = 0.05$ and $\eta = 0.05$. In this case, III.4.1. becomes

$$n > \frac{2}{(0.05)^2} \log_e \frac{2}{0.05} + 2, \text{ whence } n > 4796.$$

The results of these examples when compared with the minimum number of trials necessary when using Bernoulli's theorem show that Cantelli's theorem requires more trials. This is because Cantelli's theorem gives a value for *all* n's greater than N while Bernoulli's theorem gives a value for *any single n* greater than n_0. In either case, as the number of trials is increased, the probability that the experimental error ε has a specified upper limit becomes greater and greater, and η becomes smaller and smaller.

The theorems of Bernoulli and Cantelli are based upon the idea that there is definite probability that the values of a stochastic variable will fall within a specified range.

Another approach is to find the probability that a stochastic value taken at random will differ from some chosen value a by as much as a specified amount, D. This probability is given by the *Bienaymé-Tchebycheff Criterion.*[3]

III. 5. *The Bienaymé-Tchebycheff Criterion.*[3] This criterion is independent of the form of distribution of given measurements and in addition is independent of the origin. If X is the stochastic variable which may assume the values X_i ($i = 1, 2, \ldots, n$), and if p_i ($i = 1, 2, \ldots, n$) are the corresponding probabilities, where $\Sigma p_i = 1$ and if a is any number (origin) from which the differences of the X's are measured, then

$$D^2 = E(X_i - a)^2 = \Sigma p_i x_i^2 \qquad \text{III.5.1.}$$

where $x_i = x_i - a$ and D^2 is the expected value of the squares of the differences of the X's from a.

Under these conditions, it is found that, if $\lambda > 1$,

$$P(\lambda D) \leq 1/\lambda^2 \qquad \text{III.5.2.}$$

This expression, wherein (λD) means λ times D and λ equals the multiple of the differences D from the chosen number a, is the *Bienaymé-Tchebycheff Criterion.*

The criterion, to state it in words, says that the probability P (λ D) is not more than $1/\lambda^2$ that a stochastic variable taken at random will differ from some chosen number a by as much as λ ($\lambda > 1$) times the value of D. A very useful special case is when a is the probable or expected value.

Example 1. If the probability P (λ D) $= \eta \leq .01$ and $\varepsilon = .01$, then for any a and p, λ must be $\sqrt{100} = 10$. It will be seen later that n must be greater than 250,000.

Example 2. If the probability P (λ D) $= \eta \leq .05$ and $\varepsilon = .01$, then for any a and p, λ must be $\sqrt{20}$. In this case $n > 50{,}000$.

Example 3. If the probability P (λ D) $= \eta \leq .05$ and $\varepsilon = .05$, then for any a and p, λ must be $\sqrt{20}$. In this case $n > 2000$.

These illustrations demonstrate that quite frequently the experimenter gathers more data than is necessary for the accuracy required. This makes the cost of the study unnecessarily large and demonstrates a lack of efficiency as well as an approach that is scientifically unsound.

If we have a limit definition of probability, Bernoulli's theorem is an immediate consequence thereof. In case we have any definition of probability p for the event E happening as specified, it is possible to prove Bernoulli's theorem by the use of the Bienaymé-Tchebycheff criterion. This will be shown later in this chapter.

In general, the evaluation of the probability of a given chance event necessitates the enumeration of all possible outcomes. These outcomes as shown by the tossing of a penny or the drawing of a card involve combinations and arrangements (permutations) of happenings.

III.6. *Permutations and Combinations.* There are two basic principles in combinations:

1. If an event A can occur in a total of a ways and an event B can occur in a total of b ways, then A and B can occur in $a + b$ ways, provided they cannot occur at the same time.
2. If an event A can occur in a total of a ways and an event B can occur in a total of b ways, then A and B can occur together in $a \cdot b$ ways.

These two principles can be generalized to take account of any

number of events. Three independent events A, B, or C can occur in a + b + c ways and three events A, B, and C can occur together in a · b · c ways.

These ideas may be illustrated by letting A represent the drawing of a heart from a deck of cards and B the drawing of a spade. Since there are 13 hearts, there are 13 ways of drawing a heart, and likewise for spades. The number of ways in which a heart *or* a spade can be drawn is 13 + 13 = 26. The second principle is also illustrated by the drawing of a heart and a spade together. There are 13 ·13 ways of doing this, for with any one of the 13 hearts we may put one of the 13 spades, and with any one of the 13 spades, we may put one of the 13 hearts and so on.

A more general illustration of the second principle is that of a room in which there are n seats and x individuals to be seated, and where $x < n$. We wish to know, in how may different ways (arrangements or permutuations) these x individuals may be seated in the room. To find out we may proceed as follows: Assume that all the x individuals are outside the room. The first one to come in has n choices. He seats himself. When a second individual comes in, he has (n — 1) choices, or one choice less than the first individual. For the third individual there are (n — 2) choices, or one less than for the second person. Hence, there are n (n — 1) (n — 2) choices (arrangements or permutations) for the first three. This illustration brings out the fact that permutations have to do with single items or groups of items treated as units and that the choice for each succeeding individual (item or group) is reduced by one.

If we continue until all the x individuals are seated and if ${}_nP_x$ is the number of choices, then

$$ {}_nP_x = n\,(n-1)\,(n-2)\,(n-3)\,\ldots\,(n-x+1) \qquad \text{III.6.1.} $$

This expression may be shortened by multiplying it by

$$ \frac{(n-x)\,(n-x-1)\,(n-x-2)\,\ldots.\,3.2.1}{(n-x)\,(n-x-1)\,(n-x-2)\,\ldots.\,3.2.1} = \frac{(n-x)!}{(n-x)!} $$

It then becomes

$$ {}_nP_x = \frac{n!}{(n-x)!} \qquad \text{III.6.2.} $$

In the case when $x = n$, III.6.1. becomes

$$ {}_nP_x = n\,(n-1)\,(n-2)\,(n-3)\ldots 3.2.1. = n! \qquad \text{III.6.3.} $$

and this is the number of permutations (arrangements) of n things taken n or all at a time.

Let us now turn to the question of how many different combinations of x things are possible if n things are available. A combination is an unarranged or unordered set of things, while a permutation is an arranged or ordered set of things.

Definition: The number of different unordered sets of x ($x < n$) things which can be selected from a set of n things is called the number of combinations of the n things taken x at a time; and is designated by the symbol ${}_nC_x$.

To find ${}_nC_x$ it is only necessary to keep in mind that we may have permutations of groups (or combinations) as well as of individuals. After all the different groups have been obtained, the individuals in each group may be arranged to give the total number of permutations.

The number ${}_nP_x$ is thus the number of ways we can make ${}_nC_x$ group choices followed by x! independent individual choices. That is

$$ {}_nP_x = {}_nC_x \cdot x! $$

hence

$$ {}_nC_x = \frac{{}_nP_x}{x!} = \frac{n!}{(n-x)!\,x!} \qquad \text{III.6.4.} $$

since from III.6.2. $\quad {}_nP_x = \dfrac{n!}{(n-x)!}$

Example: Let us find (a) the number of permutations and (b) the number of combinations of *15* things taken *3* at a time.

(a) From III.6.1., ${}_{15}P_3 = 15 \cdot 14 \cdot 13 = 2730$

(b) From III.6.4., ${}_{15}C_3 = (15!)/(3!)\,(12!) = 455$.

Until now we have dealt with the simple probability of whether a single event would happen or would not happen. But we are also interested in finding the probability that two or more events will occur together.

For an illustration of a compound event, we may toss two pennies. The number of ways in which two pennies may lie are:

HH, HT, TH, TT. The probability of two pennies falling heads up is thus $\frac{1}{4}$. Now we recall that the probability of one penny falling heads up is $\frac{1}{2}$ and that $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. This indicates that the probability of the compound event, two pennies falling heads up, is under certain conditions the product of the probabilities of the two separate events, each event being a penny falling heads up. This is precisely what the situation is if the separate events are *independent.*

If it is kept in mind that for every event there is a corresponding probability p, then the theorem of compound probability follows immediately from basic principle number two in article III.6.

III.7. *Theorem of Compound Probability. If the probability that an event will occur is* p_1 *and if after this event has occurred the probability that a second event will occur is* p_2 *then the probability that both events will occur in the order stated, is* $p_1 \cdot p_2$.

If the events are independent, as in the case of the pennies, it is not necessary that they happen in any definite order. The combination a "head and a tail" is the same as a "tail and a head".

Corollary: If the separate elementary events are independent, the probability of the compound event is the product of the probabilities of the separate events.

If there are x independent events and if p is the probability of the occurrence of each independent event, the probability that the event will occur x times in x trials is p^x. If in n trials q is the probability that the event does not occur, and if x ($x < n$) is the number of times the event occurs, then $n - x$ is the number of times the event does not occur. Clearly, if p^x is the probability that the event will occur x times as specified, q^{n-x} is the probability that it will not occur the remaining $(n - x)$ times. Hence the combined probability that in n trials a specific x of the n events will occur as specified is

$$p(x) = p^x \cdot q^{n-x} \qquad \text{III.7.1.}$$

This theorem applies to a set of events as well as to a single event for the probability for the occurrence of any specific set of x events is the same as the probability for any other set of x events.

Consequently, the probability of the event's occurring exactly x times without the restriction of its being a specific x is equal to the product of the probability for any specific x occurrences by the number of combinations of x sets there are in n events. This value has been shown to be (III.6.4.) equal to

$$ {}_nC_x = \frac{n!}{x!\,(n-x)!} $$

Hence, the probability P (x) of the event's occurring exactly x times in n trials is

$$ P(x) = \frac{n!}{x!\,(n-x)!}\, p^x q^{n-x} = {}_nC_x\, p^x\, q^{n-x} \qquad \text{III.7.2.} $$

where x may assume the values 0, 1, 2, ..., n. This is a fundamental law in probability, and if we let x take on all integral values from *0* to *n*, we obtain the respective probability for each of the possible and mutually exclusive events.

A more general theorem in which combinations are involved is known as the Binomial Theorem.

III. 8. *The Binomial Theorem* (*applied to probability*). The Binomial Theorem states that if the probability that an action will take place in a particular way is p, and the probability that it will not be so performed is q, then the probability that it will take place in exactly n, (n — 1), (n — 2), ... 3, 2, 1, 0 out of n trials is given by the successive terms of the binomial expansion:

$$ (p+q)^n = p^n + n \cdot p^{n-1} q + \frac{n\,(n-1)}{1 \cdot 2}\, p^{n-2} q^2 + \ldots\ldots \qquad \text{III.8.1.} $$

which is known as the Binomial Distribution.

It will be noted that the generating term is of the form ${}_nC_x\, p^r\, q^{n-r}$. For the purpose of illustration, let a coin be tossed 3 times. In this case $p = q = \frac{1}{2}$. The probabilities of getting 0, 1, 2, or 3 heads are:

$$ (\tfrac{1}{2})^3,\ 3\,(\tfrac{1}{2})^3,\ 3\,(\tfrac{1}{2})^3,\ (\tfrac{1}{2})^3 $$

and these are the successive terms of

$$ (p+q)^3 = p^3 + 3\,p^2 q + 3\,pq^2 + q^3 $$

Similarly the probabilities of getting 0, 1, 2, 3, or 4 heads are:

$$(\tfrac{1}{2})^4,\ 4\,(\tfrac{1}{2})^4,\ 6\,(\tfrac{1}{2})^4,\ 4\,(\tfrac{1}{2})^4,\ (\tfrac{1}{2})^4.$$

We might represent the possible results of tossing a penny four times graphically, as shown in Figure III.1.

FIGURE III.1

GRAPHICAL REPRESENTATION OF THE POSSIBLE RESULTS OF TOSSING A PENNY

The possibility of each number of heads is represented on the vertical ordinate. The width of each rectangle is equal to one unit $= \Delta x$. The area of each rectangle expressed in general terms is

$$ {}_nC_x\, p^x\, q^{n-x}\, \Delta x$$
$$= {}_nC_x\, p^x\, q^{n-x}$$

This means that the area of each rectangle equals the probability of getting the number of heads corresponding with the mid-point of its base. The entire area = the probability of getting 0, 1, 2, 3, or 4 heads $= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16} = 1$, so that the probability of getting a given number of heads is equal to

$$\frac{\text{Area of rectangle}}{\text{Area of whole figure}}$$

Expressed mathematically, the probability of getting any number of heads, x

$$P(x) = \frac{{}_nC_x\, p^x\, q^{n-x}}{\Sigma\, {}_nC_x\, p^x\, q^{n-x}} = {}_nC_x\, p^x\, q^{n-x} \qquad \text{III.8.2.}$$

since $$\Sigma_n C_x\, p^x\, q^{n-x} = 1.$$

In the example given $p = q = \frac{1}{2}$ with the result that the graph of the distribution is symmetrical. If p is not equal to q the distribution is not symmetrical but skewed. It is also clear that as n is increased, the area can be accurately represented by a smooth curve. It is only in the *long run* that the relative frequency with which an event happens as specified may be compared to probability. It is only when a man has large capital that he can play long enough to take advantage of the odds in his favor.

A quicker and more efficient way of obtaining the probabilities for an event happening as specified x times out of n trials is by the use of a *recursion formula*. As in III.8.2., let

$$P(x) = \frac{n!}{x!\ (n-x)!}\, p^x\, q^{n-x}$$

Then,

$$P(x+1) = \frac{n!}{(x+1)!\ (n-x-1)!}\, p^{x+1}\, q^{n-x-1} \qquad \text{III.8.3.}$$

Dividing III.8.3. by III.8.2., we get

$$\frac{P(x+1)}{P(x)} = \frac{(n-x)}{x+1} \cdot \frac{p}{q} \qquad \text{III.8.4.}$$

$$\text{whence, } P(x+1) = \frac{(n-x)}{x+1} \cdot \frac{p}{q} \cdot P(x) \qquad \text{III.8.5.}$$

To obtain the values shown in the tabular form, we proceed as follows: Let $x = 0$, then from III.8.2 it is found that $P(x) = P(0) = q^n$. Next, from III.8.5., we find that where $x = 0$,

$$P(1) = \frac{n}{1} \cdot \frac{p}{q} P(0)$$

$$= \frac{n}{1} \cdot \frac{p}{q} \cdot q^n = nq^{n-1}\, p.$$

Then, let $x = 1$ in III.8.5., and

$$P(2) = \frac{n-1}{2} \cdot \frac{p}{q} P(1)$$

$$= \frac{n-1}{2} \cdot \frac{p}{q} \cdot nq^{n-1} p$$

$$= \frac{n(n-1)}{2!} q^{n-2} p^2$$

Continuing in this way, all the probabilities of happenings may be obtained and they are shown in the following table for the different possibilities.

Table III.1

BINOMIAL DISTRIBUTION

Number of Happenings	*Probability of Happenings*
0	q^n
1	$nq^{n-1} p$
2	$\frac{n(n-1)}{2!} q^{n-2} p^2$
3	$\frac{n(n-1)(n-2)}{3!} q^{n-3} p^3$
.	.
.	.
.	.
x	$\frac{n!}{x!\,(n-x)!} q^{n-x} p^x$
.	.
.	.
.	.
n	p^n

Such a description of happenings is designated a probability distribution or a relative frequency distribution in the case of a sample. If each of the probabilities were multiplied by the number of individuals (number of cases or number of trials), we would have the corresponding theoretical (absolute) frequency distribution.

III. 9. *Modal Term of Binomial Distribution.* The Binomial distribution is analyzed by finding the modal term, the arithmetic mean, and the variance. To find the modal term we take the generating term,

$$P(x) = \frac{n!}{x!\,(n-x)!}\,p^x\,q^{n-x}$$

of the binomial distribution and find the value of x such that the xth term will be a maximum and hence be greater than or equal to either the $(x+1)$th term or the $(x-1)$th term. In other words, the ratio of the xth to the $(x+1)$th term or the $(x-1)$th term is equal to or greater than one. Thus

$$\frac{P(x)}{P(x+1)} = \frac{\dfrac{n!}{x!\,(n-x)!}\,p^x\,q^{n-x}}{\dfrac{n!}{(x+1)!\,(n-x-1)!}\,p^{x+1}\,q^{n-x-1}} \geq 1 \text{ and}$$

$$\frac{P(x)}{P(x-1)} = \frac{\dfrac{n!}{x!\,(n-x)!}\,p^x\,q^{n-x}}{\dfrac{n!}{(x-1)!\,(n-x+1)!}\,p^{x-1}\,q^{n-x+1}} \geq 1$$

Simplifying these two inequalities, we find, respectively, that

$$\frac{x+1}{n-x}\cdot\frac{q}{p} \geq 1 \text{ or } x \geq pn - q \quad \text{and}$$

$$\frac{n-x+1}{x}\cdot\frac{p}{q} \geq 1 \text{ or } x \leq pn + p$$

Now, if $\tilde{x}$ is the *modal* or *maximum* value of x,

$$pn - q \leq \tilde{x} \leq pn + p \qquad \text{III.9.1.}$$

Thus neglecting a proper fraction, pn is the most probable or modal value. If $pn - q$ and $pn + p$ are integers, then there exist two equal terms which are larger than all the others. This is the same as saying that if the chance of n events happening is $\frac{1}{3}$, then in 30 trials it is most likely to happen 10 times.

Examples: (a) What is the greatest number of times the event

will happen as specified when there are $n = 11$ trials and when $p = q = \frac{1}{2}$. From III.9.1., we find that $\tilde{x}$ is either 5 or 6.

(b) If $n = 12$ trials and $p = q = \frac{1}{2}$, $\tilde{x} = 6$.
(c) If $n = 15$ trials and $p = \frac{1}{6}$ and $q = \frac{5}{6}$, $\tilde{x} = 2$.
(d) If $n = 18$ trials and $p = \frac{1}{6}$ and $q = \frac{5}{6}$, $\tilde{x} = 3$.
(e) If $n = 23$ trials and $p = \frac{1}{6}$ and $q = \frac{5}{6}$, $\tilde{x} = 3$ or 4.

III. 10. *Arithmetic Mean of Binomial Distribution.* Let $\bar{x}$ be the arithmetic mean (*mathematical expectation – probable* or *expected* number of times the event will happen as specified in n trials under the law of repeated trials). By definition, the arithmetic $\bar{x}$ of x is

$$\bar{x} = \frac{\sum_0^n x \left(\frac{n!}{x!\,(n-x)!}\, p^x\, q^{n-x} \right)}{\sum_0^n x \frac{n!}{x!\,(n-x)!}\, p^x\, q^{n-x}} \qquad \text{III.10.1.}$$

But the denominator is the total probability which is equal to 1. Simplifying,

$$\bar{x} = 0 \cdot q^n + 1 \cdot n q^{n-1} p + 2 \cdot \frac{n\,(n-1)}{2!}\, q^{n-2}\, p^2 + \ldots$$

$$= np \left(q^{n-1} + (n-1)\, q^{n-2} p + \frac{(n-1)\,(n-2)}{2!}\, q^{n-3}\, p^2 + \ldots \right)$$

$$= np\,(q+p)^{n-1} = np\,(1) = np. \qquad \text{III.10.2.}$$

Illustrative Example 1: Given $p = \frac{1}{6}$ and $n = 18$, and $q = \frac{5}{6}$ required to find the mean $\bar{x}$.

Substituting in III.10.2,

$$\bar{x} = 18 \cdot \tfrac{1}{6} = 3.$$

The answer may be interpreted to mean that in the long run the event will happen one time in 6 trials and therefore in 18 trials we would *expect* the number of occurrences to be 3, while the actual number of occurrences in a single trial may be $x = 0, 1, 2, 3, \ldots, 18$.

Illustrative Example 2: Suppose that it has been ascertained from a traffic count that on the average 30 per cent of the vehicles turn

left, what is the probability that (a) a specific 3 out of 5 (say the first 3) vehicles will turn left, (b) any three (exactly 3), out of 5 vehicles will turn left.

(a) In the first case, III.7.1., $p(x) = p^x q^{n-x}$ becomes

$$p(3) = (.3)^3 (.7)^2 = .01323 \qquad \text{III.10.3.}$$

(b) In the second case, III.7.2., $P(x) = \dfrac{n!}{x!\,(n-x)!} p^x q^{n-x}$

becomes

$$P(3) = \frac{5!}{3!\,2!} (.3)^3 (.7)^2 = .1323 \qquad \text{III.10.4.}$$

The answer found in III.10.3. means that in the long run, 1323 times out of 100,000, a specific 3 (say the first 3) out of each group of 5 vehicles will turn left. The answer found in III.10.4. means that in the long run, 1323 times out of 10,000, any 3 out of each group of 5 vehicles will turn left.

III.11. *Variance of Binomial Distribution.* Another important measure is the arithmetic mean of the squares of the differences between the number of times the event will happen as specified and the expected number of times the event will happen as specified. Recall that in Chapter II in discussing frequency diagrams we spoke of this as being similar to the square of the radius of gyration. This quantity is called the *variance*. To obtain its value, if σ^2 is the symbol for variance, then

$$E(x - np)^2 = \sigma^2 = \sum_0^n{}_x \left(\frac{n!}{x!\,(n-x)!}\right) p^x q^{n-x} (x - np)^2 \qquad \text{III.11.1.}$$

But

$$E(x - np)^2 = E(x^2) - [E(x)]^2 \qquad \text{III.11.2.}$$

Since, we have already found the value of $E(x)$ to be np, it suffices to obtain the value of $E(x^2)$. By the definition of expected value,

$$E(x^2) = \sum_0^n{}_x x^2 \left(\frac{n!}{x!\,(n-x)!} p^x q^{n-x}\right)$$

$$= 0.q^n + 1.nq^{n-1}p + 4\frac{n(n-1)}{2!}q^{n-2}p^2$$

$$+ 9\frac{n(n-1)(n-2)}{3!}q^{n-3}p^3 + \ldots\ldots\ldots\ldots$$

$$= np\left[q^{-n1} + 2(n-1)q^{n-2}p + \frac{3(n-1)(n-2)}{2!}q^{n-3}p^2 + \ldots\right]$$

$$= np\left[(q+p)^{n-1} + (n-1)p\left\{q^{n-2} + (n-2)q^{n-3}p\right.\right.$$

$$\left.\left. + \frac{(n-2)(n-3)}{2!}q^{n-4}p^2 + \ldots\ldots\ldots\ldots\right\}\right]$$

$$= np\left[1 + (n-1)(p)(q+p)^{n-2}\right]$$

$$= np\left[1 + (n-1)p\right] = np + n^2p^2 - np^2 \qquad \text{III.11.3.}$$

Substituting the values from III.11.3. and III.10.2 in III.11.2., we find

$$\sigma^2 = E(x - np)^2 = E(x^2) - [E(x)]^2 \quad \text{becomes}$$

$$\sigma^2 = np + n^2p^2 - np^2 - n^2p^2$$

$$= np - np^2 = np(1-p) = npq \qquad \text{III.11.4.}$$

Illustrative example: Given $p = \frac{1}{6}$, $q = \frac{5}{6}$, and $n = 18$. From III.11.4. we find that $\sigma^2 = 18\left(\frac{1}{6}\right)\left(\frac{5}{6}\right) = 2.5$. This means that in 18 trials we would expect the number of occurrences to differ from 3 by 2.5. In other words, we would expect the actual number of occurrences to lie between $3 - 2.5 = 0.5$ and $3 + 2.5 = 5.5$, namely, between 1 and 5.

In the case of relative frequency or relative number of occurrences, if $(x/n - p)$ is the difference between the observed number of occurrences out of n and the probability p of occurrence, then it is not hard to show that

$$E(x/n - p)^2 = \frac{E(x - np)^2}{n^2} = \frac{\sigma^2}{n^2} = \frac{pq}{n}. \qquad \text{III.11.5.}$$

III. 12. *Size of Sample Required for Stability.* At this point it should be noted that we are thinking of the relative frequencies in many random samples, and that we are concerned about the degree of

stability or the degree of dispersion of such a series of relative frequencies. This is a fundamental problem in statistics. In the binomial distribution, sometimes called the *Bernoulli distribution*, we assume that the underlying probability remains constant from trial to trial and from sample to sample and that the drawings are mutually independent. This assumption is implied in so-called *simple sampling*.

Returning to Bernoulli's theorem, III.3.1., let $\varepsilon = \lambda\sqrt{\frac{pq}{n}}, (\lambda > 1)$.

In the Bienaymé-Tchebycheff inequality, III.5.2., let $D = \sqrt{pq/n}$. Then

$$P(\lambda D) \leq \frac{1}{\lambda^2} \text{ becomes } P(\varepsilon) \leq \frac{pq}{n\varepsilon^2} \qquad \text{III.12.1}$$

It may be seen from III.12.1. that as n tends to infinity, $\eta = P(\varepsilon)$ tends toward zero. This proves Bernoulli's theorem for any distribution law of probability by the use of Bienaymé-Tchebycheff criterion as was suggested in III.5.

In order to get a comparison of the results obtained by articles III.3., III.4., III.5., let $\varepsilon = 0.01$, $p = 0.1$, $q = 0.9$, $\lambda = 2\sqrt{5} = 4.472$ and $\eta = 0.05$. Substituting these values in III.12.1.,

$$\eta = P(\varepsilon) \leq \frac{pq}{n\varepsilon^2}$$

$$= P(.01) = 0.05 \leq \frac{(.1)(.9)}{n(.01)^2}$$

whence $$n \geq 18{,}000.$$

Again let $\varepsilon = 0.05$, $p = 0.1$, $q = 0.9$, $\lambda = 2\sqrt{5} = 4.472$ and $\eta = 0.05$. Substituting these values in III.12.1., we get

$$\eta = P(\varepsilon) \leq \frac{pq}{n\varepsilon^2}$$

$$= P(.05) = 0.05 \leq \frac{(.1)(.9)}{n(.05)^2}$$

whence $$n \geq 718.$$

Comparing these results with those previously found, it is seen that they are materially less as was indicated previously. It is

noted that n is a maximum when $p = q = \frac{1}{2}$ for then pq is the maximum. Hence, it is always safe to take the value of n when p and q equal $\frac{1}{2}$ as the minimum value of n. That is, in case the values of p and q are not known, it is safe to use $p = q = \frac{1}{2}$ in determining the size of sample required. In many traffic problems, p is very small and q very near unity which will require a smaller sample for stability than if p were equal or nearly equal to q.

Additional means of characterizing the binomial distribution are moments about the mean. These are:

$$\begin{aligned}
&\mu_1 = 0 \\
&\mu_2 = npq \\
&\mu_3 = npq\,(q - p) \\
&\mu_4 = 3\,p^2 q^2 n^2 - pqn\,(1 - 6\,pq) \qquad \text{III.12.2.} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\mu_x = \sum_0^n (j - np)^x \binom{n}{j} q^{n-j} p^j \\
&\mu_{x+1} = pq\left(nx\mu_{x-1} + \frac{d\,\mu_x}{dp}\right)
\end{aligned}$$

where $\binom{n}{j}$ is the number of combinations of n things taken j at a time and n is very large.

Other characterizing means are the β coefficients:

$$\begin{aligned}
\beta_1 &= \frac{(q-p)^2}{npq} \\
\beta_2 &= 3 + \frac{1 - 6\,pq}{npq} \qquad \text{III.12.3.}
\end{aligned}$$

β_1 is a coefficient of skewness, while β_2 is a coefficient of kurtosis or "peakedness".

The theorems of Bernoulli and Cantelli and the Bienaymé-Tchebycheff criterion are devoted to obtaining a *lower limit* to the probability that the experimental error will not exceed a given amount.

The *binomial* distribution and particularly its generating function P (x) given in III.7.2. gives the actual probability of the

event's occurring exactly x times in n trials, so that it is possible to determine the actual probability of the event's occurring between any two specified number of times in n trials. This is accomplished by adding the respective separate probabilities involved since the events are mutually exclusive.

The function P (x) is given by

$$P(x) = \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

The function P (x) is a fundamental law of probability for all positive values of x, integral or fractional. The function is continuous almost everywhere (i. e. except for negative integers) and has a unique value for every positive value of x. It is simple enough to handle if x is an integer. It is quite difficult and cumbersome if x is not a positive integer.

In practice it is most usable when x is a whole number. Many times, however, x is not a whole number. It then becomes imperative, if possible, to derive from the function given in III.7.2. another continuous function which is easier to use and also gives us the actual probabilities (not lower limits only) that are desired to be known.

Two such functions are the *Normal Distribution* and the *Poisson Distribution*. We shall now develop and discuss these two functions.

III. 13. *The Normal Distribution.* The normal distribution is a continuous approximation to the binomial distribution when n is large and p and q are not small.

Let us reexamine the generating term P (x) of the binomial distribution, namely,

$$P(x) = \frac{n!}{x!\,(n-x)!} p^x q^{n-x} \qquad \text{III.13.1.}$$

The graph of this equation is a set of points whose abscissas are x values and ordinates are the corresponding P (x) values for all values of x from zero to plus infinity. The function P (x) is continous almost everywhere (i. e., except for negative integers).

For our purpose, it is convenient to translate the origin to the mean or expected value of x. This requires that we substitute $x = x' + np$ for x in III.13.1. It then becomes

$$P(x') = \frac{n!}{(x' + np)!\,(nq - x')!}\, p^{pn+x'} q^{qn+x'} \qquad \text{III.13.2.}$$

If we consider unit intervals only, this probability that the number of occurrences will lie between $np - k$ and $np + k$, inclusive of end values, is

$$\sum_{-k}^{k} {}_{x'} P(x') = P(-k) + P(-k+1) + \ldots + P(0) + P(1) + \ldots + P(k) \qquad \text{III.13.3}$$

This follows from the fact that the resultant event is obtained by compounding a set of mutually exclusive events in which case the resultant probability is the sum of the probabilities of the set of mutually exclusive events.

To simplify III.13.2., if the number of trials n is *large*, it is convenient to use Stirling's asymptotic approximation for n! which is

$$n! = n^n e^{-n} (2n)^{\frac{1}{2}} \left(1 + \tfrac{1}{12n} + \tfrac{1}{288n^2} + \ldots\right) \qquad \text{III.13.4.}$$

or

$$n! \cong \sqrt{2\pi}\, e^{-n} n^{n+\frac{1}{2}} \qquad \text{III.13.5.}$$

if the first term of III.13.4. only is used. If III.13.5. is used, the result obtained is equal to the true value divided by a number having a value between 1 and $\frac{1}{10n}$.

Remembering that n is large and using III.13.5. for all the factorials in III.13.2.,

$$P(x') = \frac{1}{(2\pi npq)^{\frac{1}{2}}} \left(1 + \frac{x'}{pn}\right)^{-pn-x'-\frac{1}{2}} \left(1 - \frac{x'}{qn}\right)^{-qn+x'-\frac{1}{2}} \qquad \text{III.13.6.}$$

Transforming III.13.6. by taking logarithms of both sides of the equality,

$$\log_e P(x') = -\log_e (2\pi npq)^{\frac{1}{2}} - (np + x' + \tfrac{1}{2}) \log_e \left(1 + \frac{x'}{pn}\right) - (qn - x + \tfrac{1}{2}) \log_e \left(1 - \frac{x'}{qn}\right) \qquad \text{III.13.7.}$$

Expanding $\log_e\left(1+\frac{x'}{pn}\right)$ and $\log_e\left(1-\frac{x'}{qn}\right)$ in power series of x',

III.13.7. becomes

$$\log_e[P(x')]\,[2\pi npq]^{\frac{1}{2}} = -(np+x'+\tfrac{1}{2})\left[\frac{x'}{np}-\frac{x'^2}{2n^2p^2}-\frac{x'^3}{n^3}R(x')\right]$$
$$-(qn-x'+\tfrac{1}{2})\left[\frac{x'}{nq}-\frac{x'^2}{2n^2q^2}-\frac{x'^3}{n^3}S(x')\right] \qquad \text{III.13.8.}$$

To make this expansion valid, it is necessary to assume that n is sufficiently large so that $\frac{x'}{n}$ is sufficiently small. It follows that R (x') and S (x') are finite.

Simplifying III.13.8., and performing the multiplying operations indicated, we find that

$$\log_e[P(x')]\,[2\pi npq]^{\frac{1}{2}} = \frac{(p-q)\,x'}{2npq}-\frac{x'^2}{2npq}+\frac{x'^2}{n^2}T(x') \qquad \text{III.13.9.}$$

The equation III.13.9. may be written in the form

$$\log_e[P(x')]\,[2\pi npq]^{\frac{1}{2}} = -\frac{x'^2}{2npq}-\frac{x'}{n}U(x') \qquad \text{III.13.10.}$$

where U (x') is also finite.

Now if n is large enough (in other words, n must be very large) so that $\left(\frac{x'}{n}\right)U(x')$ is very small (negligible or within the allowable error), then ignoring this term, III.13.10. may be written as

$$P(x') = \frac{1}{(2\pi npq)^{\frac{1}{2}}}e^{\frac{-x'^2}{2npq}} \qquad \text{III.13.11.}$$

which is called the *normal distribution.*

It appears that this was first known to DeMoivre in November, 1732. Multiply both sides of the equality III.13.3. by $\Delta x'$, then,

$$\sum_{-k}^{k}{}_{x'} P(x')\Delta x' = P(-k)\Delta x' + P(-k+1)\Delta x' + \ldots\ldots + P(0)\Delta x'$$
$$+ P(1)\Delta x' + P(k)\Delta x'$$

and on the assumption that P (x') is continuous,

$$\lim_{\Delta x'\to 0}\sum_{-k}^{k}{}_{x'} P(x')\Delta x' = \frac{1}{(2\pi npq)^{\frac{1}{2}}}\int_{-k}^{k} e^{-\frac{x'^2}{2npq}}dx' \qquad \text{III.13.12.}$$

The right hand member of III.13.12 is known as the *probability integral.* It gives the probability that a random variable x′ has the value $-k \leq x' \leq k$.

If P (x′) is discontinuous and the ordinates are at unit intervals, then in III.13.3. there is one more ordinate than intervals of area. Hence,

$$\sum_{-k}^{k}{}_{x'} P(x') = \frac{1}{(2\pi npq)^{\frac{1}{2}}} \int_{-k-\frac{1}{2}}^{k+\frac{1}{2}} e^{\frac{-x'^2}{2npq}} dx' \quad \text{approximately.} \qquad \text{III.13.13.}$$

The above results summarized lead to the well-known DeMoivre-Laplace theorem, namely[8]:

The probability that the difference $x' = x - np$ *between the number of occurrences x and the expected number of occurrences will not exceed a positive number k is given to a first approximation by III.13.12 and to closer approximation by III.13.13.*

III. 14. *Interpretation of the Properties of Normal Distribution.* The special form of the normal distribution as given in III.13.11. is restricted to the conditions that n is large and p and q are not small thus giving a continuous approximation to the binomial distribution.

$$\text{Now consider } P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \qquad \text{III.14.1.}$$

where σ is the standard deviation with the restriction that it is finite such that $0 \leq \sigma \leq k$.

The graph of the equation is shown in Figure III.2.

From III.14.1., it is seen that the curve is symmetrical with respect to the y-axis. Likewise the curve has a maximum point at $x = 0$, namely at the point whose abscissa is the arithmetic mean. There are two points of inflection, namely P_1 and P_2 each of which are at a distance σ from the arithmetic mean. The curve is asymptotic to the x-axis at both plus and minus infinity.

From III.14.1. or from tables, it is found that the total area under the curve is unity, the area between $x = -\sigma$ and $x = +\sigma$

is 0.6827, the area between $x = -2\sigma$ and $x = +2\sigma$ is 0.9545, and the area between $x = -3\sigma$ and $x = +3\sigma$ is 0.9973. If

$$\frac{2}{\sigma\sqrt{2\pi}}\int_0^{x} e^{-\frac{x^2}{2\sigma^2}}\,dx = \frac{1}{2}$$

then $$x = 0.67449\,\sigma \qquad \text{III.14.2.}$$

which is known as the *probable error.*

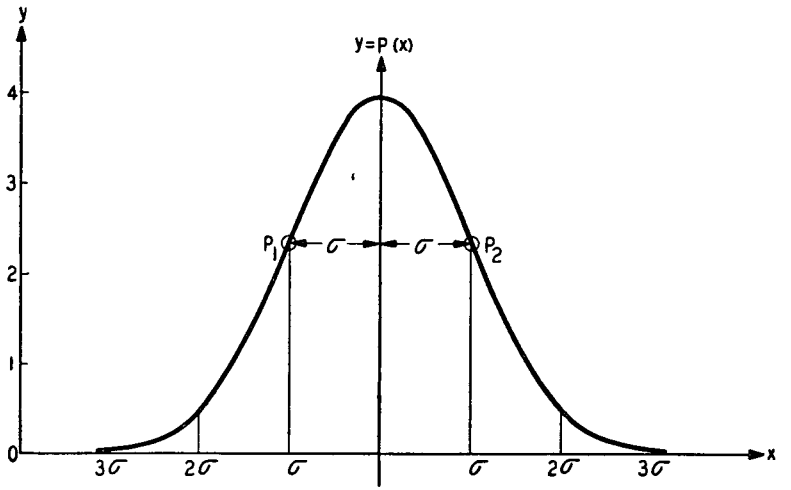


FIGURE III.2

GRAPH OF THE EQUATION $P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$

As an illustration, consider again the case $\eta = 0.05$, $\varepsilon = 0.01$. From the Bienaymé-Tchebycheff inequality, $\lambda = t = 4.472$. Now, let $p = q = \frac{1}{2}$. Then, from III.11.5. and III.12.1.,

$$t\sqrt{\frac{pq}{n}} \leq \varepsilon$$

becomes $$4.472\sqrt{\frac{(\frac{1}{2})(\frac{1}{2})}{n}} \leq 0.01$$

whence $n \geq 500$

Similarly, if $\eta = 0.05$ and $\varepsilon = 0.05$

$$t\sqrt{\frac{pq}{n}} \leq \varepsilon$$

becomes $$4.472\sqrt{\frac{(\frac{1}{2})(\frac{1}{2})}{n}} \leq 0.05$$

whence $n \geq 100$.

Again, let $p = \frac{1}{2}$ and $\varepsilon = 0.01$. The value of t such that

$$\frac{2}{\sqrt{2\pi}}\int_{-t}^{t} e^{-\frac{x^2}{2}}\,dx = 0.99 = 1 - \eta$$

is 2.58. But $n \geq \frac{pqt^2}{2}$. Hence, solving for n, it is found that $n \geq 166$ and if

$$\frac{2}{\sqrt{2\pi}}\int_{-t}^{t} e^{-\frac{x^2}{2}}\,dx = 0.95 = 1 - \eta,$$

$t = 1.96$ and $n \geq 97$, if $\varepsilon = 0.01$.

Under certain conditions where $p = q$, the equation of the continuous approximation curve is given by

$$y = \frac{Np^{p+1}}{ae^p\,\Gamma(p+1)}\,e^{-\gamma x}\left(1 + \frac{x}{a}\right)^{\gamma a} \qquad \text{III.14.3.}$$

where the origin is at the mode.

The question is often raised: How is it known that the distribution is normal? A very good answer is: If it can be justified axiomatically that the arithmetic mean is the most probable value, then the distribution is normal. This is known as the postulate of the arithmetic mean. Another way is: If $\beta_1 = 0$ and $\beta_2 = 3$ (See II.25.17. and II.25.18.), the distribution is normal.

III. 15. *Poisson Distribution.* This distribution is frequently thought of as the law of small probabilities or the law of rare events. It appears to be especially useful in solving many traffic problems (see Chap. V).

Consider again the generating term of the binomial expansion,

$$P(x) = \frac{n!}{x!\,(n-x)!}\, p^x q^{n-x} \qquad \text{III.15.1.}$$

the probability that in n trials exactly x of them will take place as specified, where p is the probability that the event in a single trial will occur as specified.

Equation III.15.1. may be written as

$$P(x) = \frac{n\,(n-1)\,(n-2)\ldots\ldots\ldots(n-x+1)}{x!}\, p^x\,(1-p)^{n-x} \qquad \text{III.15.2.}$$

Write $p = \frac{m}{n}$ where m is the number of times a given happening occurs in n trials. Substituting this value of p for p in III.15.2.,

$$P(x) = \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\ldots\left(\frac{n-x+1}{n}\right)\left(\frac{m^x}{x!}\right)\left(1-\frac{m}{n}\right)^n\left(1-\frac{m}{n}\right)^{-x} \qquad \text{III.15.3.}$$

Now, hold both x and m fixed and let n approach infinity. Then, in the limit,

$$\frac{n}{n} = 1,\ \frac{n-1}{n} = 1,\ldots\ldots,\ \frac{n-x+1}{n} = 1,\ \text{and}\ \left(1-\frac{m}{n}\right)^{-x} = 1.$$

To obtain the limiting value of $\left(1-\frac{m}{n}\right)^n$ we set

$$\left(1-\frac{m}{n}\right)^n = \left[\left(1-\frac{m}{n}\right)^{\frac{n}{m}}\right]^m. \qquad \text{III.15.4.}$$

The limiting value of $\left[\left(1-\frac{m}{n}\right)^{\frac{n}{m}}\right]$ as n approaches infinity is e^{-1}. Hence

$$\lim_{n\to\infty}\left[\left(1-\frac{m}{n}\right)^{\frac{n}{m}}\right]^m = e^{-m}. \qquad \text{III.15.5.}$$

Substituting all the limiting values just found in III.15.2., we obtain

$$P(x) = (1)\,(1)\,(1)\ldots\ldots(1)\,\frac{m^x}{x!}\,e^{-m}.\,(1) \qquad \text{III.15.6.}$$

which may be written as

$$P(x) = \frac{m^x e^{-m}}{x!} \qquad \text{III.15.7.}$$

which is *Poisson's distribution* or the *Poisson Exponential Function*. This function is a continuous approximation to the binomial distribution when p is small and n is *large*.

The function is continuous almost everywhere and has a real value for all values of x except negative integers. For negative integral values of x, P (x) is not defined. The continuity is obvious if it is recalled that x! is related to the *Gamma Function*,[9] that is:

$$x! = \int_0^\infty y^x e^{-y}\, dy = \Gamma(x + 1) \qquad \text{III.15.8.}$$

The graph of the function is shown in Figure III.3. Also tables (Tables for Biometricians and Statisticians, pp. 122–124) of values for P_x exist.

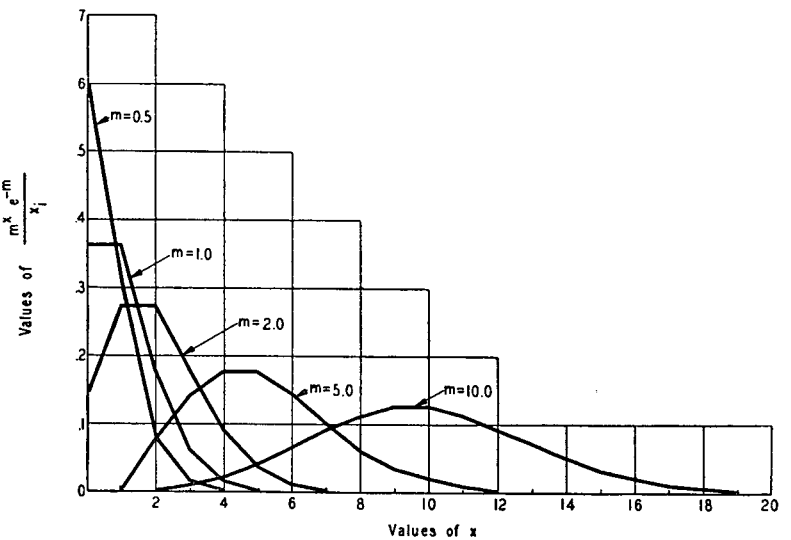


FIGURE III.3

GRAPH OF THE FUNCTION $P(x) = \frac{m^x e^{-m}}{x!}$

From the figure it is seen that for small values of m the curve is highly skewed and that as the values of m increase the curve becomes more symmetrical.

In all cases, p must be small and n must be large, but small values of m as well as large values of m are possible under these conditions. It is also quite important to note that as m becomes larger, the agreement between III.15.7. and III.13.11. becomes closer.

III.16. *The Sum of the Terms of the Poisson Distribution.* Since each term is the probability for the event's happening x times, the sum of the probabilities for each of these possibilities should equal unity because some one of the possibilities is certain to take place. Letting x take successively the values 0, 1, 2, ..., the sum of the respective terms is

$$\sum_0^\infty{}_x \frac{m^x e^{-m}}{x!} = \frac{m^0 e^{-m}}{0!} + \frac{me^{-m}}{1!} + \frac{m^2e^{-m}}{2!} + \cdots\cdots$$

$$= e^{-m}\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots\cdots\right) \qquad \text{III.16.1.}$$

The series in parentheses has the value e^m. Hence

$$\sum_x^\infty \frac{m^x e^{-m}}{x!} = e^{-m}e^m = e^0 = 1 \qquad \text{III.16.2.}$$

III.17. *The Arithmetic Mean of Poisson Distribution.* If $\bar{x}$ is the arithmetic mean number of happenings, then

$$\bar{x} = \sum_0^\infty{}_x \frac{m^x e^{-m}}{x!} x$$

$$= \frac{m^0 e^{-m}}{0!} 0 + \frac{me^{-m}}{1!} 1 + \frac{m^2 e^{-m}}{2!} 2 + \frac{m^3 e^{-m}}{3!} 3 + \cdots.$$

$$= me^{-m}\left[1 + \frac{m}{1!} + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots\cdots\right]$$

$$= me^{-m}e^m = m. \qquad \text{III.17.1.}$$

III.18. *The Variance of Poisson Distribution.* Since variance is the expected value of the squares of the measurements minus the square of the expected value of the measurements, we will first obtain the expected value of the squares of the measurements. It is given by,

$$E(x^2) = \sum_{0}^{\infty}{}_{x} \frac{m^x e^{-m}}{x!} x^2$$

$$= \frac{m^0 e^{-m}}{0!} 0 + \frac{m e^{-m}}{1!} 1 + \frac{m^2 e^{-m}}{2!} 4 + \frac{m^3 e^{-m}}{3!} 9 + \cdots\cdots$$

$$= m e^{-m} \left[1 + \frac{2m}{1!} + \frac{3m^2}{2!} + \cdots\cdots\right]$$

$$= m e^{-m} \left[e^m + \left(m + \frac{m^2}{1!} + \frac{m^3}{2!} + \cdots\cdots\right)\right]$$

$$= m e^{-m} \left[e^m + m\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \cdots\cdots\right)\right]$$

$$= m e^{-m} \left[e^m + m e^m\right] = m + m^2 \qquad \text{III.18.1.}$$

But the square of the expected value is m^2. Hence

$$\sigma^2 = E(x^2) - [E(x)]^2$$
$$= m + m^2 - m^2 = m \qquad \text{III.18.2.}$$

Example 1. There occurred at a certain highway intersection 6 accidents during the passing of 10,000 vehicles. In this case $p = 0.0006$ and $n = 10000$. Suppose we wish to know the probability that the number of accidents lies between 3 and 9 per 10000 vehicles. Making use of III.13.13., we find that

$$P(x) = \frac{1}{(2\pi npq)^{\frac{1}{2}}} \int_{-k-\frac{1}{2}}^{k+\frac{1}{2}} e^{\frac{-x'^2}{11.9928}} dx' = 0.02654 \int_{-3\frac{1}{2}}^{3\frac{1}{2}} e^{\frac{-x'^2}{11.9928}} dx'$$

From tables of the normal probability function it is found that if

$$z = \frac{x'_1}{\sigma} = \frac{3.5}{2.449} = 1.429$$

then

$$0.02654 \int_{-3\frac{1}{2}}^{3\frac{1}{2}} e^{\frac{-x'^2}{11.9928}} dx' = 0.847$$

the desired probability.

To calculate the probability from the Poisson distribution with $m = 6$, we add the probabilities for the event's happening 3, 4, 5, 6, 7, 8, and 9 times as taken from the Poisson tables[10] for individual terms:

Happenings	Probability
3	.089235
4	.133853
5	.160623
6	.160623
7	.137677
8	.103258
9	.068838
Total Probability	.854107

We may also use the table for cumulated terms and substract the probability for 10 or more happenings from the probability for 3 or more happenings with $m = 6$.

Happenings	Probability
3 or more	.938031
10 or more	.083924
	.854107 = probability of 3 to 9.

Again if the binomial distribution is used, the value of the desired probability is 0.854.

These results show that there is little difference between the use of the so-called normal distribution and the Poisson exponential function, while the Poisson exponential function is a better approximation than the Bernoulli distribution for rare events, that is events with small probability.

Example 2. For a given period of time, at a certain point on a highway, it is observed that on the *average* three heavy trucks per

100 vehicles pass the point. A subsequent sample contains six heavy trucks per 100 vehicles. Using the Poisson exponential distribution, compute the probabilities of 0, 1, 2, 3, 4, 5, 6, 7, and 8 heavy trucks per 100 vehicles using $m = np = 3$.

The probability distribution is shown in Table III.2.

Table III.2.

x	p_x	x	P_x
0	.0498	5	.1008
1	.1494	6	.0504
2	.2240	7	.0216
3	.2240	8	.0081
4	.1680		

This table shows that (1) the probability of obtaining one heavy truck in a sample of 100 vehicles is 0.1494; (2) the probability of getting more than three heavy trucks is .5768; (3) the probability of getting at least six heavy trucks is .3080.

The probability of six or less than six, being .9664 with a level of significance of $1 - .9664 = .0336$, indicates that on a 5 per cent level we have grounds to reject the hypothesis that this number of heavy trucks is not significant.

In obtaining the size of the sample so that the error from the arithmetic mean is one heavy truck, namely, that the number of heavy trucks is between 2 and 4, the reasoning is:

The standard deviation is

$$\sigma = m = np = n\,(.03)$$

and since $\varepsilon = 1$, it is clear that

$$\varepsilon = tm$$

becomes $1 = (1/3)\,n\,(.03)$

which gives $n = 100$

and the sum of the probabilities, namely

$.2240 + .2240 + .1680 = .6160$, the measure of certainty.

Example 3. Required to find the probability of n cars appearing within an interval of time r beginning at the instant t. Then

p (n, r, t), the probability of n cars within an interval of time r beginning at the instant t, is given by

$$p(n, r, t) = \frac{K^n e^{-K}}{n!}$$

where K is the expected number of cars in the interval.

III. 19. *Dispersion and Variance.* Thus far it has been assumed that the relative frequency (sample) or the probability (universe) that an event will happen as specified remains constant throughout the entire field of observation. There are many cases where the underlying probability (relative frequency) does not remain constant. This indicates that it is necessary that the statistician obtain all the available knowledge from the data by properly classifying them into subsets for analysis and comparison. In other words, it is valuable to know whether the relative frequencies or probabilities vary from case to case or from set to set.

Consider the following: Given N independent quantities $X_1, X_2, \ldots, X_N$ such that the mean or expected value $E(X_1)$ of X_1 is a_1 and the mean or expected value $E(X_1^2)$ of X_1^2 is A_1. Then, if $\bar{X} = \left(\frac{X_1 + X_2 + \ldots + X_N}{N}\right)$ and $a = (a_1 + a_2 + \ldots + a_n)/N$, it has been shown ("Probability," by J. L. Coolidge, Oxford Press, 1925, p. 67) that

$$E\left[\sum_1^N (X_1 - \bar{X})^2\right] = \frac{N-1}{N}\sum_1^N (A_1 - a_1^2) + \sum_1^N (a_1 - a)^2 \qquad \text{III.19.1.}$$

If the observations are from homogeneous data, $a_1 = a$, $A_1 = A$. In such a case, III.19.1., reduces to

$$E\left[\sum_1^N (X_1 - \bar{X})^2\right] = \frac{N-1}{N} \cdot N(A - a^2) = (N-1)\sigma^2 \qquad \text{III.19.2.}$$

since

$$\sigma^2 = E(X^2) - [E(X)]^2 = A - a^2.$$

The relationship given in III.19.2. reduces to

$$\sigma^2 = E\left[\sum_1^N (X_1 - \bar{X})^2/(N-1)\right] \qquad \text{III.19.3.}$$

Suppose now that a set $N = lk$ independent items has been observed and classified in some relevant manner, say, in l rows of k items each as shown in Table III.3.

Table III.3.

$X_{11}, X_{12}, \ldots, X_{1j}, \ldots, X_{1k}$	$T_1. \quad \bar{X}_1.$
$X_{21}, X_{22}, \ldots, X_{2j}, \ldots, X_{2k}$	$T_2. \quad \bar{X}_2.$
.....	.....
$X_{l1}, X_{l2}, \ldots, X_{lj}, \ldots, X_{lk}$	$T_l. \quad \bar{X}_l.$
$X_{l1}, X_{l2}, \ldots, X_{lj}, \ldots, X_{lk}$	$T_l. \quad \bar{X}_l.$
$T_{\cdot 1}, T_{\cdot 2}, \ldots, T_{\cdot j}, \ldots, T_{\cdot k}$	T
$\bar{X}_{\cdot 1}, \bar{X}_{\cdot 2}, \ldots, \bar{X}_{\cdot j}, \ldots, \bar{X}_{\cdot k}$	$\bar{X}$

In the table, $T_i.$ is the total and $\bar{X}_i.$ is the arithmetic mean of the ith row; $T_{\cdot j}$ is the total and $\bar{X}_{\cdot j}$ is the arithmetic mean of the jth column; and T is the total and $\bar{X}$ is the arithmetic mean of the whole *sample* of $N = lk$ items.

Let $E(X_{ij}) = a_{ij}$; $E(X^2_{ij}) = A_{ij}$; $\sum_1^k{}_j a_{ij} = ka_i$; $\sum_1^l{}_i a_i = la$; $\sum_1^l{}_i \bar{X}_i = l\bar{X}$; $\sum_1^k{}_j \bar{X}_j = k\bar{X}$.

Then, by III.19.1., for the ith row

$$E\left[\sum_1^k{}_j (X_{ij} - \bar{X}_i)^2\right] = \frac{k-1}{k} \sum_1^k{}_j (A_{ij} - a_{ij}^2) + \sum_1^k{}_j (a_{ij} - a_i)^2 \qquad \text{III.19.4.}$$

Summing III.19.4. for all the l rows, it is found that

$$E\left[\sum_1^l{}_i \sum_1^k{}_j (X_{ij} - \bar{X}_i)^2\right] = \frac{k-1}{k} \sum_1^l{}_i \sum_1^k{}_j (A_{ij} - a_{ij})^2 + \sum_1^l{}_i \sum_1^k{}_j (a_{ij} - a_i)^2 \qquad \text{III.19.5.}$$

Since $E(\bar{X}_{1.}) = a_1$, we note that

$$E(\bar{X}_{1.} - a_1)^2 = E(\bar{X}_{1.}^2) - 2a_1 E(\bar{X}_{1.}) + a_1^2 = E(\bar{X}_{1.}^2) - a_1^2 \text{ or}$$

$$E(\bar{X}_{1.}^2) = E(\bar{X}_{1.} - a_1)^2 + a_1^2 \qquad \text{III.19.6.}$$

Applying III.19.1. to $\bar{X}_{1.}$ $(i = 1, 2, \ldots, l)$,

$$E\left[\sum_1^l {}_1(\bar{X}_{1.} - \bar{X})^2\right] = \frac{l-1}{l} \sum_1^l {}_1 [E(\bar{X}_{1.}^2) - a_1^2] + \sum_1^l {}_1(a_1 - a)^2 \qquad \text{III.19.7.}$$

But

$$E(\bar{X}_{1.}^2) - a_1^2 = E(\bar{X}_{1.} - a_1)^2 = \frac{1}{k^2} \sum_1^k {}_j (A_{1j} - a_{1j}^2)$$

so that

$$E\left[\sum_1^l {}_1(\bar{X}_{1.} - \bar{X})^2\right] = \frac{l-1}{lk^2} \sum_1^l {}_1 \sum_1^k {}_j (A_{1j} - a_{1j})^2 + \sum_1^l {}_1 (a_1 - a)^2 \qquad \text{III.19.8.}$$

Applying III.19.1. to the $N = lk$ values, we get

$$E\left[\sum_1^l {}_1 \sum_1^k {}_j (X_{1j} - \bar{X})^2\right] = \frac{lk-1}{lk} \sum_1^l {}_1 \sum_1^k {}_j (A_{1j} - a_{1j}^2) + \sum_1^l {}_1 \sum_1^k {}_j (a_{1j} - a)^2 \qquad \text{III.19.9.}$$

By starting with the jth column and proceeding as in III.19.5., III.19.6., and III.19.7., it is found that

$$E\left[\sum_1^k {}_j \sum_1^l {}_1 (X_{1j} - \bar{X}_{.j})^2\right] = \frac{l-1}{l} \sum_1^k {}_j \sum_1^l {}_1 (A_{1j} - a_{1j}^2) + \sum_1^k {}_j \sum_1^l {}_1 (a_{1j} - b_j)^2 \qquad \text{III.19.10.}$$

and

$$E\left[\sum_1^k (\bar{X}_{.j} - \bar{X})^2\right] = \frac{k-1}{kl^2} \sum_1^k {}_j \sum_1^l {}_1 (A_{1j} - a_{1j}^2) + \sum_1^k {}_j (b_j - a)^2 \qquad \text{III.19.11.}$$

If the $N = lk$ values are statistically homogeneous or are all observations from the same population, then $A_{ij} = A$, $a_{ij} = a_i = b_j = a$ so that III.19.5., III.19.8., III.19.9., and III.19.10., and III.19.11., become, respectively

$$E\left[\sum_1^l{}_i \sum_1^k{}_j (X_{ij} - \bar{X}_{i.})^2\right] = \frac{k-1}{k} \cdot lk\,(A - a^2) = l\,(k-1)\,(A - a^2)$$

III.19.12.

$$E\left[\sum_1^l{}_i (\bar{X}_{i.} - \bar{X})^2\right] = \frac{l-1}{lk^2} \cdot lk\,(A - a^2) = \frac{l-1}{k}(A - a^2) \quad \text{III.19.13.}$$

$$E\left[\sum_1^l{}_i \sum_1^k{}_j (X_{ij} - \bar{X})^2\right] = \frac{lk-1}{lk} \cdot lk\,(A - a^2) = (lk-1)\,(A - a^2)$$

III.19.14.

$$E\left[\sum_1^k{}_j \sum_1^l{}_i (X_{ij} - \bar{X}_{.j})^2\right] = \frac{l-1}{l} \cdot lk\,(A - a^2) = k\,(l-1)\,(A - a^2)$$

III.19.15.

$$E\left[\sum_1^k{}_j (\bar{X}_{.j} - \bar{X})^2\right] = \frac{k-1}{kl^2} \cdot lk\,(A - a^2) = \frac{k-1}{l}(A - a^2) \quad \text{III.19.16.}$$

To summarize, it has been shown that in a statistically homogeneous set of $N = lk$ observations arranged in l rows and k columns, the following estimates of variance (or the following mean sums of squares) all have the same expected value:

$$(1)\ \frac{\sum_1^l{}_i \sum_1^k{}_j (X_{ij} - \bar{X})^2}{lk-1} \qquad (2)\ \frac{\sum_1^l{}_i \sum_1^k{}_j (X_{ij} - \bar{X}_{i.})^2}{l\,(k-1)}$$

III.19.17.

$$(3)\ \frac{\sum_1^k{}_j \sum_1^l{}_i (X_{ij} - \bar{X}_{.j})^2}{k\,(l-1)} \qquad (4)\ \frac{k \sum_1^l{}_i (\bar{X}_{i.} - \bar{X})^2}{l-1}$$

$$(5)\ \frac{l \sum_1^k{}_j (\bar{X}_{.j} - \bar{X})^2}{k-1}$$

Any significant differences between the estimates given in III.19.17. indicate lack of homogeneity of the set of items. The tests for this will be described in Chapter IV.

Let us now consider several special cases. Let p_{ij} be the probability that X has the value X_{ij} and let p_i be the average probability for the i th set, then

$$kp_i = \sum_1^k p_{ij}; \qquad lp = \sum_1^l p_i$$

and it can be shown by the use of III.19.1. that

$$\sum_1^l \sum_1^k (X_{ij} - \bar{X})^2 = lkpq - \sum_1^l \sum_1^k (p_{ij} - p_i)^2 + (k^2 - k) \sum_1^l (p_i - p)^2 \qquad \text{III.19.18.}$$

The special cases are:

(1) *Bernoulli series*: $p_{ij} = p_i = p$. Here III.19.18 becomes

$$\sum_1^l \sum_1^k (X_{ij} - \bar{X})^2 = lkpq$$

(2) *Lexis series*: $p_{ij} \equiv p_i$; $p_i \neq p$. Here III.19.18. becomes

$$\sum_1^l \sum_1^k (X_{ij} - \bar{X})^2 = lkpq + (k^2 - k) \sum_1^l (p_i - p)^2.$$

(3) *Poisson series*: $p_{ij} \neq p_i$; $p_i \equiv p$. Here III.19.18. becomes

$$\sum_1^l \sum_1^k (X_{ij} - \bar{X})^2 = lkpq - \sum_1^l \sum_1^k (p_{ij} - p)^2$$

The special cases expressed verbally are:

(1) *Bernoulli series*: The underlying probability p is constant from trial to trial and set to set or is constant throughout the whole field of observation and we have statistical homogeneity.

(2) *Lexis series*: The probability is constant from trial to trial within a set but varies from set to set and we do not have statistical homogeneity.

(3) *Poisson series*: The probability varies from trial to trial within a set of k trials, but the several probabilities for one set of k trials are identical to those of every other of l sets of k trials and we do not have homogeneity.

Illustrations of such series exist in the study of traffic on a given route at l different crossings at k different times with a total of $N = lk$ observations.

III. 20. *The Multinomial Distribution*: Let samples of size n be drawn from a specified universe with each sample divided into the k classes or cells with the distribution *random* among these classes or cells.

The probability, P, that there are f_{01} individuals in the first cell, f_{02} in the second cell, and so forth, is

$$P = \pi_1^{f_{01}} \pi^{f_{02}} \ldots\ldots \pi_k^{f_{0k}} \frac{n!}{f_{01}!\, f_{02}!\, \ldots\ldots\, f_{0k}!} \qquad \text{III.20.1.}$$

where π_1 is the probability that an individual falls in the first class or cell, π_2 the probability that it falls in the second cell, and so forth; and

$$\frac{n!}{f_{01}!\, f_{02}!\, \ldots\ldots\, f_{0k}!}$$

is the number of combinations of n things taken f_{01} of one kind, f_{02} of another kind, f_{0k} of the k-th kind.

To illustrate: At an intersection point it has been determined that the probability of turning left is $\frac{2}{5}$, of going straight ahead is $\frac{1}{2}$, and of turning right is $\frac{1}{10}$. Of 6 vehicles, what is the probability that one will turn left, two will go straight ahead, and 3 will turn right?

Solution: Here $\pi_1 = \frac{2}{5} = 0.4$, $\pi_2 = \frac{1}{2} = 0.5$, and $\pi_3 = \frac{1}{10} = 0.1$. Also $f_{01} = 1, f_{02} = 2, f_{03} = 3$. Substituting these values in III.20.1.,

$$P = (0.4)^1\,(0.5)^2\,(0.1)^3 \frac{6!}{1!\,2!\,3!}$$

$$= 0.0001\,(60) = 0.006$$

which means that 6 times in 1000 the event will happen as specified.

Let us now *assume* that each f_{0i} ($i = 1, 2, \ldots, k$) is large. Then, by the use of Stirling's asymptotic approximation to the factorials in III.20.1., it is found that

$$P \cong \pi_1^{f_{01}}\, \pi_2^{f_{02}} \ldots\ldots \pi_k^{f_{0k}} \frac{n^{n+\frac{1}{2}} e^{-n} \sqrt{2\pi}}{f_{01}^{f_{01}+\frac{1}{2}} e^{-f_{01}} \sqrt{2\pi} \ldots f_{0k}^{f_{0k}+\frac{1}{2}} e^{-f_{0k}} \sqrt{2\pi}} \qquad \text{III.20.2.}$$

where the symbol $\cong$ means "approximately equal to".

Since $\sum_1^k f_{0i} = n$, it is not hard to show that

$$P \cong \left(\frac{n\pi_1}{f_{01}}\right)^{f_{01}+\frac{1}{2}} \left(\frac{n\pi_2}{f_{02}}\right)^{f_{02}+\frac{1}{2}} \ldots\ldots \left(\frac{n\pi_k}{f_{0k}}\right)^{f_{0k}+\frac{1}{2}} \qquad \text{III.20.3.}$$

Now, let $f_{ti} = n\pi_i$ $(i = 1, 2, \ldots, k)$ and

$$X_i = \frac{f_{0i} - n\pi_i}{\sqrt{n\pi_i}} = \frac{f_{0i} - f_{ti}}{\sqrt{f_{ti}}} \qquad \text{III.20.4.}$$

for $i = 1, 2, \ldots, k$.

Substituting from III.20.4 in III.20.3, and transforming to logarithms, it is found that

$$\begin{aligned} \log P - \log K &= \sum_1^k \left\{ (f_{0i} + \tfrac{1}{2}) \log \frac{f_{ti}}{f_{0i}} \right\} \\ &= \sum_1^k (f_{0i} + \tfrac{1}{2}) \log \frac{f_{ti}}{f_{ti} + X_i\sqrt{f_{ti}}} \\ &= -\sum_1^k (f_{ti} + \tfrac{1}{2} + X_i\sqrt{f_{ti}}) \log \left(1 + \frac{X_i}{\sqrt{f_{ti}}}\right) \qquad \text{III.20.5.} \end{aligned}$$

It is next assumed that f_{ti} and f_{0i} for each i are of the *same order of magnitude*. It then follows that X_i will be small compared with f_{ti}. Expanding the logarithm in III.20.5 into a series, we have, to first order,

$$\log P - \log C \cong -\sum_1^k (f_{ti} + \tfrac{1}{2} + X_i\sqrt{f_{ti}}) \left(\frac{X_i}{\sqrt{f_{ti}}} - \tfrac{1}{2}\frac{X_i^2}{f_{ti}}\right) \qquad \text{III.20.6.}$$

$$\cong -\sum_1^k \left\{ \tfrac{1}{2} X_i^2 + X_i\sqrt{f_{ti}} \right\}$$

But $$\sum_1^k (X_i\sqrt{f_{ti}}) = \sum_1^k (f_{0i} - f_{ti}) = n - n = 0.$$

Hence

$$\log P - \log C \cong -\tfrac{1}{2}\sum_1^k X_i^2 \text{ and}$$

$$P = e^{-\frac{1}{2}\sum_1^k X_i^2} \qquad \text{III.20.7.}$$

From III.20.7, it is clear that P varies directly as the sum of k

normal independent variates of unit variance which are subject to the single constraint that $\sum_1^k (X_1/\sqrt{f_{t1}}) = 0$.

This is precisely χ^2 (Chi-square) as will be seen in Chapter IV. Hence,

$$\chi^2 = \sum_1^k X_1^2 = \sum_1^k \frac{(f_{01} - f_{t1})^2}{f_{t1}} \qquad \text{III.20.8.}$$

and is the probability of the sum of the squares of $(k - 1)$ independent normal variates each of unit variance.

The criterion given in III.20.8 is known as the *Chi-square test of goodness of fit* and is useful in testing the hypothesis that a sample at hand came from a universe of specified type.

The algebraic form of the distribution of χ^2 is

$$P(\chi^2) = \frac{1}{2^{\frac{k-1}{2}} \prod\left(\frac{k-1}{2}\right)} e^{-\frac{1}{2}\chi^2} (\chi^2)^{(\frac{1}{2})(k-3)} \qquad \text{III.20.9.}$$

Using the table on page 220 for this function an application is shown in Chapter V, page 163.

Thus far the underlying probability of success has been assumed constant. Suppose now that the probability of success is not constant, but depends on what has previously happened such as the case of finding r white balls from an urn that contains np white balls and nq black balls when s balls are drawn one at a time from the urn without replacements.

The solution of such a situation is given by the Hypergeometric Distribution.

III.21. *Hypergeometric Distribution*: Consider an urn in which there are np white balls and nq black balls. Draw s balls one at a time without replacements. The probability, P_r, that r ($r = 0, 1, 2, \ldots, s$) of the s balls are white is

$$P_r = y_r = \frac{\dfrac{(np)!}{r!\,(np-r)!} \cdot \dfrac{(nq)!}{(s-r)!\,(nq-s+r)!}}{\dfrac{n!}{s!\,(n-s)!}}$$

$$= \frac{(np)!\,(nq)!\,s!\,(n-s)!}{(np-r)!\,(nq-s+r)!\,n!\,r!\,(s-r)!} \qquad \text{III.21.1.}$$

To illustrate: Consider the case of 100 vehicles approaching an intersection of which np = 30 are trucks and nq = 70 are not trucks. Consider any s = 5 of these vehicles one at a time. The probability, $P_r = P_3$ that 3 of the 5 vehicles are trucks is

$$P_3 = \frac{30!\ 70!\quad 5!\ 95!}{27!\ 68!\ 100!\ 3!\ 2!} = 0.117$$

which means that 117 times out of 1000 sets of 5 vehicles the probability is that 3 vehicles out of 5 will be trucks.

Now, let

$$x = r + \tfrac{1}{2} \quad \text{and} \quad y = (y_r + y_{r+1})/2 \quad \text{and} \quad \left(\frac{dy}{dx}\right)_{(x,\,y)} = y_{r+1} - y_r.$$

Then,

$$\frac{dy}{dx_{(x,\,y)}} = y_r \frac{s + nps - nq - 1 - r(n+2)}{(r+1)(r+1+nq-s)} \qquad \text{III.21.2.}$$

From $y = (y_r + y_{r+1})/2$, it is found that

$$y = \tfrac{1}{2}\, y_r \frac{nps + nq + 1 - s - r(nq + 2 - np - 2s) + 2r^2}{(r+1)(r+1+nq-s)} \qquad \text{III.21.3.}$$

Replacing r by $x - \frac{1}{2}$, III.21.3. becomes

$$\left(\frac{1}{y}\right)\left(\frac{dy}{dx}\right) = \frac{2s + 2nps - 2nq - 2 - (2x-1)(n+2)}{nps + nq + 1 - s + (x - \frac{1}{2})(nq + 2 - np - 2s) + 2(x - \frac{1}{2})^2} \qquad \text{III.21.4.}$$

The equation given in III.21.4. is the equation of the system of curves which are continuous approximations to the law of probability given in III.21.1.

The curves are usually known as the Pearson system of frequency curves which are the particular solutions of the differential equation III.21.4.

The equation III.21.4., may be written in the form

$$\left(\frac{dy}{dx}\right) = \frac{y(x+a)}{b_0 + b_1 x + b_2 x^2} \qquad \text{III.21.5.}$$

which has 12 particular solutions or 12 specific types of curves dependent upon the values of the constants.[11]

The moments about the arithmetic mean of the distribution III.21.1., are

$$\mu_2 = \frac{spq\,(n-s)}{n-1}$$

$$\mu_3 = \frac{spq\,(q-p)\,(n-s)\,(n-2\,s)}{(n-1)\,(n-2)} \qquad \text{III.21.6.}$$

$$\mu_4 = \frac{spq\,(n-s)}{(n-1)\,(n-2)\,(n-3)}\left[n\,(n+1)-6s(n-s)+3pq\left\{n^2(s-2)-ns^2+6s\,(n-s)\right\}\right]$$

$$n\mu_{r+1} = \left\{(1+E)^r-E^r\right\}\left[\mu_2-\left\{np+s\,(q-p)\right\}\mu_1+\left\{spq\,(n-s)\,\mu_0\right\}\right] \qquad \text{III.21.7.}$$

where E is an operator and means that

$$E\mu_r = \mu_{r+1}\ (r = 0, 1, 2, \ldots.).$$

The maximum term of III.21.1. is approximately[5]

$$\sqrt{\frac{n}{2\,pqs\,(n-s)}} \qquad \text{III.21.8.}$$

If in III.21.6. and III.21.7., $n \to \infty$, the respective moments become the moments of the binomial distribution which shows that the binomial or Bernoulli distribution is the limiting case (or the case of a large or infinite universe) of the hyper-geometric distribution (or the case of a finite universe).

III. 22. *Correlation*[6]: The theory of correlation is devoted to the endeavor of finding laws of relationship (dependence) between two or more variables. Suppose a group of individuals is measured in regard to a certain attribute. It is found that the individuals differ in their measurements. It is desired to explain these differences in terms of factors on which this attribute is dependent and to obtain laws connecting the attribute with one or more such factors. The better the law of connection explains the variability in the attribute in question, the higher is the correlation.

To illustrate: One may wish to know whether the height of an individual can be explained or measured by the weight of an in-

dividual. In other words, are tall people heavy and short people not heavy. It is well known that weight alone does not measure height or explain the difference in the height of individuals. In this instance there are more factors than the one factor weight.

There are three main types of correlation: *simple correlation*, *multiple correlation*, and *partial correlation*. These will now be developed and discussed in the order named.

The Correlation Coefficient r-Linear Regression or *Linear Trend.* The regression or trend line is necessarily the best fitting line in the sense of least squares. The line may be curved or straight. To start with, let it be assumed that the regression (trend) line is a straight line. The equation of this line is

$$y = mx + b \qquad \text{III.22.1.}$$

The values of m and b must be determined and they are, respectively, the slope and y-intercept of the line. The x and y values are observed in pairs and they are the coordinates of any point on the line. The formula III.22.1. describes an infinite number of lines, each with its m as well as its b. No two different lines have the same m as well as the same b. If the lines are parallel, they have the same m but different b's. If the lines pass through the same point on the y-axis, they have the same b but different m's. We assume that any one of the possible lines has the same weight as any other one in arriving at a particular line, namely, the line that fits the data best in the theory of Least Squares. The *Principle of Least Squares*, used to determine the line of best fit, states that the line of best fit for a series of values is a line such that the sum of the squares of the vertical distances from it will be a minimum. There can obviously be only one line having this qualification. Another such line exists for the horizontal distances. However, the one for vertical distances is sufficient for most practical purposes.

In Figure III.4., suppose that the line $\overrightarrow{RR'}$ is the straight line of best fit for the plotted points (scatter diagram) shown, and that its equation is

$$y = mx + b \qquad \text{III.22.1.}$$

The y-distance, namely, y′, of any point (x_1, y_1) from this line is equal to

$$y_1 - (mx_1 + b) \qquad \text{III.22.2.}$$

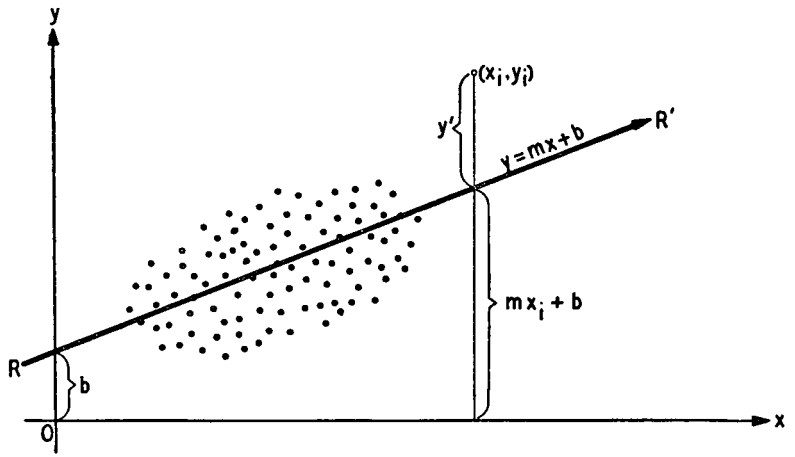


FIGURE III.4
ILLUSTRATION OF PRINCIPLE OF LEAST SQUARES

The sum of these distances squared must be a minimum. Symbolically,

$$d^2 = \sum_1^n (mx_1 + b - y_1)^2 \qquad \text{III.22.3.}$$

is to be a minimum. This necessitates that

$$\frac{\partial d}{\partial b} = +2 \sum_1^n (mx_1 + b - y_1) = 0 \qquad \text{III.22.4.}$$

and

$$\frac{\partial d}{\partial m} = +2 \sum_1^n x_1 (mx_1 + b - y_1) = 0 \qquad \text{III.22.5.}$$

From III.22.4.:

$$\sum_1^n y_1 = nb + m \sum_1^n x_1 \qquad \text{III.22.6.}$$

where n equals the number of cases or number of points. From III.22.5.:

$$\sum_1^n x_1 y_1 = b \sum_1^n x_1 + m \sum_1^n x_1^2 \qquad \text{III.22.7.}$$

Equations III.22.6 and III.22.7 are so-called "normal" equations for finding the least-square straight line. The two equations can be solved simultaneously to find the unknowns m and b. These two equations are all that are needed to determine the equation of the line of best fit. This line gives the relationship between the two variables x and y.

The procedure can be illustrated by an example. The required calculations can be done quite rapidly with tables and a calculating machine.

Example: Given the associated pairs of values for x and y:

x: 3, 5, 8, 12, 17, 23, 30
y: 1, 2, 6, 23, 40, 50, 60

Using these values in equations III.22.6 and III.22.7, it is found that

$$182 = 7\,b + 98\,m$$
$$3967 = 98\,b + 1960\,m$$

Solving these equations for b and m, we find that $m = 2.41$ and $b = -7.78$ whence

$$y = 2.41\,x - 7.78 \qquad \text{III.22.8.}$$

is the equation of the best fitting straight line. From III.22.6

$$m\bar{x} + b - \bar{y} = 0 \qquad \text{III.22.9.}$$

The equation III.22.9. expresses the fact that the linear function (straight line) passes through the point whose coordinates are $(\bar{x}, \bar{y})$.

Now measure all the x's and y's from their respective means as origin and replace every x by its deviation x′ from $\bar{x}$, and y by its deviation y′ from $\bar{y}$. Then III.22.9. becomes, since b now is zero,

$$y' = mx' \qquad \text{III.22.10.}$$

and III.22.7 becomes

$$m \sum_1^n x_1'^2 - \sum_1^n x_1' y_1' = 0$$

from which

$$m = \frac{\sum_1^n x_1' y_1'}{\sum_1^n x_1'^2} = \frac{np}{n\sigma_x^2} = \frac{p}{\sigma_x^2} \qquad \text{III.22.11.}$$

It follows that

$$y' = \frac{p}{\sigma_x^2} x'$$

whence

$$\bar{y}_x - \bar{y} = \frac{p}{\sigma_x^2}(x - \bar{x}) \qquad \text{III.22.12.}$$

It is important to note that $\bar{y}_x$ is the computed value of y for a given x from the equation of the least-square line. For the line to be a regression (trend) line, it is necessary that $\bar{y}_x$ is the arithmetic mean (or close to being so) of the values of y associated with a given value of x.

Similarly

$$\bar{x}_y - \bar{x} = \frac{p}{\sigma_y^2}(y - \bar{y}) \qquad \text{III.22.13.}$$

The coefficient p/σ_x^2 gives the deviation in y from the mean y corresponding to unit deviation in x from the mean x, for when $x - \bar{x} = 1$, $\bar{y}_x - \bar{y} = p/\sigma_x^2$. Likewise, p/σ_y^2 gives the deviation in x from the mean x corresponding to unit deviation in y from the mean y.

But, in general, $p/\sigma_y^2 \neq p/\sigma_x^2$. This demands the necessity of altering the unit of measure so that unit change in x and y are of the same magnitude. Then

$$\frac{\bar{y}_x - \bar{y}}{\sigma_y} = \frac{p}{\sigma_x \sigma_y}\left(\frac{x - \bar{x}}{\sigma_x}\right) \qquad \text{III.22.14.}$$

and

$$\frac{\bar{x}_y - \bar{x}}{\sigma_x} = \frac{p}{\sigma_x \sigma_y}\left(\frac{y - \bar{y}}{\sigma_y}\right) \qquad \text{III.22.15.}$$

Next, write

$$\frac{p}{\sigma_x \sigma_y} = r$$

the *coefficient of correlation.* Hence

$$\bar{y}_x - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \text{III.22.16.}$$

and

$$\bar{x}_y - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \text{III.22.17.}$$

which are the regression (trend) lines. The numbers $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$ are called the coefficients of regression or of the trend.

Consider

$$\bar{y}_x - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad \text{or} \quad y' = r\frac{\sigma_y}{\sigma_x}x'.$$

Then

$$\begin{aligned} d &= \sum_1^n{}_1 \left(y'_1 - r\frac{\sigma_y}{\sigma_x}x'_1\right)^2 \\ &= \sum_1^n{}_1 y'^2_1 - 2r\frac{\sigma_y}{\sigma_x}\sum_1^n{}_1 x'_1 y'_1 + r^2\frac{\sigma^2_y}{\sigma^2_x}\sum_1^n{}_1 x'^2_1 \\ &= n\sigma^2_y - 2r\frac{\sigma_y}{\sigma_x}(nr\sigma_y\sigma_x) + r^2\frac{\sigma^2_y}{\sigma^2_x}(n\sigma^2_x) \\ &= n\sigma^2_y(1 - r^2) \qquad \text{III.22.18.} \end{aligned}$$

Since d being the sum of squares is positive, we have

$$n\sigma^2_y(1 - r^2) > 0 \quad \text{and}$$

$$-1 \leq r \leq 1 \qquad \text{III.22.19.}$$

and

$$r = \pm 1 \text{ when } \left|\frac{y'_1}{x'_1}\right| = \left|\frac{\sigma_y}{\sigma_x}\right|.$$

Now

$$np = \sum_1^n{}_1 x'_1 y'_1 \text{ and } x'_1 = x_1 - \bar{x};\ y'_1 = y_1 - \bar{y}.$$

Hence

$$np = \sum_1^n{}_1 (x_1 - \bar{x})(y_1 - \bar{y}) = \sum_1^n{}_1 (x_1 y_1) - n\bar{x}\bar{y}.$$

Hence

$$p = \frac{\overset{n}{\underset{1}{\Sigma}} x_1 y_1}{n} - \bar{x}\bar{y}.$$

But

$$r = \frac{p}{\sigma_x \sigma_y}.$$

Hence

$$r = \frac{\dfrac{\overset{n}{\underset{1}{\Sigma}} x_1 y_1}{n} - \bar{x}\bar{y}}{\sigma_x \sigma_y} = \frac{\dfrac{\overset{n}{\underset{1}{\Sigma}} x_1 y_1}{n} - \bar{x}\bar{y}}{\sqrt{\dfrac{\overset{n}{\underset{1}{\Sigma}} x_1^2}{n} - (\bar{x})^2}\sqrt{\dfrac{\overset{n}{\underset{1}{\Sigma}} y_1^2}{n} - (\bar{y})^2}}$$

$$= \frac{\overset{n}{\underset{1}{\Sigma}} (x_1 - \bar{x})(y_1 - \bar{y})}{n \sigma_x \sigma_y} = \frac{\overset{n}{\underset{1}{\Sigma}} x_1'' y_1''}{n} = \frac{\overset{n}{\underset{1}{\Sigma}} x_1' y_1'}{n \sigma_x \sigma_y} \qquad \text{III.22.20.}$$

From this relation, it is fairly clear that r may be considered as the cosine of the angle between two vectors in Euclidean n space. Again, from this fact, it follows that $-1 \leq r \leq 1$. Also, r is the arithmetic mean of the products of the deviations of the corresponding values from the respective arithmetic means when measured in standard deviation units; also, r is sometimes called the product-moment coefficient.

The formulas useful in finding the value of the coefficient of correlation are as follows:

(1) If the variables are in original units with respect to their natural origin, then

$$r = \frac{\dfrac{\overset{n}{\underset{1}{\Sigma}} X_1 Y_1}{n} - \bar{X}\bar{Y}}{\sigma_x \sigma_y} \qquad \text{III. 22. 21.}$$

(2) If the variables are referred to a class mid-point as an origin and in terms of the class interval as a unit, then

$$r = \frac{\frac{\sum_1^n x_1 y_1}{n} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} \qquad \text{III.22.22.}$$

These formulas are readily obtained algebraically from III. 22. 20.

To interpret r, it is necessary to use r^2 which is called the determining coefficient.

If r, say, equals 0.70, we find that $r^2 = 0.49$ which means that 49 per cent of the variability in the y-values is determined or explained by the potential determining or measuring factor x and the linear theory connecting y with x. In other words, the theory used or tested is but 49 per cent efficient as an estimator or forecasting or predicting theory.

III. 23. *Basic theory of correlation.* To explain the Basic Theory of Correlation let us suppose that we have given n pairs of values for the variables x and y. The problem is to determine the nature and degree of the dependence between the x values and their corresponding y values.

To determine the amount of interdependence that exists between the pairs of variables it is convenient to represent them by points in a two dimensional Euclidean manifold (scatter diagram). To facilitate a description of the dependence we partition the data into classes. This is accomplished by selecting class intervals of size dx. We recall that the set of y values associated with a given value of x on an interval of size dx is called an x array of y's. If it is desired to describe the behavior of the expected values of the y values associated with the x values, it is necessary to find the equation of the curve $y = f(x)$ that passes through these points. This curve is known as the estimate of the true regression curve. The limiting curve that is approached as dx tends toward zero is the true regression curve (trend) of y on x and is actually the locus of the arithmetic mean of arrays of y values of the theoretical distribution as dx tends toward zero. The description of the theoretical

law of behavior appertaining to the arrangement of y is the solution of the problem of statistical dependence (regression or trend analysis) of y on x.

To illustrate: Consider the related value of minimum spacing, center to center in feet, with speed in miles per hour.

Table III.4. is a correlation table which shows numerically as well as graphically the two-way distribution connecting minimum spacing, center to center in feet with speed in miles per hour as found by actual observation. The first question to be answered is: How dependent upon the speed of a vehicle is the minimum spacing? The answer to this question is found in interpreting the value of the *determining coefficient* which is the square of the *correlation coefficient.*

Substituting in III.22.22 the required values from Table III..,4 it is found that

$$r = \frac{\frac{\Sigma(xy)}{n} - \bar{x}\,\bar{y}}{\sigma_x\,\sigma_y}$$

becomes

$$r = \frac{\frac{47440}{1336} - \left(\frac{-3321}{1336}\right)\left(\frac{-9849}{1336}\right)}{\sqrt{\frac{58771}{1336} - \left(\frac{-3321}{1336}\right)^2}\sqrt{\frac{113049}{1336} - \left(\frac{-9849}{1336}\right)^2}}$$

$$= \frac{35.509 - (-2.486)(-7.372)}{\sqrt{44.090 - 6.180}\sqrt{84.618 - 54.346}}$$

$$= \frac{35.509 - 18.327}{(6.149)(5.502)} = \frac{17.182}{33.832}$$

$$= 0.5079 = 0.51 \qquad \text{III.23.1.}$$

This result means that $(0.5079)^2 = 0.2580 = .26 = 26$ per cent of the variability in minimum spacing is explained by or dependent upon the speed of the vehicle and the assumed linear connection between spacing and speed. In other words, it appears that speed is an unimportant or minor factor for determining minimum spacing.

Table III.4.

Speed in miles per hour

Spacing center to center in feet	0–1.9	2–3.9	4–5.9	6–7.9	8–9.9	10–11.9	12–13.9	14–15.	16–17.9	18–19.9	20–21.9	22–23.9	24–25.9	26–27.9	28–29.9	30–31.9	32–33.9	34–35.9	36–37.9	38–39.9	40–41.9	42–43.9	44–45.9	46–47.9	48–49.9	50–51.9	52–53.9	fy	y	f_y (y)	f_y (y^2)	Σx	y Σx
139–142															1			1										2	16	32	512	5	80
135–138															1		1	1										3	15	45	675	8	120
131–134															1			1										2	14	28	392	5	70
127–130																													13				
123–126																		1						1				2	12	24	288	14	168
119–122													1															1	11	11	121	−1	−11
115–118													1			1	1											3	10	30	300	4	40
111–114									1			1		1					1	1								5	9	45	405	4	36
107–110										1		2		1	1		2							1				8	8	64	512	9	72
103–106									3			1	1	1	1		2	1				1						11	7	77	539	1	7
99–102							1						1							2		1						5	6	30	180	12	72
95–98											2	3	1		2	1	2		1	2	1	2						17	5	85	425	37	185
91–94											1	1	4		1		2		1									10	4	40	160	3	12
87–90								1				4	5	4	1	5	1			1	1							23	3	69	207	8	24
83–86										1	1	3	3	2	1	5	2	2	3									23	2	46	92	24	48
79–82					1			1	1	1	1		1	2	2	1	2	3		2			2					20	1	20	20	24	24
75–78								1		2	1		2		3	3	5	4	1	1		1						24	0			40	
71–74											1		3		5	5	1	7	3	4	1	2						32	−1	−32	32	102	−102
67–70							1			1	2	4	2	2	1	3	4	3	5	4	1	2						35	−2	−70	140	76	−152
63–66					2			1	3	1	2	3	1	1	11	3	12	2	6	3	2	1						54	−3	−162	486	75	−225
59–62					1			1		1	1	2	3		6	10	9	6	2	4	1		2	1		1		51	−4	−204	816	129	−516
55–58				1		3	2		1	1	3	4	2		9	9	11	5	4	4	1		1					61	−5	−305	1525	64	−320
51–54			2	1	2	1	1		2	3	3	2	9		8	6	9	6	2	3				1				61	−6	−366	2196	0	
47–50			3	4	2		2	2	1	3	3	4	4	3	17	6	9	4	6	7		1			1			82	−7	−574	4018	8	−56
43–46		1	1		6	2	6	4	7	1	3	1			14	8	11	11	3	4			1		1			85	−8	−680	5440	−43	344
39–42	2	3	2	7	13	9	9	3	6	2	6	6	7	2	19	9	8	6	2	3	3		1		1			129	−9	−1161	10449	−345	3105
35–38	1	4	12	9	14	2	5	5	3	2	3	7	11	14	5	3	7	10	2	1	2		1	1		1	1	126	−10	−1260	12600	−401	4010
31–34		6	6	9	6	6	11	3	5	5	11	2	6	6	10	9	5	5	3	3	1							118	−11	−1298	14278	−410	4510
27–30	4	22	26	17	8	7	8	6	1	8	6	8	6	1	10	5	2	5	1		1			1				153	−12	−1836	22032	−1001	12012
23–26	19	24	19	17	4	2	2	5	6	3	1	4	5	5	3		1	2			1							123	−13	−1599	20787	−1047	13611
19–22	23	13	3	3	4	3	2		1	1		1	2			1												57	−14	−798	11172	−603	8442
15–18	6	2		2																								10	−15	−150	2250	−122	1830
																												1336		−9849	113049	−3321	47440

Speed in miles per hour	f_x	x	f_x (x)	f_x (x^2)	Σy	xΣy	$\bar{y}_x$
0–1.9	55	−13	−715	9295	−735	9555	23.05
2–3.9	75	−12	−900	10800	−929	11148	26.95
4–5.9	74	−11	−814	8954	−846	9306	30.62
6–7.9	70	−10	−700	7000	−788	7880	30.66
8–9.9	63	−9	−567	5103	−610	5490	39.66
10–11.9	35	−8	−280	2240	−356	2848	35.81
12–13.9	50	−7	−350	2450	−476	3332	38.42
14–15.	33	−6	−198	1188	−296	1776	40.62
16–17.9	41	−5	−205	1025	−301	1505	47.14
18–19.9	37	−4	−148	592	−292	1168	44.93
20–21.9	51	−3	−153	459	−366	1098	47.79
22–23.9	63	−2	−126	252	−332	664	55.42
24–25.9	81	−1	−81	81	−441	441	54.72
26–27.9	45				−287	0	50.99
28–29.9	133	1	133	133	−797	−797	51.91
30–31.9	93	2	186	372	−490	−980	55.42
32–33.9	109	3	327	981	−493	479	58.41
34–35.9	86	4	344	1376	−444	−1776	55.85
36–37.9	46	5	230	1150	196	−980	59.46
38–39.9	49	6	294	1764	−190	−1140	61.00
40–41.9	16	7	112	784	−93	−651	53.25
42–43.9	11	8	88	704	7	56	79.05
44–45.9	8	9	72	648	−38	−342	60.94
46–47.9	6	10	60	600	−12	−120	68.50
48–49.9	3	11	33	363	−24	−264	45.83
50–51.9	2	12	24	288	−14	−168	48.50
52–53.9	1	13	13	169	−10	−130	36.50
	1336		−3321	58771	−9849	47440	

This means that either there are several other factors which together would explain 74 per cent of the variability or that there exists a possible single other factor or that the relationship is not linear. Of these, it appears that the former is the most likely.

A second question that needs to be answered is: What is the equation of the linear law of relationship which is useful to predict the expected minimum spacing when the speed is known.

To answer this, it is necessary to use the regression equation III.22.16, namely:

$$\bar{y}_x - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

Substituting the values indicated by the use of Table III.4. and III.23.1, it is found that

$$\bar{y}_x - 47.0 = 0.508\,\frac{22.008}{12.300}(x - 22.0) \qquad \text{III.23.2.}$$

whence

$$\bar{y}_x = 0.909\,x + 27.0$$

The graph of this equation is shown in Figure III.3. To illustrate the use of III.23.2, suppose it is desired to know the minimum spacing in feet if the speed is, say, 30 miles per hour. To answer this question, substitute 30.0 for x in equation III.23.2, whence the minimum spacing $\bar{y}_x$ is found to be 54.3 feet. This means that the expected minimum spacing center to center in feet or on the average the minimum spacing center to center in feet is 54.3 feet when the speed is 30.0 miles per hour.

A very important question now to be answered is: How typical or reliable is the expected minimum spacing of 54.3 feet. This question will be answered in article III.25.

III. 24. *Coefficient of Regression*: Consider

$$f = \sum_1^n n_{x_1}\,(y_{n_{x_1}} - mx_1 - b)^2$$

For f to be minimum

$$\frac{\partial f}{\partial m} = 0 \quad \text{and} \quad \frac{\partial f}{\partial b} = 0. \qquad \text{III.24.1.}$$

From equations III.24.1.,

$$m = \frac{\sum_1^n n_{x_1} \bar{y}_{n_{x_1}} x_1}{\sum_1^n n_{x_1} x_1^2} = \frac{\sum_1^n n_{x_1} x_1 \bar{y}_{n_{x_1}}/n}{\sum_1^n n_{x_1} x_1^2/n}$$

$$= \frac{\sum_1^n (x_1 y_1)/n}{\sigma_x^2} = \frac{r\,\sigma_x \sigma_y}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x}.$$

III. 25. *Standard Deviation of Arrays*:

Consider

$$nS_y^2 = \sum_1^n \left(y_1 - r\frac{\sigma_y}{\sigma_x} x_1\right)^2$$

$$= \sum_1^n y_1^2 - 2r\frac{\sigma_y}{\sigma_x}\sum_1^n (y_1 x_1) + r^2\frac{\sigma_y^2}{\sigma_x^2}\sum_1^n x_1^2$$

$$= n\sigma_y - 2nr^2\sigma_y^2 + nr^2\sigma_y^2$$

$$= n\sigma_y^2(1 - r^2)$$

Hence:

$$S_y^2 = \sigma_y(1 - r^2) \qquad \text{III.25.1.}$$

S_y may be regarded as a sort of average value of the standard deviations of the arrays of y's and is sometimes called the root-mean-square error of estimate of y, or more briefly, the standard error of estimate of y. The factor $(1 - r^2)^{\frac{1}{2}}$ is called the coefficient of alienation or the measure of the failure to improve the estimate of y from the knowledge of correlation.

If S_y is regarded as a function of x, say S (x), the curve

$$y = S(x)\,\sigma_y$$

is called the scedastic curve. Its ordinates measure the scatter in the arrays of y's in comparison to the scatter of all the y's. If S (x) is a constant, the regression system of y on x is called a homoscedastic system. If S (x) is not a constant, the system is said to be heteroscedastic. For a homoscedastic system with linear regression, $S_y = \sigma_y\,(1 - r^2)^{\frac{1}{2}}$ is the standard deviation of each erray of y's.

Similarly, for the dispersion of x on y, we have $S_x^2 = \sigma_x^2 (1 - r^2)$. Going back to the spacing speed illustration given in article III.22 where it was found that the expected spacing is 54.3 feet when the speed is 30.0 miles per hour. To determine the dependability of the value found for spacing, it is necessary to obtain its standard error or its measure of variability. This is given by III.25.1, namely: if S_y^2 is the variance of the expected values for spacing, then

$$S_y^2 = \sigma_y (1 - r^2).$$

Substituting the values for σ^2_y and r^2 found earlier in this chapter, we find that

$$S_y^2 = 484.35 (1 - .2580)$$
$$= 359.39$$

whence $S_y = 19.0$

This means that on the average, when the speed is 30.0 miles per hour, the spacing differs from the expected spacing of 54.3 feet by 19.0 feet. In other words, the probable or expected spacing lies between $54.3 - 19.0 = 35.3$ feet, and $54.3 + 19.0 = 73.3$ feet when the speed is 30.0 miles per hour. It is fairly obvious that the ability to predict the spacing knowing the speed is very poor and of very little practical value.

III. 26. *Correlation Ratio*: *Non-Linear Regression*: From III.25. it may be seen that

$$r^2 = 1 - S_y^2/\sigma_y \qquad \text{III.26.1.}$$

If $S_y = 0$, $r = 1$ and all the dots on the scatter diagram fall exactly on the line of regression $y = r\frac{\sigma_y}{\sigma_x}$. If $S_y = \sigma_y$, $r = 0$ and the regression line is of no aid in predicting y from an assigned x.

Now, let S'_y be the mean square of the deviations from the means of arrays. Then $S'^2_y = S_y^2$ when the regression is linear and $S'^2_y \neq S_y^2$ when the regression is not linear. This fact suggests the use of

$$\eta^2_{yx} = 1 - \frac{S'^2_y}{\sigma_y^2} \qquad \text{III.26.2.}$$

where η_{yx} is the correlation ratio of y on x and $S_y'^2$ is the mean square of the deviations from the means of arrays whether these means are near to or far from the proposed line of regression. For linear regression of y on x, we have $\eta^2_{yx} = r^2$. Similarly for x on y, we have

$$\eta^2_{xy} = 1 - \frac{S_x'^2}{\sigma_x^2} \qquad \text{III.26.3.}$$

To illustrate the finding of the value of correlation ratio which actually is the true measure of correlation, the procedure is to find η^2_{yx} from equation III.26.2. where

$$\eta^2_{yx} = 1 - \frac{S_y'^2}{\sigma_y^2}$$

As was explained, $(S_y')^2$ is the mean square of the deviations from the means of arrays, namely

$$(S_y')^2 = \frac{f_1 s_1^2 + f_2 s_2^2 + \ldots\ldots + f_1 s_1^2 + \ldots + f_k^2 s_k^2}{n} \qquad \text{III.26.4.}$$

where f_1 is the frequency of the ith vertical array – the array when x has the value x_1 and s_1^2 is the variance of the ith array. From III.26.1., it is clear that $f_1 s_1^2$ is actually the sum of the squares of the deviations of the values for the ith array of y's from the arithmetic mean of the ith array of y's.

Making use of Table III.4., it is found that, beginning with the first array of y's, namely, the array of y's when $x = 0.95$, then the second array when $x = 2.95$ and so on...,

$$\begin{aligned} f_1 s_1^2 = {} & \\ 2\,(40.5 - 23.1)^2 + {} & \\ 1\,(36.5 - 23.1)^2 + {} & \\ 4\,(28.5 - 23.1)^2 + {} & \\ 19\,(24.5 - 23.1)^2 + {} & \\ 23\,(20.5 - 23.1)^2 + {} & \\ 6\,(16.5 - 23.1)^2 & \\ = 1355.9 & \end{aligned} \qquad \begin{aligned} f_2 s_2^2 = {} & \\ 1\,(44.5 - 27.0)^2 + {} & \\ 3\,(40.5 - 27.0)^2 + {} & \\ 4\,(36.5 - 27.0)^2 + {} & \\ 6\,(32.5 - 27.0)^2 + {} & \\ 22\,(28.5 - 27.0)^2 + {} & \\ 24\,(24.5 - 27.0)^2 + {} & \\ 13\,(20.5 - 27.0)^2 + {} & \\ 2\,(16.5 - 27.0)^2 & \\ = 2364.7 & \end{aligned}$$

Similarly, it is found that

$f_3\ s_3^2 = 4108.8$ $\quad$ $f_{15}\ s_{15}^2 = 59855.0$
$f_4\ s_4^2 = 5272.5$ $\quad$ $f_{16}\ s_{16}^2 = 33508.7$
$f_5\ s_5^2 = 5489.2$ $\quad$ $f_{17}\ s_{17}^2 = 45523.0$
$f_6\ s_6^2 = 3891.0$ $\quad$ $f_{18}\ s_{18}^2 = 49788.0$
$f_7\ s_7^2 = 8295.6$ $\quad$ $f_{19}\ s_{19}^2 = 14902.0$
$f_8\ s_8^2 = 1069.8$ $\quad$ $f_{20}\ s_{20}^2 = 19500.7$
$f_9\ s_9^2 = 22976.7$ $\quad$ $f_{21}\ s_{21}^2 = 6950.7$
$f_{10}\ s_{10}^2 = 15353.5$ $\quad$ $f_{22}\ s_{22}^2 = 2578.5$
$f_{11}\ s_{11}^2 = 18564.5$ $\quad$ $f_{23}\ s_{23}^2 = 2068.6$
$f_{12}\ s_{12}^2 = 40986.3$ $\quad$ $f_{24}\ s_{24}^2 = 7680.0$
$f_{13}\ s_{13}^2 = 50938.5$ $\quad$ $f_{25}\ s_{25}^2 = 37.1$
$f_{14}\ s_{14}^2 = 29733.6$ $\quad$ $f_{26}\ s_{26}^2 = 288.0$
$f_{27}\ s_{27}^2 = 0$

Substituting the values of the $f_i\ s_i^2$ just found in III.26.1, it is found that

$$(S'_y)^2 = \frac{453080.9}{1336} = 339.1$$

From Table III.4, and III.23.1 it was found that

$$S_y^2 = 16\,[84.618 - 54.346]$$
$$= 16\,(30.272) = 484.4$$

Substituting the values just found for $(S'_y)^2$ and S_y^2 in III.26.2., it is found that

$$\eta_{yx}^2 = 1 - \frac{339.1}{484.4} = 1 - 0.70 = 0.30$$

Previously in III.23.1 it was found that, on the hypothesis of linear regression, the determining coefficient $r^2 = .26$. If the regression is not linear, we have found that the determining ratio – the real and proper measure of correlation – is 0.30. A legitimate question: Is the difference between the determining ratio and the determining coefficient large enough to justify the rejection of the hypothesis of linear regression? The technique to answer this question will be shown in Chapter IV.

The reader is cautioned *not* to follow the usual practice of tacitly assuming linear regression and in this sense finding the value of r^2. The proper procedure is to find η^2 first. Then it should be

determined whether η^2 is large enough to justify the obtaining of the actual regression (trend) function as well as whether η^2 is large enough to indicate that a significant correlation exists. The former is discussed and shown in III.29. and the latter in Chapter IV.

In the case just illustrated it is true that $\eta^2 = 0.30$ indicates real correlation, but it is much too small for predicting or estimation purposes. It is also true that there are sufficient grounds, as will be seen in III.29. to reject the hypothesis of linear regression.

A mean square of the deviations in each array is a minimum when the deviations are taken from the mean of the array. Hence, the $(S'_y)^2$ in III.26.2. must be equal to or less than S_y^2 in III.26.1. for the same data, since the deviations in III.26.1. are measured from the proposed line of regression. Hence, we have shown that

$$1 \geq \eta_{yx}^2 \geq r^2$$

It follows from III.26.2. that $\eta_{yx} \leq 1$.

If regression of y on x is linear, $\eta_{yx}^2 - r^2$ found from the sample differs from zero by an amount not greater than fluctuations due to random sampling. A comparison of $\eta_{yx}^2 - r^2$ with its sampling error is a useful criterion for testing linearity of regression. A better and more powerful method, however, to test linearity of regression is by the use of the *Analysis of Variance.*

III. 27. *Multiple Correlation*: Suppose we have given N sets of corresponding values of n variables $x_1, x_2, \ldots, x_n$. Now separate the values of x_1 into classes by selecting class intervals $dx_2, dx_3, \ldots, dx_n$ of the remaining variables.

The locus of means of such arrays of x_1's in the theoretical distribution, as $dx_2, \ldots dx_n$ approach zero is called the regression surface (trend) of x_1 on the remaining variables. We now assume, for convenience, that any variable, x_j, is measured from its arithmetic mean as origin. Let σ_j be its standard deviation and let r_{pq} be the correlation coefficient of the n given pairs of values of x_p and x_q. We now seek to find $b_{12}, b_{13}, \ldots, b_{1n}$ of the linear regression surface

$$x_1 = b_{12}\,x_2 + b_{13}\,x_3 + \ldots + b_{1n}\,x_n + c \qquad \text{III.27.1.}$$

of x_1 on the remaining variables so that x_1 computed from III.27.1. will give the best estimates in the sense of Least Squares

of the values of x_1 that correspond to any assigned values of $x_2, \ldots, x_n$. It follows that

$$U = \Sigma (x_1 - b_{12} x_2 - b_{13} x_3 - \ldots - b_{1n} x_n - c)^2 \qquad \text{III.27.2.}$$

shall be a minimum. This gives us for the linear regression surface

$$x_1 = -\sigma_1 \sum_{2}^{n} q \frac{R_{1q}}{R_{11}} \frac{x_q}{\sigma_q} \qquad \text{III.27.3.}$$

where

$$R = \begin{vmatrix} r_{11}, & r_{12}, & \ldots, & r_{1n} \\ r_{21}, & r_{22}, & \ldots, & r_{2n} \\ : & : & & \\ : & : & & \\ r_{n1}, & r_{n2}, & \ldots, & r_{nn} \end{vmatrix}$$

and R_{pq} is the cofactor of the pth row and qth column of R.

If the dispersion $\sigma_{1.23\ldots n}$ of the *observed* values of x_1 from computed values is *defined* as

$$\sigma^2_{1.23\ldots n} = \frac{1}{n} \Sigma \text{ (observed } x_1 - \text{computed } x_1)^2 \qquad \text{III.27.4.}$$

then, it can be proved that

$$\sigma^2{}_{1.23\ldots n} = \sigma_1^2 \left(\frac{R}{R_{11}} \right) \qquad \text{III.27.5.}$$

We are next interested in the dispersion of the estimated values given by III.27.3. Since the mean value of the estimates is zero, when the origin is at the mean of each system of variates, it can be shown that

$$\sigma^2_{1E} = \sigma_1^2 \left(1 - \frac{R}{R_{11}} \right) \qquad \text{III.27.6.}$$

The square of the multiple correlation coefficient $r_{1.23\ldots n}$ of order $(n - 1)$ of x_1 with the other $n - 1$ variable is given by

$$r^2_{1.23\ldots n} = 1 - \left(\frac{R}{R_{11}} \right) \qquad \text{III.27.7.}$$

The analysis of data furnished by J. S. Ellerby, Safety Director, Fort Belvoir, Virginia will serve as an example of multiple correlation. These data consist of the following information on 440 drivers:

x_1 = Road Test
x_2 = Years of Experience

x_3 = Reaction Time
x_4 = Distance Judgment
x_5 = Driver Information (Written test)

Let us assume that the road test is a measure of driver ability and let it be our problem to determine whether each of the other tests individually or collectively measure driving ability.

The first step is to determine the simple correlation between each of the tests. The procedure for this is that followed in the example of finding the correlation between speed and minimum spacing.

These correlations are shown in Table III.5. Before using these results to obtain a multiple correlation let us consider the significance of these simple correlations. It is noted immediately that none of them is large enough to be significant and therefore our conclusion is that none of the tests is of value as a measure of driving ability.

Table III.5

SIMPLE CORRELATION OF DRIVER TESTS

	(1) *Road Test*	(2) *Years Experience*	(3) *Reaction Time*	(4) *Distance Judgment*	(5) *Driver Information*
(1) *Road Test*	$r_{11}=1.0000$	$r_{12}=.0476$	$r_{13}=.0257$	$r_{14}=.05514$	$r_{15}=0.2608$
(2) *Yrs. Experience*	$r_{21}=.0476$	$r_{22}=1.0000$	$r_{23}=.006157$	$r_{24}=.00101$	$r_{25}=-0.4603$
(3) *Reaction Time*	$r_{31}=.0257$	$r_{32}=.006157$	$r_{33}=1.0000$	$r_{34}=-.0404$	$r_{35}=-.1027$
(4) *Distance Judgment*	$r_{41}=.05514$	$r_{42}=.00191$	$r_{43}=-.0404$	$r_{44}=1.0000$	$r_{45}=.1568$
(5) *Driver Information*	$r_{51}=\upsilon.2608$	$r_{52}=-0.4603$	$r_{53}=-.1027$	$r_{54}=.1568$	$r_{55}=1.0000$

At least one of the correlations is opposite to what one might expect. A driver with an increase in experience apparently knows less about driving since the correlation is negative (—.46). However, since $r^2 = (.46^2) = .21 = 21$ per cent, only this amount of the variable in driving knowledge may be said to be explained or dependent upon experience, consequently it may be said that there is little or no connection between driving ability and experience.

We would not of course be justified in concluding from this one study that drivers' tests have no value, for it may be that all of the drivers tested are good drivers and their visual acuity, reaction time, and other capabilities are well within the safe range. For example, the total range of reaction time was from .350 to .560 seconds. A driver with a reaction time much slower than .56 might be an accident prone driver. It is fair to say that it is quite a bit more likely than not, however, that these deductions are valid.

The next question to be answered is that of whether the tests as a whole give any indication of driving ability, i. e., whether the sets of data x_2, x_3 x_4, and x_5 taken together furnish us with a measure of driving ability. To answer this question, we make use of the theory of multiple linear correlation. The first step in the analysis is to find the multiple linear regression equation. This is done by substituting the values for the r's from Table III.5, in equation III.27.3. and solving by determinants.

$$x_1 = \sigma_1 \left[\frac{R_{12}\, x_2}{R_{11}\, \sigma_2} + \frac{R_{13}\, x_3}{R_{11}\, \sigma_3} + \frac{R_{14}\, x_4}{R_{11}\, \sigma_4} + \frac{R_{15}\, x_5}{R_{11}\, \sigma_5} \right]$$

$$= -\frac{1}{2}\frac{R_{12}}{R_{11}} x_2 - \frac{1}{3}\frac{R_{13}}{R_{11}} x_3 - \frac{1}{4}\frac{R_{14}}{R_{11}} x_4 - \frac{1}{5}\frac{R_{15}}{R_{11}} x_5$$

$$= +\frac{1}{2}\frac{\begin{vmatrix} r_{21} & r_{23} & r_{24} & r_{25} \\ r_{31} & r_{33} & r_{34} & r_{35} \\ r_{41} & r_{43} & r_{44} & r_{45} \\ r_{51} & r_{53} & r_{54} & r_{55} \end{vmatrix}}{\begin{vmatrix} r_{22} & r_{23} & r_{24} & r_{25} \\ r_{32} & r_{33} & r_{34} & r_{35} \\ r_{42} & r_{43} & r_{44} & r_{45} \\ r_{52} & r_{53} & r_{54} & r_{55} \end{vmatrix}} x_2 - \frac{1}{3}\frac{\begin{vmatrix} r_{21} & r_{22} & r_{24} & r_{25} \\ r_{31} & r_{32} & r_{34} & r_{35} \\ r_{41} & r_{42} & r_{44} & r_{45} \\ r_{51} & r_{52} & r_{54} & r_{55} \end{vmatrix}}{\begin{vmatrix} r_{22} & r_{23} & r_{24} & r_{25} \\ r_{32} & r_{33} & r_{34} & r_{35} \\ r_{42} & r_{43} & r_{44} & r_{45} \\ r_{52} & r_{53} & r_{54} & r_{55} \end{vmatrix}} x_3$$

$$+\tfrac{1}{4}\frac{\begin{vmatrix} r_{21} & r_{22} & r_{23} & r_{25} \\ r_{31} & r_{32} & r_{33} & r_{35} \\ r_{41} & r_{42} & r_{43} & r_{45} \\ r_{51} & r_{52} & r_{53} & r_{55} \end{vmatrix}}{\begin{vmatrix} r_{22} & r_{23} & r_{24} & r_{25} \\ r_{32} & r_{33} & r_{34} & r_{35} \\ r_{42} & r_{43} & r_{44} & r_{45} \\ r_{52} & r_{53} & r_{54} & r_{55} \end{vmatrix}} x_4 - \tfrac{1}{5}\frac{\begin{vmatrix} r_{21} & r_{22} & r_{23} & r_{24} \\ r_{31} & r_{32} & r_{33} & r_{34} \\ r_{41} & r_{42} & r_{43} & r_{44} \\ r_{51} & r_{52} & r_{53} & r_{54} \end{vmatrix}}{\begin{vmatrix} r_{22} & r_{23} & r_{24} & r_{25} \\ r_{32} & r_{33} & r_{34} & r_{35} \\ r_{42} & r_{43} & r_{44} & r_{45} \\ r_{52} & r_{53} & r_{54} & r_{55} \end{vmatrix}} x_5$$

$$= -9.3287\left[\frac{.0092}{.7532}\frac{x_2}{11.4434} + \frac{-.0460}{.7532}\frac{x_3}{.0452} + \frac{-.0030}{.7532}\frac{x_4}{10.2713} + \frac{-.2722}{.7532}\frac{x_5}{2.7367}\right]$$

$$= -.0016\,x_2 + .0253\,x_3 + .0036\,x_4 + 1.2318\,x_5\,.$$

The next question that is to be answered is how reliable are the expected values of the x_1's as determined from the regression equation when sets of values for x_2, x_3, x_4, and x_5 are known. The square of the multiple correlation coefficient when properly interpreted is the answer to this question.

This is equation III.27.7

$$r^2_{1.23\cdots n} = 1 - \left(\frac{R}{R_{11}}\right)$$

We first find R by substituting the values from Table III.5 for its determinant and solving.

$$R = \begin{vmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{21} & r_{22} & r_{23} & r_{24} & r_{25} \\ r_{31} & r_{32} & r_{33} & r_{34} & r_{35} \\ r_{41} & r_{42} & r_{43} & r_{44} & r_{45} \\ r_{51} & r_{52} & r_{53} & r_{54} & r_{55} \end{vmatrix} = .6774$$

Therefore, since $R_{11} = .7532$ as determined above,

$$r^2_{1.2345} = 1 - \left(\frac{R}{R_{11}}\right) = 1 - \frac{.6774}{.7532} = 1 - .8994 = .1006$$

Since this value, .1006 means that only 10.06 per cent of the variability in road tests is explained by the composite knowledge

of the factors, years of experience, reaction time, distance judgment, and driver information, it may be concluded that the composite result of these tests is practically worthless as a measure of driving ability as shown by the road test.

Another question to be answered is what is the standard error in the expected values of x. This standard error is a measure of the total variability that is not explained, or in other words, is not dependent upon the sets of values of x_2, x_3, x_4, and x_5.

The standard error in the expected value of x_1 obtained from the regression equation III.27.5 is equal to

$$\sigma^2_{1.2345} = \sigma^2_1 \left(\frac{R}{R_{11}}\right)$$

$$\sigma_{1.2345} = \sigma_1 \sqrt{\frac{R}{R_{11}}} = 9.3287 \sqrt{\frac{.6774}{.7532}}$$

$$= 0.8847$$

$$= 88.47 \text{ percent}$$

Since

$$\sigma_1 \left(\frac{R}{R_{11}}\right) + \sigma \left(1 - \frac{R}{R_{11}}\right) = \sigma^2_1 \frac{R}{R_{11}} + \sigma^2_1 - \sigma^2_1 \frac{R}{R_{11}} = \sigma^2_1$$

we may say that the proportional part of the total variability (σ^2_1) that is not explained in terms of x_2, x_3, x_4, and x_5 is $\frac{R}{R_{11}} = .8994$ = 89.94 per cent and that the explained variability

$$= 1 - \frac{R}{R_{11}} = 1 - .8994 = .1006 = 10.06 \text{ per cent.}$$

As a check:

$$\frac{R}{R_{11}} + \left(1 - \frac{R}{R_{11}}\right) = .8994 + .1006 = 1.$$

III. 28. *Partial Correlation*: Very often we wish the degree of correlation between two variables x_1 and x_2 when the other variables x_3, x_4, . . . , x_n have assigned values. Thus, we define a partial correlation coefficient $r_{12 \cdot 34 \ldots n}$ of x_1 and x_2 for assigned x_3, x_4, . . . , x_n as the

correlation coefficient of x_1 and x_2 in the part of the population for which $x_3, x_4, \ldots x_n$ have assigned values. A change in the assigned values may lead to the same or different values of $r_{12.34\ldots n}$.

Assume that the theoretical mean or expected values of x_1 and x_2 for an assigned $x_3, x_4, \ldots, x_n$ are

$$\left.\begin{aligned} x_1 &= b_{13}\,x_3 + b_{14}\,x_4 + \ldots + b_{1n}\,x_n \\ x_2 &= b_{23}\,x_3 + b_{24}\,x_4 + \ldots + b_{2n}\,x_n \end{aligned}\right\} \qquad \text{III.28.1.}$$

respectively.

Then, a partial correlation coefficient $r'_{12.34\ldots n}$ is the simple correlation coefficient of residuals

$$\left.\begin{aligned} x_{1.34\ldots n} &= x_1 - b_{13}\,x_3 - b_{14}\,x_4 - \ldots - b_{1n}\,x_n \\ x_{2.34\ldots n} &= x_2 - b_{23}\,x_3 - b_{24}\,x_4 - \ldots - b_{2n}\,x_n \end{aligned}\right\} \qquad \text{III.28.2.}$$

limited to the part of the population $n_{34\ldots n}$ of the total n for which $x_3, x_4, \ldots, x_n$ are fixed.

Suppose further that the population is such that any change in the assignment of values to $x_3, x_4, \ldots, x_n$ does not change the standard deviation of $x_{1.34\ldots n}$ nor of $x_{2.34\ldots n}$ nor the value of $r_{12.34\ldots n}$. Such a population suggests that we define

$$r_{12.34\ldots n} = \frac{x_{1.34\ldots n}\,x_{2.34\ldots n}}{n\sigma_{1.34\ldots n}\,\sigma_{2.34\ldots n}} \qquad \text{III.28.3.}$$

where the summation extends to n pairs of residuals, as the partial correlation coefficient of x_1 and x_2 for all sets of assignments of $x_3 \ldots, x_n$.

If the population is such that $r'_{12.34\ldots n}$ is not the same for each different set of assignments of $x_3, x_4, \ldots x_n$, the right hand member of III.28.3. may still be regarded as a sort of average value of correlation coefficients of x_1 and x_2 in subdivisions of a population obtained by assigning $x_3, x_4, \ldots, x_n$ or it may be regarded as the correlation coefficient between the deviations of x_1 and x_2 from the corresponding predicted values given by their linear equations on $x_3, x_4, \ldots x_n$. It can be shown that

$$r_{12.34\ldots n} = \frac{-R_{12}}{(R_{11}\,R_{22})^{\frac{1}{2}}} \qquad \text{III.28.4.}$$

To illustrate, we make use of the data for the Driver tests previously given in Table III.5 and set ourselves the problem of finding

the correlation between road test and years of experience under the assumption that each is influenced to some extent by reaction time, distance judgment and driver information. If each is thus influenced, the obtainment of the simple correlation coefficient between the road test and driver experience, assuming the existence of such influence, gives us spurious correlation. Partial correlation between road test and years of experience is the theory of correlation that removes the influence of reaction time, distance judgment, and driver information. Substituting the probable values of the R's for III.28.4, we find that

$$r_{12.34} = \frac{-R_{12}}{(R_{11} R_{22})^{\frac{1}{2}}}$$

Wherein R_{12} and R_{11} have the values already determined and R_{22} has the value .8960 found by substituting values from Table III.5. and solving the determinant.

$$R_{22} = \begin{vmatrix} r_{11} & r_{13} & r_{14} & r_{15} \\ r_{31} & r_{33} & r_{34} & r_{35} \\ r_{41} & r_{43} & r_{44} & r_{45} \\ r_{51} & r_{53} & r_{54} & r_{55} \end{vmatrix} = .8960$$

hence

$$r_{12.34} = \frac{-R_{12}}{(R_{11} R_{22})^{\frac{1}{2}}} = \frac{-.0092}{\sqrt{(.7532)(.8960)}} = \frac{-.0092}{\sqrt{.6749}} = \frac{-.0092}{.8215} = -0.001$$

therefore, there is practically no partial correlation.

III.29. *Regression (Trend) Lines*: Let

$$Y = a_0 + a_1 X + a_2 X^2 + \ldots + a_p X^p \qquad \text{III.29.1.}$$

be the equation of expected values of Y that are associated with the various values of X. It is desired to know the values of the a's such that the value of U given by

$$U = \sum_1^n (y_1 - a_0 - a_1 x_1 - a_2 x_1^2 - \ldots - a_p x_1^p)^2 \qquad \text{III.29.2.}$$

is a minimum.

This requires that

$$\frac{\partial U}{\partial a_j} = \sum_1^n (x_i^j y_i) - a_0 \sum_1^n x_i^j - a_1 \sum_1^n x_i^{j+1} - \ldots - a_p \sum_1^n x_i^{j+p} = 0 \qquad \text{III.29.3.}$$

whence

$$a_j = \frac{\Delta_j^{(p)}}{\Delta^{(p)}} \qquad \text{III.29.4.}$$

where

$$\Delta^{(p)} = \begin{vmatrix} \mu_0, & \mu_1, & \ldots, & \mu_p \\ \mu_1, & \mu_2, & \ldots, & \mu_{p+1} \\ . & . & & \\ . & . & & \\ . & . & & \\ . & . & & \\ \mu_p, & \mu_{p+1}, & \ldots, & \mu_{2p} \end{vmatrix} = \begin{vmatrix} n, & \Sigma f_x x, & \ldots. & \Sigma f_x x^p, & \Sigma f_x \bar{y}_x \\ \Sigma f_x x, & \Sigma f_x x^2, & \ldots. & \Sigma f_x x^{p+1}, & \Sigma f_x x \bar{y}_x \\ . & . & & & \\ . & . & & & \\ . & . & & & \\ . & . & & & \\ \Sigma f_x x^p, & \Sigma f_x x^{p+1}, & \ldots. & \Sigma f_x x^{2p}, & \Sigma f_x x^{2p} \bar{y}_x \end{vmatrix} \qquad \text{III.29.5.}$$

and $\Delta_j^{(p)}$ is the determinant obtained by substituting the product moments $\mu_{01}, \ldots \mu_{p1}$ for the $(j+1)$th column in $\Delta^{(p)}$.

It is not too difficult to show that the regression (trend) equation may be written in the form

$$\begin{vmatrix} Y, & 1, & X, & \ldots., & X^p \\ \mu_{01}, & \mu_0, & \mu_1, & \ldots., & \mu_p \\ \mu_{11}, & \mu_1, & \mu_2, & \ldots., & \mu_{p+1} \\ . & . & . & & \\ . & . & . & & \\ . & . & . & & \\ \mu_{p1}, & \mu_p, & \mu_{p+1}, & \ldots., & \mu_{2p} \end{vmatrix} = 0 \qquad \text{III.29.6.}$$

Now consider

$$Y = b_0 P_0 + b_1 P_1 + \ldots + b_p P_p$$

and demand that $\Sigma (P_j P_k) = 0$ when $j \neq k$, where the P's are polynomials in X, P_j being of degree j.

Again, minimizing

$$\sum_{\substack{x=x_1 \\ y=y_1}}^{\substack{x=x_n \\ y=y_n}} (Y - b_0 P_0 - b_1 P_1 - \ldots - b_p P_p)^2 \qquad \text{III.29.7.}$$

it is found that

$$\Sigma (yP_j) - b_0 \Sigma (P_0 P_j) \ldots - b_p \Sigma (P_p P_j) = 0 \qquad \text{III.29.8.}$$

Since $\Sigma (P_j P_k)$ for $j \neq k$ is zero, III.29.8. reduces to

$$\Sigma (y P_j) - b_j \Sigma (P_j^2) = 0. \qquad \text{III.29.9.}$$

Hence b_j is simply determined by P_j and if in fitting a curve of degree p, it is desired to proceed a step farther and add a term $b_{p+1} P_{p+1}$, the coefficients $b_0, \ldots, b_p$ already found remain unaltered. This method is known as the method of orthogonal polynomials.

The use of orthogonal polynomials gives a convenient method of determining step by step the goodness of fit of the regression line. Consider

$$\begin{aligned} U &= \Sigma (y - b_0 P_0 - \ldots - b_p P_p)^2 \\ &= \Sigma (y^2) - 2 b_0 \Sigma (y P_0) - \ldots - 2 b_p \Sigma (y P_p) \\ &\qquad + b_0^2 \Sigma (P_0^2) + \ldots + b_p^2 \Sigma (P_p^2) \end{aligned}$$

But, from III.29.9., we may express $\Sigma (y P_j)$ in terms of $\Sigma (P_j^2)$. Hence

$$U = \Sigma (y^2) - b_0^2 \Sigma (P_0^2) - \ldots - b_p^2 \Sigma (P_p^2) \qquad \text{III.29.10.}$$

This shows that the effect of any term $b_j P_j$ is to reduce U by $b_j^2 \Sigma (P_j^2)$ and the effect of this term on U is an independent matter. Again, if it is found that the addition of any term $b_j P_j$ does not reduce U significantly, the conclusion is that the term is redundant and therefore not necessary or that the fit is good enough.

It is now necessary to obtain the expressions for the various orthogonal polynomials. To this end, let

$$P_p = \sum_0^p{}_j C_{pj} X^j \qquad \text{III.29.11.}$$

In III.29.11., there are $(p + 1)$ unknown constants. Hence, in all the polynomials up to and including those of order p, there are $\frac{1}{2} (p + 1) (p + 2)$ constants. The orthogonal relations up to and

including order p provide $\frac{1}{2}\,p\,(p+1)$ conditions on the C's. It follows that $\frac{1}{2}\,(p+1)\,(p+2)-\frac{1}{2}\,p\,(p+1)=p+1$ constants are assignable at will. For convenience, take one constant for each P and assign it so that the coefficient of X^j in P_j has the value unity. In other words, put

$$C_{jj} = 1 \qquad \text{III.29.12.}$$

Rewriting III.29.11., we get

$$\begin{aligned}
P_0 &= 1 \\
P_1 &= C_{10} + X \\
P_2 &= C_{20} + C_{21}\,X + X^2 \\
P_3 &= C_{30} + C_{31}\,X + C_{32}\,X^2 + X^3 \qquad \text{III.29.13.} \\
&\cdots\cdots\cdots\cdots\cdots \\
P_p &= C_{p0} + C_{p1}\,X + C_{p2}\,X^2 + \ldots + X^p
\end{aligned}$$

From the orthogonal relations

$$\begin{aligned}
\Sigma\, P_p\, P_0 &= \Sigma\, P_p = 0 \\
\Sigma\, P_p\, P_1 &= 0 \qquad \text{III.29.14.} \\
&\cdots\cdots\cdots\cdots
\end{aligned}$$

This system, III.29.14., is equivalent to

$$\begin{aligned}
\Sigma\, P_p &= 0 \\
\Sigma\, x\, P_p &= 0 \\
\Sigma\, x^p\, P_p &= 0 \qquad \text{III.29.15.}
\end{aligned}$$

Substituting the values of the P's from III.29.13., it is found that

$$\begin{aligned}
&C_{p0}\,\mu_0 + C_{p1}\,\mu_1 + \ldots + C_{p,\,p-1}\,\mu_{p-1} + \mu_p = 0 \\
&C_{p0}\,\mu_1 + C_{p1}\,\mu_2 + \ldots + C_{p,\,p-1}\,\mu_p + \mu_{p+1} = 0 \qquad \text{III.29.16.} \\
&\cdots\cdots\cdots\cdots\cdots \\
&C_{p0}\,\mu_{p-1} + C_{p1}\,\mu_p + \ldots + C_{p,\,p-1}\,\mu_{2\,p-2} + \mu_{2\,p-1} = 0
\end{aligned}$$

From these equations,

$$C_{pj} = \frac{\Delta^{(p)}_{pj}}{\Delta^{(p-1)}} \qquad \text{III.29.17.}$$

where $\Delta^{(p-1)}$ has the same meaning as before and $\Delta^{(p)}_{pj}$ is the minor of the term in the last row and $(j+1)$th column of $\Delta^{(p)}$. It follows that

$$P_p = \frac{1}{\Delta^{(p-1)}} \begin{vmatrix} \mu_0, & \mu_1, & \cdots\cdots, & \mu_p \\ \mu_1, & \mu_2, & \cdots\cdots, & \mu_{p+1} \\ \mu_{p-1}, & \mu_p, & \cdots\cdots, & \mu_{2p-1} \\ 1, & X, & \cdots\cdots, & X^p \end{vmatrix} \qquad \text{III.29.18}$$

It is clear, because of diagonal symmetry of $\Delta^{(p)}$ that $C_{jk} = C_{kj}$.

III.29.19.

From III.29.15.

$$\Sigma (P_p^2) = \Sigma (x^p P_p)$$

and hence from III.29.18. if we multiply the last row and sum

$$\Sigma (P_p^2) = \frac{n \Delta^{(p)}}{\Delta^{(p-1)}} \qquad \text{III.29.20.}$$

Likewise

$$\Sigma (y P_p) = \frac{n \Delta p^{(p)}}{\Delta^{(p-1)}} \qquad \text{III.29.21.}$$

Finally, from III.29.9.

$$b_p = \frac{\Delta p^{(p)}}{\Delta^{(p)}} \qquad \text{III.29.22.}$$

and the problem is completed.

Specifically, if $\mu_0 = 1$, $\mu_1 = 0$, $\mu_2 = 1$, then

$$P_0 = 1$$

$$P_1 = \frac{\begin{vmatrix} 1 & 0 \\ 1 & x \end{vmatrix}}{1} = X \qquad \text{III.29.23.}$$

$$P_2 = \frac{\begin{vmatrix} 1 & 0 & 1 \\ 0 & 1 & \mu_3 \\ 1 & X & X^2 \end{vmatrix}}{\begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}} = X^2 - \mu_3 X - 1$$

$$P_3 = \frac{\begin{vmatrix} 1 & 0 & 1 & \mu_3 \\ 0 & 1 & \mu_3 & \mu_4 \\ 1 & \mu_3 & \mu_4 & \mu_5 \\ 1 & X & X^2 & X^3 \end{vmatrix}}{\begin{vmatrix} 1 & 0 & 1 \\ 0 & 1 & \mu_3 \\ 1 & \mu_3 & \mu_4 \end{vmatrix}}$$

$$= \frac{1}{\mu_4 - \mu_3^2 - 1}\left\{ (\mu_4 - \mu_3^2 - 1)\,X^3 - (\mu_5 - \mu_4\,\mu_3 - \mu_3)\,X^2 + (\mu_3\,\mu_5 - \mu_4^2 + \mu_4 - \mu_3^2)\,X + (\mu_5 - 2\,\mu_4\,\mu_3 + \mu_3)\right\}$$

To illustrate: From Table III. 4. the regression data are obtained and placed in the first three colums of Table III. 6.

Table III. 6.

(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\bar{Y}_x$	X	f_x	$f_x\bar{Y}_x$	$f_xX\bar{Y}_x$	f_xX	f_xX^2
23.1	1	55	1270.5	1270.5	55	55
27.0	3	75	2025.0	6075.0	225	675
30.6	5	74	2264.4	11322.0	370	1850
30.7	7	70	2149.0	15043.0	490	3430
39.7	9	63	2501.1	22509.9	567	5103
35.8	11	35	1253.0	13783.0	385	4235
38.4	13	50	1920.0	24960.0	650	8450
40.6	15	33	1339.8	20097.0	495	7425
47.1	17	41	1931.1	32828.7	2009	11849
44.9	19	37	1661.3	31564.7	703	13357
47.8	21	51	2437.8	51193.8	1071	22491
55.4	23	63	3490.2	80274.6	1449	33327
54.7	25	81	4430.7	110767.5	2025	50625
51.0	27	45	2295.0	61965.0	1215	32805
51.9	29	133	6902.7	200178.3	3857	111853
55.4	31	93	5152.2	159718.2	2883	89373
58.4	33	109	6365.6	210064.8	3597	118701
55.9	35	86	4807.4	168259.0	3010	105350
59.5	37	46	2737.0	101269.0	1702	62974
61.0	39	49	2989.0	116571.0	1911	74529
53.3	41	16	852.8	34964.8	656	26896
79.1	43	11	870.1	37414.3	473	20339
60.9	45	8	487.2	21924.0	360	16200
68.5	47	6	411.0	19317.0	282	13254
45.8	49	3	137.4	6732.6	147	7203
48.5	51	2	97.0	4947.0	102	5202
36.5	53	1	36.5	1934.5	53	2809
			62814.8	1566949.2	29430	850360

To obtain the various regression (trend) functions for the data of Table III.4., it is necessary to compute the following values, the obtainment of the first four being shown in columns (4), (5), (6), (7) of Table III.6.:

$$\begin{array}{llll} \Sigma f_x \bar{Y}_x = 62814.8 & & \Sigma f_x X^4 = 917057464 \\ \Sigma f_x X \bar{Y}_x = 1566949.2 & & \Sigma f_x X^5 = 32132903385 \\ \Sigma f_x X = 29430 & & \Sigma f_x X^6 = 1180837278435 \\ \Sigma f_x X^2 = 850360 & & \Sigma f_x X^2 \bar{Y}_x = 47175422.8 \\ \Sigma f_x \bar{Y}_x^2 = 2867513.03 & & \Sigma f_x X \bar{Y}^3{}_x = 1535815847.1 \\ \Sigma f_x X^3 = 27146214 \end{array}$$

First, it is necessary to compute the value of the b_j's from III.29.22. These are found to be as follows:

$$b_0 = \frac{\Delta_0^{(0)}}{\Delta^{(0)}} = \frac{|\mu_{01}|}{|\mu_0|} = \frac{|\Sigma f_x \bar{Y}_x|}{|n|} = \frac{62814.8}{1336} = 47.017 \qquad \text{III.29.24.}$$

$$b_1 = \frac{\Delta_1^{(1)}}{\Delta^{(1)}} = \frac{\begin{vmatrix} \mu_0 & \mu_{01} \\ \mu_1 & \mu_{11} \end{vmatrix}}{\begin{vmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{vmatrix}} = \frac{\begin{vmatrix} n & \Sigma f_x \bar{Y}_x \\ \Sigma f_x X & \Sigma f_x X \bar{Y}_x \end{vmatrix}}{\begin{vmatrix} n & \Sigma f_x X \\ \Sigma f_x X & \Sigma f_x X^2 \end{vmatrix}}$$

$$= \frac{\begin{vmatrix} 1336, & 62814.8 \\ 29430, & 1566949.2 \end{vmatrix}}{\begin{vmatrix} 1336, & 29430 \\ 29430, & 850360 \end{vmatrix}} = \frac{(1336)\,(1566949.2) - (29430)\,(62814.8)}{(1336)\,(850360) \quad - (29430)\,(29430)}$$

$$= \frac{244804567.2}{269956060} = 0.909 \qquad \text{III.29.25.}$$

$$b_2 = \frac{\Delta_2^{(2)}}{\Delta^{(2)}} = \frac{\begin{vmatrix} \mu_0 & \mu_1 & \mu_{01} \\ \mu_1 & \mu_2 & \mu_{11} \\ \mu_2 & \mu_3 & \mu_{21} \end{vmatrix}}{\begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{vmatrix}} = \frac{\begin{vmatrix} n & \Sigma f_x X & \Sigma f_x \bar{Y}_x \\ \Sigma f_x X & \Sigma f_x X^2 & \Sigma f_x X \bar{Y}_x \\ \Sigma f_x X^2 & \Sigma f_x X^3 & \Sigma f_x X^2 \bar{Y}_x \end{vmatrix}}{\begin{vmatrix} n & \Sigma f_x X & \Sigma f_x X^2 \\ \Sigma f_x X & \Sigma f_x X^2 & \Sigma f_x X^3 \\ \Sigma f_x X^2 & \Sigma f_x X^3 & \Sigma f_x X^4 \end{vmatrix}}$$

$$= \frac{\begin{vmatrix} 1336, & 29430, & 62815 \\ 29430, & 850360, & 1566949 \\ 850360, & 27146214, & 47175423 \end{vmatrix}}{\begin{vmatrix} 1336, & 29430, & 850360 \\ 29430, & 850360, & 27146214 \\ 850360, & 27146214, & 917057464 \end{vmatrix}}$$

$$= \frac{1336 \begin{vmatrix} 850360, & 1566949 \\ 27146214, & 47175423 \end{vmatrix} - 29430 \begin{vmatrix} 29430, & 1566949 \\ 850360, & 47175423 \end{vmatrix}}{1336 \begin{vmatrix} 850360, & 27146214 \\ 27146214, & 917057464 \end{vmatrix} - 29430 \begin{vmatrix} 29430, & 27146214 \\ 850360, & 917057464 \end{vmatrix}}$$

$$+ 62815 \begin{vmatrix} 29430, & 850360 \\ 850360, & 27146214 \end{vmatrix}$$

$$+ 850360 \begin{vmatrix} 29430, & 850360 \\ 850360, & 27146214 \end{vmatrix}$$

$$= \frac{(1336)(-24206)(10^8) - (29430)(55901)(10^6)}{(1336)(42912) \quad (10^9) - (29430)(39049)(10^8)}$$

$$+ (62815) \quad (75801)(10^6)$$

$$+ (850360)(75801)(10^6)$$

$$= -\frac{1176482}{68673633} = -0.01713 \qquad \text{III.29.26.}$$

$$b_3 = \frac{\Delta_3^{(3)}}{\Delta^{(3)}} = \frac{\begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \mu_{01} \\ \mu_1 & \mu_2 & \mu_3 & \mu_{11} \\ \mu_2 & \mu_3 & \mu_4 & \mu_{21} \\ \mu_3 & \mu_4 & \mu_5 & \mu_{31} \end{vmatrix}}{\begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \mu_3 \\ \mu_1 & \mu_2 & \mu_3 & \mu_4 \\ \mu_2 & \mu_3 & \mu_4 & \mu_5 \\ \mu_3 & \mu_4 & \mu_5 & \mu_6 \end{vmatrix}} = \frac{\begin{vmatrix} n & \Sigma f_x X & \Sigma f_x X^2 & \Sigma f_x \bar{Y}_x \\ \Sigma f_x X & \Sigma f_x X^2 & \Sigma f_x X^3 & \Sigma f_x X \bar{Y}_x \\ \Sigma f_x X^2 & \Sigma f_x X^3 & \Sigma f_x X^4 & \Sigma f_x X^2 \bar{Y}_x \\ \Sigma f_x X^3 & \Sigma f_x X^4 & \Sigma f_x X^5 & \Sigma f_x X^3 \bar{Y}_x \end{vmatrix}}{\begin{vmatrix} n & \Sigma f_x X & \Sigma f_x X^2 & \Sigma f_x X^3 \\ \Sigma f_x X & \Sigma f_x X^2 & \Sigma f_x X^3 & \Sigma f_x X^4 \\ \Sigma f_x X^2 & \Sigma f_x X^3 & \Sigma f_x X^4 & \Sigma f_x X^5 \\ \Sigma f_x X^3 & \Sigma f_x X^4 & \Sigma f_x X^5 & \Sigma f_x X^6 \end{vmatrix}}$$

III.29.27.

Note: To evaluate determinants, the reader is referred to "A Textbook of Determinants, Matrices, and Algebraic Forms," by W. L. Ferrar, Oxford University Press, 1941.

Next, it is necessary to obtain the various orthogonal polynomials. They are

$$P_1 = \frac{\begin{vmatrix} n & \Sigma f_x X \\ 1 & X \end{vmatrix}}{n} = \frac{\begin{vmatrix} 1336 & 29430 \\ 1 & X \end{vmatrix}}{|1336|}$$

$$= \frac{1336\,X - 29430}{1336} = X - 22.03 \qquad \text{III.29.28.}$$

$$P_2 = \frac{\begin{vmatrix} n & \Sigma f_x X & \Sigma f_x X^2 \\ \Sigma f_x X & \Sigma f_x X^2 & \Sigma f_x X^3 \\ 1 & X & X^2 \end{vmatrix}}{\begin{vmatrix} n & \Sigma f_x X \\ \Sigma f_x X & \Sigma f_x X^2 \end{vmatrix}}$$

$$= 1\,\frac{\begin{vmatrix} \Sigma f_x X & \Sigma f_x X^2 \\ \Sigma f_x X^2 & \Sigma f_x X^3 \end{vmatrix} - X \begin{vmatrix} n & \Sigma f_x X^2 \\ \Sigma f_x X & \Sigma f_x X^3 \end{vmatrix} + X^2 \begin{vmatrix} n & \Sigma f_x X \\ \Sigma f_x X & \Sigma f_x X^2 \end{vmatrix}}{\begin{vmatrix} n & \Sigma f_x X \\ \Sigma f_x X & \Sigma f_x X^2 \end{vmatrix}}$$

$$= 1\,\frac{\begin{vmatrix} 29430 & 850360 \\ 850360 & 27146214 \end{vmatrix} - X \begin{vmatrix} 1336 & 850360 \\ 29430 & 27146214 \end{vmatrix} + X^2 \begin{vmatrix} 1336 & 29430 \\ 29430 & 850360 \end{vmatrix}}{\begin{vmatrix} 1336 & 29430 \\ 29430 & 850360 \end{vmatrix}}$$

$$= \frac{75800948420 - 11241247104\,X + 269956060\,X^2}{269956060}$$

$$= 280.7899 - 41.6410\,X + X^2 \qquad \text{III.29.29.}$$

The linear regression (trend) function is

$$\bar{Y}_{x'} = b_0 + b_1 P_1 = b_0 + b_1 (X - 22.03)$$

$$= 47.017 + 0.909\,(X - 22.03)$$

$$= 26.99 + 0.909\,X \qquad \text{III.29.30.}$$

which agrees with result obtained in III.23.2., p. 115 as it should.

The quadratic regression (trend) function is

$$\overline{Y}_x = b_0 + b_1 P_1 + b_2 P_2$$
$$= 471017 + 0.909\,(X - 22.03) - 0.01713\,(280.7899 - 41.6410\,X + X^2)$$
$$= 22.18 + 1.622\,X - 0.01713\,X^2 \qquad \text{III.29.31}$$

Likewise

$$P_3 = \frac{\begin{vmatrix} n & \Sigma f_x X & \Sigma f_x X^2 & \Sigma f_x X^3 \\ \Sigma f_x X & \Sigma f_x X^2 & \Sigma f_x X^3 & \Sigma f_x X^4 \\ \Sigma f_x X^2 & \Sigma f_x X^3 & \Sigma f_x X^4 & \Sigma f_x X^5 \\ 1 & X & X^2 & X^3 \end{vmatrix}}{\begin{vmatrix} n & \Sigma f_x X & \Sigma f_x X^2 \\ \Sigma f_x X & \Sigma f_x X^2 & \Sigma f_x X^3 \\ \Sigma f_x X^2 & \Sigma f_x X^3 & \Sigma f_x X^4 \end{vmatrix}} \qquad \text{III.29.32.}$$

Since the effect of adding the second degree term is rather small, it follows that the addition of the third degree term is negligible and redundant. In III.29.30. and III.29.31., $\overline{Y}_x$ is the probable or expected minimum spacing for a particular speed X.

Suppose $X = 10$ miles per hour, then from III.29.30. we find that the expected minimum spacing in feet is $\overline{Y}_x = \overline{Y}_{10} = 36.08$ feet, and from III.29.31., we find $\overline{Y}_x = \overline{Y}_{10} = 36.69$ feet.

Again, if $X = 30$ miles per hour, III.29.30. gives $\overline{Y}_{30} = 54.26$ feet and III.29.31. gives $\overline{Y}_{30} = 55.42$ feet.

If $X = 50$ miles per hour, III.29.30. gives $\overline{Y}_{50} = 72.44$ feet and III.29.31. gives 60.45 feet.

It is to be emphasized that because of the scarcity of data beyond a speed of 40 miles per hour, it is not possible or scientifically sound to use the regression functions to predict the minimum spacing beyond that speed. In any event, however, the use of the quadratic function, III.29.31., gives the better estimate of the minimum spacing in so far as we are able to use either theory. For the lower speeds, III.29.30. gives an underestimate and for the higher speeds an overestimate.

It also appears very likely that the actual minimum spacing is not expressible in terms of a single regression function. In other words, it appears that there may be one regression function for lower speeds and a different one for higher speeds.

REFERENCES, CHAPTER III

[1] Uspensky, J. V., "*Introduction to Mathematical Probability*," First Edition, McGraw-Hill Book Co., 1937, page 101.

[2] Ibid., page 101.

[3] Ibid., page 204.

Tchebycheff, P. S., "*Des Valeurs Moyennes*," Journal de Mathematique (2), Volume 12 (1867), pages 174–184.

Bienayme, M., "*Considerations a l'appui de la decouverte de Laplace sur la loi de probabilite dans la methode des moindres carres*," Comptes Rendus, Vol. 37 (1853), pages 309–24.

[4] Zoch, R., "*On the Postulate of the Arithmetic Mean*," Annals of Mathematical Statistics, Vol. VI., No. 4, December 1935, pages 171–187.

Zoch, R., "*Invariants and Covariants of Certain Frequency Curves*," Annals of Mathematical Statistics, Vol. VI, No. 1, March 1935, pages 124–135.

[5] Weida, F. M., "*Maximum Term of Hypergeometric Series*," American Mathematical Monthly, Vol. XXXIII, No. 6, June–July 1926, page 339.

[6] Weida, F. M., "*On Various Conceptions of Correlation*," Annals of Mathematics, Second Series, Vol. 29, No. 3, July 1928, pages 276–312.

Rietz, H. L., "*Mathematical Statistics*," Open Court Publishing Co. 1927, pages 77–113.

Rietz, H. L., "*Handbook* of *Mathematical Statistics*," Houghton-Mifflin Co., 1924, pages 120–165.

[7] Saculy, M., "*Trend Analysis of Statistics*," Brookings Institution, 1934, pages 33–37.

Kendall, M. G., "*The Advanced Theory of Statistics*," Charles Griffin and Co. Ltd., London, 1946, pages 145–152.

[8] Rietz, H. L., "*Mathematical Statistics*," Open Court Publishing Co., 1927, pages 31–38.

[9] Fry, Thornton G., "*Probability and Its Engineering Uses*," D. Van Nostrand Co., New York, 1928.

[10] Molina, E. C., "*Poissons's Exponential Binomial Limits*," D. Van Nostrand Co., New York, 1942.

[11] Elderton, W. F., "*Frequency Curves and Correlation*," C. and E. Layton, London, 1927.

CHAPTER IV

SAMPLING THEORY

Reliability and Significance

IV. 1. *Objective.* In this chapter it is proposed to show how to use the mathematical models of distribution that were developed in Chapter III as a basis for making inferences from a limited number of happenings that will apply to all such happenings. This process of reasoning from the particular to the general is known as *inductive inference* and in a broader sense is called *sampling theory.*

Inductive inference is a means by which scientific progress comes about. The research worker obtains data through planned experiments or through the observation of natural happenings such as the occurrence of accidents at certain types of highway intersections. From the data obtained he infers that certain things are so. But it is well known that exact inductive inference is theoretically impossible. One of the functions of statistics is to provide techniques for making inferences and for measuring the degree of certainty of the inferences.

In order to make the idea of inference somewhat more concrete, let us suppose that we have observed the speeds of one hundred vehicles at a given location and have found that five were traveling over seventy miles per hour. We might estimate from this sample that five per cent of all vehicles travel over seventy miles per hour, but we would not be very sure as to the correctness of our estimate for we know that a different sample of this limited size would undoubtedly lead to a different estimate. At best the sample contains but partial information about the law of behavior of the total population of drivers. Population is used in its statistical sense meaning a collection of results or objects. Summary numbers calculated from the sample accurately characterize the sample, but the important question is, how good are these same summary numbers when used as estimates of the characteristics of the population? What is the error committed by the use of

sample characterizing numbers in place of the associated population characterizing numbers?

The role of statistics in providing a measure of the uncertainty of inferences from samples is confined to sampling errors. It must be assumed that the experimenter has guarded against accidents in recording the data. In gathering data the first consideration is the obtaining of a random sample.

IV. 2. *Random Sampling*: In order to demonstrate what is meant by random sampling let us suppose that we have a given population and that the attribute or attributes of the population to be measured are specified. The problem is to find a sampling method for the given population and the stochastic variable being measured that will yield a random or unbiased sample. The answer lies partly in theory and partly in techniques that have been proven in practice or may have to be devised to meet a given situation.

The first requirement is that there be no obvious connection between the method of selection and the properties being studied. The method and the properties must be independent in so far as our prior knowledge enables us to make them so.

To meet the second requirement that the sample be a random selection, we rely on our previous experience with a given method as well as our intuition to justify its use on new occasions. A very reliable method of drawing random samples consists of constructing a model of the population and sampling from the model.

Actually, randomness is largely a matter of intuition. The theory of probability considers the set of all possible different samples that may be drawn from a specified universe and enables us to derive their distribution law for any desired characterizing summary number. This theory requires that it be made certain that the sampling method will tend to yield all possible different samples with equal frequency. A method that does this is called a *random* method.

IV. 3. *Distribution of Sample Arithmetic Means*. For the purpose of illustrating the law of the distribution of sample arithmetic means, let us suppose that we have a normal universe, and that from this universe, we draw a large number of samples all of the same size,

n. If the samples are random and drawn independently, then the distribution of sample arithmetic means is also normal. Furthermore, the arithmetic mean of the distribution of sample arithmetic means is the true arithmetic mean of the universe and the standard deviation of the distribution of sample arithmetic means is the standard deviation of the universe divided by the square root of the size of the sample. Expressed symbolically: If $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, ..., $\bar{X}_1$, ... $\bar{X}_k$ are the sample arithmetic means and if $\bar{\bar{X}}$ is the arithmetic mean of the universe from which the samples were drawn, then

$$\bar{\bar{X}} = \frac{\Sigma \bar{X}_1}{k}. \qquad \text{IV.3.1.}$$

If σ is the standard deviation of the universe of measures and $s_{\bar{x}}$ is the standard deviation of the distribution of sample arithmetic means, then

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \qquad \text{IV.3.2.}$$

The value $s_{\bar{x}}$ is frequently called the standard error of the arithmetic mean. Actually it is the *measure of reliability* of the arithmetic mean and is in fact the expected error committed when a particular sample arithmetic mean is used in place of the true arithmetic mean of the universe. The smaller the expected error, the more reliable or the more precise is the sample arithmetic mean.

The measure of reliability given by IV.3.2. is exact in theory but not usable in practice because the value of σ depends upon the population which is not known. Consequently it is necessary to obtain from the sample an *unbiased estimate* of the universe variance σ^2, indicated by the symbol $\hat{\sigma}^2$. This is equal to:

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2 \qquad \text{IV.3.3.}$$

where s^2 is the variance of the sample. Substituting this value $\hat{\sigma}^2$ for σ^2 in IV.3.2., we obtain

$$s_{\bar{x}} = \frac{s}{\sqrt{n-1}} \qquad \text{IV.3.4.}$$

which is usable as the standard error of the arithmetic mean.

It is to be noted that IV.3.3. gives an estimate of universe variance.

Using the data of Table II.1. it was found that the arithmetic mean was 38.2 miles per hour and the standard deviation, 8.9 miles per hour. In II.22., page 50, it was also found that the expected speed of 38.2 miles per hour was in error at most 23.3 per cent with a measure of confidence of 71 per cent. To find out how near the true value of the arithmetic mean our sample mean is, we substitute in IV.3.4. and find that

$$s_{\bar{x}} = \frac{s}{\sqrt{n-1}} = \frac{8.9}{\sqrt{299}} = 0.52 \text{ miles per hour.} \qquad \text{IV.3.5.}$$

which is the expected error in the sample arithmetic mean. In other words, it is 68.27 per cent certain that the true arithmetic mean in the universe has a value between $38.2 - 0.5 = 37.7$ and $38.2 + 0.5 = 38.7$ miles per hour. (68.27 is the per cent of area contained within one standard deviation on each side of the mean). In this case the maximum expected relative error is $0.52/38.7 = 1.3$ per cent with 68.27 per cent certainty. In like manner it is 95.45 per cent certain that the maximum relative error does not exceed 2.6 per cent and similarly it is 99.73 per cent certain that the error does not exceed 3.9 per cent. The conclusion then is that the sample arithmetic mean is fairly reliable (precise) but as found before, it is not usable as a typical or characterizing speed.

IV. 4. *Inference Concerning Population Mean.* Let μ be the population mean and $\bar{X}$ the sample mean. It is desired to test the hypothesis: The sample whose mean is $\bar{X}$ could have come from a population with mean μ. If this is so, how certain are we that it did ? This question is answered by using the t-distribution where in this case

$$t = \frac{|\bar{X} - \mu|}{s_{\bar{x}}} \qquad \text{IV.4.1.}$$

For example: Could our sample with arithmetic mean of 38.2 miles per hour have come from a population whose arithmetic

mean is 40 miles per hour? Substituting the values already found in IV.4.1., we have

$$t = \frac{|38.2 - 40.0|}{0.52} = 1.54$$

Making use of the t-table in "Statistical Methods for Research Workers"[5] with in *this* case $n - 1 = 299$ degrees of freedom it is found that 5 per cent of the time the difference as expressed by t would be at least 1.97. Only one degree of freedom is lost because the only restriction is that the deviations are taken from the mean of the sample. However, our value of $t = 1.54$ is less than 1.97. Hence it is concluded that on the 5 per cent level of significance we have insufficient grounds to reject the hypothesis. In other words, if the hypothesis is rejected, it would be rejected when it is true slightly more than 5 per cent of the time. This means that we would have a slightly greater than 5 per cent risk in rejecting the hypothesis. To put it in another way the odds are a bit less than 95 to a bit more than 5 per cent in favor of rejection of the hypothesis. The level of significance and risk are synonymous, for the level of significance is the probability that the hypothesis is true and its complement is the probability that the hypothesis is not true.

IV. 5. *Confidence Limits.* Since it is impossible to estimate or predict the true value exactly it is necessary to obtain two numbers between which the true value will fall. These two numbers are known as *confidence limits.* To obtain them, it is necessary first to determine the value of t associated with the relevant degrees of freedom (number of possible values variable assumes minus number of rigorous conditions or constraints the values must obey) and a desirable probability level of significance.

The sample arithmetic mean may be greater or less than the population arithmetic mean. From IV.4.1, it was found that

$$t = \frac{|\bar{X} - \mu|}{s_{\bar{x}}}$$

It is not hard to see from this equation that $\pm t = (\bar{X} - \mu)/s_{\bar{x}}$, or

$$\mu = \bar{X} \pm ts_{\bar{x}} \qquad \text{IV.5.1.}$$

which gives the two values (confidence limits) between which the true sample arithmetic mean will fall. These values are based upon the specific degrees of freedom and level of significance as demanded by the subjective problem. The limit of significance and the degree of reliability may be of any desired value.

To illustrate: Suppose we have a sample whose arithmetic mean is 52, whose standard deviation is 5 and whose size is 101. It is desired to find the confidence limits on a 5 per cent level.

Making use of the t-table with $(n - 1) = 100$ degrees of freedom and IV.5.1., it is found that

$$\mu = 52 \pm 1.98\left(\frac{5}{10}\right)$$
$$= 52 \pm 0.99$$

whence the two values of μ are 51.01 and 52.99.

This means that it is 95 per cent certain that the true arithmetic mean of the universe lies between 51.01 and 52.99. Again, it is 95 per cent certain that if we take the arithmetic mean of 52 as the value of the population (true) arithmetic mean the error committed will not exceed $0.99/52 = .019 = 1.90$ per cent. If the error that may be tolerated (which is obtained from the subjective material) is not less than 1.90 per cent, then for the pertinent purpose the sample arithmetic mean may be used as the population arithmetic mean. Otherwise, it may not be used.

IV. 6. *Difference Between Sample Arithmetic Means.* Frequently the arithmetic means are computed from two independent samples. The question that needs to be answered is: Are these samples independent and from the same normal universe? To answer this question we again make use of the t-distribution, but in this case we use for t the value t′ given by

$$t' = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{(N_1 + N_2)(N_1 S_1^2 + N_2 S_2^2)}{(N_1 N_2)(N_1 + N_2 - 2)}}} \qquad \text{IV.6.1.}$$

where

$\bar{X}_1$ is the arithmetic mean of the first sample

$\bar{X}_2$ is the arithmetic mean of the second sample

S_1^2 is the variance of the first sample

S_2^2 is the variance of the second sample

N_1 is the size of the first sample

N_2 is the size of the second sample

$N_1 + N_2 - 2$ are the degrees of freedom and

$\sqrt{\dfrac{(N_1 + N_2)\,(N_1 S_1^2 + N_2 S_2^2)}{(N_1 N_2)\,(N_1 + N_2 - 2)}}$ is the standard deviation of the distribution of differences between independent sample arithmetic means from the same normal universe.

To illustrate: Suppose we have the following two samples:

	Sample I	*Sample II*
Arithmetic mean	$\bar{X}_1 = 145$	$\bar{X}_2 = 150$
Standard Deviation	$S_1 = 5$	$S_2 = 6$
Number of Individuals	$N_1 = 12$	$N_2 = 20$

We wish to test the hypothesis: The difference between the sample arithmetic means is insignificant, therefore, these two samples are independent and from the same normal universe.

To make the test we use IV.6.1. Substituting the given values in IV.6.1., it is found that in numerical value

$$t' = \frac{|145 - 150|}{\sqrt{\dfrac{32\,[12\,(25) + 20\,(36)]}{240\,(30)}}} = \frac{5}{\sqrt{4.53}} = \frac{5}{2.13} = 2.35$$

Making use of the t-table with $(N_1 + N_2 - 2) = (12 + 20 - 2) = 30$ degrees of freedom it is found that when $t = 2.042$ the probability that the two samples came from the same normal universe is 0.05 and when $t = 2.750$ the probability is 0.01. The value of $t = 2.35$ lies between the 5 per cent and 1 per cent levels of significance, hence, we conclude that the two sample arithmetic means are significantly different on the 5 per cent level but not so on the 1 per cent level. This means that the odds are between 95 and

99 to between 5 and 1 in favor of rejecting the hypothesis that the two samples came from the same normal universe.

It is important to note that if the two means had not been significantly different it would have been necessary to investigate the significance of the difference between the variances. The method of doing this will be shown later.

If the variances or the means, or both, are significantly different, we have grounds to reject the hypothesis; but if the variances and means each are not significantly different, we do not have grounds to reject the hypothesis. This is true because the normal distribution is a two-parameter family of curves.

IV.7. *Size of Sample for Arithmetic Mean.* Suppose we require, within a specified degree of certainty, that the sample arithmetic mean shall differ from true mean by not more than a given ε.

Consider again

$$t = \frac{\bar{X} - \mu}{s_{\bar{x}}} \qquad \text{IV.7.1.}$$

Since the error is ε, it follows that $\bar{X} - \mu = \varepsilon$. Hence IV.7.1. becomes

$$t = \frac{\varepsilon}{s_{\bar{x}}} = \varepsilon \Big/ \frac{s}{\sqrt{N-1}} \qquad \text{IV.7.2.}$$

Rewriting IV.7.2., we obtain

$$\frac{N-1}{t^2} = \frac{s^2}{\varepsilon^2}. \qquad \text{IV.7.3.}$$

Suppose we wish to know the size of the sample such that it is 95 per cent certain that the sample mean is within 2 units of the true mean of the universe. In this case, if the variance of the sample is 100, $s^2 = 100$, $\varepsilon^2 = 4$ and from IV.7.3.,

$$\frac{N-1}{t^2} = \frac{100}{4} = 25$$

From the t-table, it is found that when $N = 101$, $\frac{N-1}{t^2}$ $= 25.508$ and when $N = 91$, $\frac{N-1}{t^2} = 22.727$. Hence, the size of the sample is 101.

IV.8. *Reliability of Sample Standard Deviation.* The test for the reliability of a sample standard deviation is defined as χ^2 (Chi-square) and is

$$\chi^2 = \frac{NS^2}{\sigma^2} \qquad \text{IV.8.1.}$$

where N is the size of the sample, S^2 is the sample variance and σ^2 is the population variance. Thus χ^2 is the sum of the squares of N—1 independent normal deviates divided by their common variance.

This criterion is useful for comparing a sample variance with a population variance.

To illustrate: Take a sample of size 10 whose variance is 25, could this sample have come from a universe whose variance is 16 ?

Using IV.8.1., it is found that

$$\chi^2 = \frac{10\ (25)}{16} = \frac{250}{16} = 15.63$$

From a χ^2 table for $(N - 1) = 9$ degrees of freedom, it is found that the probability of $\chi^2 > 14.684$ is 0.10 and the probability of $\chi^2 > 16.919$ is 0.05.

It follows that a population (universe) having a variance of 16 could yield a sample with variance of 25 or more between 5 and 10 times out of 100.

Sometimes it is desirable to obtain from the sample an unbiased estimate of the true universe variance. This is accomplished by using

$$\sigma^2 = \frac{N}{N-1} S^2 \qquad \text{IV.8.2.}$$

which in this case becomes

$$\sigma^2 = \frac{10}{9} 25 = 27.8$$

which means that the expected value of the universe variance is 27.8 when the sample variance is 25 and the size of the sample is 10.

IV. 9. *Significance of Difference Between Sample Variances.* The test here is to determine, with respect to variance, whether two samples are independent and from the sample normal universe. The criterion is the F-test which is given by

$$F = \frac{S_1'^2}{S_2'^2} \qquad \text{IV.9.1.}$$

where $S'^2 = \frac{N_1 S_1^2}{N_1 - 1}$ and $S_2'^2 = \frac{N_2 S_2^2}{N_2 - 1}$ and the degrees of freedom for S_1^2 is $N_1 - 1$ and for S_2^2 is $N_2 - 1$. Having two unbiased estimates of variance, always use for $S_1'^2$ the greater of the two variances.

To illustrate: Let there be given two samples of 10 and 12 individuals respectively. Let their variances be 10 and 5 respectively. Are these two samples independent and from the same normal universe? In other words, is the variance 10 significantly greater than the variance 5?

Substituting in IV.9.1., it is found that F becomes

$$F = \frac{N_1 S_1^2}{N_1 - 1} \Big/ \frac{N_2 S_2^2}{N_2 - 1} = \frac{10\,(10)}{9} \Big/ \frac{12\,(5)}{11}$$
$$= 2.04$$

From the F-table with $n_1 = N_1 - 1 = 9$ degrees of freedom and $n_2 = N_2 - 1 = 11$ degrees of freedom, we find that at the 5 per cent level of significance F is 2.90 and at the 1 per cent level of significance F is 4.63.

Hence we conclude that, since our value of F is 2.04 which is less than the F for the 5 per cent level, the larger variance is not significantly greater than the smaller. In other words, there are not sufficient grounds to reject the hypothesis that the two samples could have come from the same normal universe.

IV. 10. *Significance of a Correlation Coefficient.* The question here is: Could the sample whose coefficient of correlation is r have come from a non-correlated universe? We use

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \qquad \text{IV.10.1.}$$

where the degrees of freedom are N — 2.

To illustrate: Suppose we have a sample of size 11 whose coefficient of correlation is 0.60. Could this sample have come from a non-correlated universe?

Substitute these values in IV.10.1., and we obtain

$$t = \frac{0.60\sqrt{11-2}}{\sqrt{1-.36}}$$

$$= \frac{1.80}{.8} = 2.25$$

From the t-table with 9 degrees of freedom we find that at the 5 per cent level of significance $t = 2.262$ and at the 1 per cent level of significance $t = 3.250$. Hence we conclude that a little more than 5 per cent of the time the sample could have come from a non-correlated universe and a little less than 95 per cent of the time, it could not. In other words, the odds are about 95 to 5 in favor of rejecting the hypothesis that the sample could have come from a non-correlated universe.

In the case of a multiple correlation coefficient, if we wish to test whether the sample came from a non-correlated universe, the criterion is

$$F = \frac{r^2_{1.23\ldots n}/(m-1)}{1-r^2_{1.23\ldots n})/(N-m)} \qquad \text{IV.10.2.}$$

where m is the number of parameters in the regression function, N is the size of the sample and $N_1 = m-1$, $N_2 = N-m$ are the respective degrees of freedom.

To illustrate: Assume that $r_{1.23} = 0.60$ and that the regression function is a plane that is, $m = 3$ and that the size of the sample is 103.

Substituting in IV.10.2., we have

$$F = \frac{.36/2}{.64/100} = 28.1$$

From the F-table we find that at the 5 per cent level, $F = 3.09$ and at the 1 per cent level $F = 4.82$ when $n_1 = m-1 = 2$ and $n_2 = N-m = 100$. Hence we conclude that there are ample grounds to reject the hypothesis that the sample came from a non-correlated universe.

To test the hypothesis concerning a partial correlation coefficient the procedure is the same as that for a simple correlation coefficient with the exception that the number of variables held constant must be substracted from the size of the sample N. Hence, if k-variables are held constant the test is

$$F = \frac{r^2_{12.\overbrace{34\ldots n}^{k}}/1}{(1 - r^2_{12.\overbrace{34\ldots n}^{k}})/(N-k-1)} \qquad \text{IV.103 .}$$

REFERENCE, CHAPTER IV

[1] Yule, G. Udney, and Kendall, M. C., "*An Introduction to the Theory of Statistics*," C. Griffin & Co., London, 1937.

[2] Croxton, F. E., and Cowden, D. J., "*Applied General Statistics*," Prentiss-Hall Inc., New York, 1946.

[3] Rider, Paul, "*Statistical Methods*," John Wiley & Sons Inc., New York, 1939.

[4] Kendall, M. C., "*The Advanced Theory of Statistics*," Charles Griffin & Co., London, 1946, Vol. I, page 40.

[5] Fisher, R. A., "*Statistical Methods for Research Workers*," Oliver and Boyd, Ltd., Edinburgh.

CHAPTER V

SOME APPLICATIONS OF STATISTICAL METHODS

V. 1. *Objective.* This chapter illustrates some of the applications of statistical methods to problems of most interest to traffic engineers. Usually a statistical approach *is more rational than any other* and leads to a better understanding of the factors involved. The methods apply to all types of traffic problems, but first we shall study those that have to do with highway capacity. These problems are of primary concern, for they are connected with the main purpose of a highway which is to serve traffic.

V. 2. *Confusion As to Meaning of Highway Capacity.* Before attempting any analysis, it is necessary that certain terms be defined. There is some confusion as to what is meant by highway capacity. This is brought out by the *Highway Capacity Manual*[1], which states that the term perhaps most widely misunderstood and improperly used in the field of highway capacity is the word *capacity* itself. Considerable work went into the preparation of this manual, and it offers the most authentic and complete information extant on capacity. In Part I, Definitions, is found the statement that "the term *capacity* without modification, is simply a generic expression pertaining to the ability of a roadway to accommodate traffic." The manual gives three levels of capacity:

1. *Basic Capacity*: "The maximum number of passenger cars that can pass a given point on a lane or roadway during one hour under the most nearly ideal roadway and traffic conditions which can be attained."
2. *Possible Capacity*: "The maximum number of vehicles that can pass a given point on a lane or roadway during one hour under the prevailing roadway and traffic conditions."
3. *Practical Capacity*: "The maximum number of vehicles that can pass a given point on a roadway or in a designated lane

during one hour without the traffic density being so great as to arouse unreasonable delay, hazard, or traffic conditions." Prevailing roadway conditions include roadway alignment, number and width of lanes.

From a practical standpoint, speed should be included in any definition of traffic capacity. The driver is interested primarily in the amount of time it takes him to arrive at his destination. Perhaps capacity, meaning vehicles per hour, should be supplemented by a dimensionless index number similar to the Reynolds number in hydraulics. This number would indicate critical limits.

Since the term *capacity* has a variable meaning, we shall in most cases use the word volume and *define it as the number of vehicles passing a given point per unit of time. Density will refer to the number of vehicles in a given length of lane.* With these definitions,

Average Volume = Average Density times Average Speed.

V.3. *Theoretical Maximum Capacity* (*Volume*). The amount of traffic per unit of time depends on the speed and the spacing between vehicles. The greater the speed the larger is the volume, and the greater the spacing the less is the volume. Therefore,

$$\text{Volume} = \frac{\text{Speed}}{\text{Spacing}}$$

This same reasoning applies to any number of lanes in the same direction, but with more than one lane, passing takes place, which adds another factor to be considered. For the sake of simplicity, we shall first take up the theoretical capacity of a single lane.

In general, anyone who has observed traffic knows that as speeds increase, the spacing between vehicles increases. If the spacing increases at a greater rate than the speed, then there is an optimum speed that gives a maximum volume. If the spacing increases at a rate equal to or less than the speed, then the higher the speed the greater the volume. The question of minimum spacing needs to be examined critically.

The original assumption was that drivers should and did maintain a safe stopping distance behind the vehicle ahead. This safe stopping distance was based on the possibility that the car ahead

might stop instantaneously. This, of course, practically never happens for it can take place only through some unusual occurrence such as the head-on collision of two vehicles. That the original assumption of minimum spacing persists is evidenced by an article in *Traffic Engineering* for August, 1950, by Dr. Victor F. Hess, Physics Department, Fordham University, New York.[2] It should be mentioned that Dr. Hess is deriving a formula for safe travel at a maximum efficiency. This article states accurately that the stopping distance includes (1) a, the distance the vehicle travels during the "reaction time", (time interval between the stop signal observed and the instant the brakes are applied) and (2) b, the distance the vehicle travels after the brakes are applied. The distance a is proportional to the speed of the car v.

$$a = tv$$

Distance b, the braking distance, is the distance required to absorb the kinetic energy of the vehicle ($^1/_2\, mv^2$), and therefore must vary with the square of the velocity; that is

$$b = kv^2$$

in which the constant k is a factor depending upon the efficiency of the brakes and the coefficient of friction between the tires and the pavement. The stopping distance is equal to

$$a + b = tv + kv^2$$

in which t = reaction time, which is usually taken as .75 second.

V.4. *Stopping Distance And Minimum Spacing.* Observations have proved that the stopping distance is not the minimum spacing between vehicles. This fact may also be arrived at by inductive reasoning.

If we assume that two vehicles are mechanically equivalent and traveling at the same speed, then one can be stopped in the same distance as the other, and if they both start to stop at the same instant, they will come to rest at the same distance apart as when the brakes were applied. The fact that the brakes cannot be applied at the same time results from the rear driver's needing time to react. What takes place is that the driver sees the car ahead start to stop and then reacts and applies his brakes. This

reasoning leads to the conclusion that the minimum spacing between vehicles consists of the distance required for reaction plus an additional distance which the driver maintains as a safety factor. This factor of safety distance may be quite small.

From photographic observations of vehicles traveling in queues so that each one could be assumed to be traveling at minimum spacing, it was found that the average minimum spacing in feet was approximately $s = 1.1v + 21$ in which v = speed in miles per hour*.[3] The factor 1.1 corresponds to the reaction time of .75 seconds if the speed is given in feet per second. The 21 feet is the spacing when $v = 0$, and includes the length of the vehicle. This factor was determined in 1933, for a given composition of traffic and would evidently not apply in all conditions. It may be noted that if the spacing is expressed in time, it tends to be a constant. At 20 m.p.h. the time spacing would be 1.46 seconds; at 30 m.p.h., 1.2 seconds; and at 40 m.p.h., 1.1 seconds.

Observations in urban traffic have shown that the average minimum spacing between vehicles expressed in time is practically a constant, regardless of speed. In one case, it was found to be 1.1 seconds for all speeds which were low.[4]

In Part 3 of the Capacity Manual, Figure I shows the minimum spacings given in the table below. These spacings, if we assume a reaction time of .75 seconds, may be divided into a reaction-judgment distance plus a braking distance.

Table V.1

Speed	*Observed Minimum Spacing*	*Reaction Distance .75 Seconds*	*Additional Braking Distance*	*Ratio of Braking Distances*	*Ratio of* v^2*'s*
10	44	11	33	$^{33}/_{38} = .87$	$10^2/20^2 = 0.25$
20	60	22	38	$^{38}/_{47} = .81$	$20^2/30^2 = 0.45$
30	80	33	47	$^{47}/_{64} = .73$	$30^2/40^2 = 0.56$
40	108	44	64	$^{64}/_{85} = .75$	$40^2/50^2 = 0.64$
50	140	55	85		

* Coupare with the formula $s = 0.909\ v$ (III. 23.2) which was based on data which did not include zero speeds.

The braking distances for stopping should be proportional to the square of the speeds, but as shown in the table, the minimum spacings are not proportional to this amount. This is additional evidence that minimum spacings do not depend on braking ability.

V. 5. *Interpretation of Minimum Spacing Formula.* The formula $s = 1.1v + 21$ would give a maximum traffic flow of about 4000 vehicles per hour per lane. This, of course, is *never realized except momentarily*. If a stream of traffic were moving at this minimum spacing, the slowing or stopping of any vehicle would immediately affect all following vehicles. The formula is not given because of its practicability but because it points to two significant facts.

a. The volume increases with speed, but apparently approaches a maximum point at about 40 miles per hour where the constant 21 ceases to be significant.
b. The minimum spacing depends primarily on "reaction-perception-judgment" time.

V.6. *Limiting Factors.* To summarize: The factors that limit the capacity of a highway are:

1. Necessary minimum clearance between vehicles.
2. Slow moving vehicles that retard others, when passing is not possible, due to lack of space on the opposite lane or to restricted sight distance.
3. Reduced overall speeds caused by the physical features of the highway, the mechanical characteristics of vehicles, or the desire of drivers.

These factors need to be studied in as much detail as possible if we are to reach a clear conception of the problem of measuring the ability of a highway to accommodate traffic.

V.7. *Additional Relationships of Spacing and Speed.* In a study made in Ohio in 1934,[4] it was found that there is a straight line relationship between average density in vehicles per mile (spacing)

and average speed. As the density increases, the speed decreases. Expressed in the form of an equation

$$\frac{\text{Speed}}{\text{Density}} = k$$

where k is a constant for a given roadway and composition of traffic. If this relationship is true, and it was based on observations of over 220 groups of 100 vehicles each, it means that with a given highway and composition of traffic the potential capacity range can be obtained by getting the speeds at a low density and at a high density since two points determine a straight line.

That the relationship $\frac{\text{Speed}}{\text{Density}} = k$ may be only approximately true is indicated by information given in Figure 5, page 31, of the Highway Capacity Manual.

This figure indicates that there is a straight-line relationship between speed and volume of vehicles per hour. The equation of

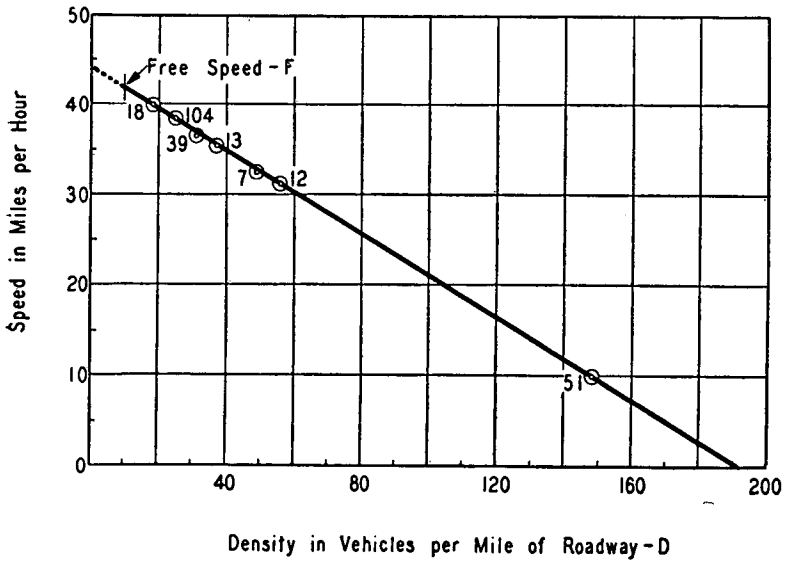


FIGURE V.1

SPEED IN MILES PER HOUR CORRESPONDING TO A GIVEN AVERAGE DENSITY IN VEHICLES PER MILE OF ROADWAY

the curve for "the majority of existing highways" as nearly as may be judged from the Figure, is

$$S = 43 - .009\,V,$$

where S equals speed in miles per hour and V equals volume.

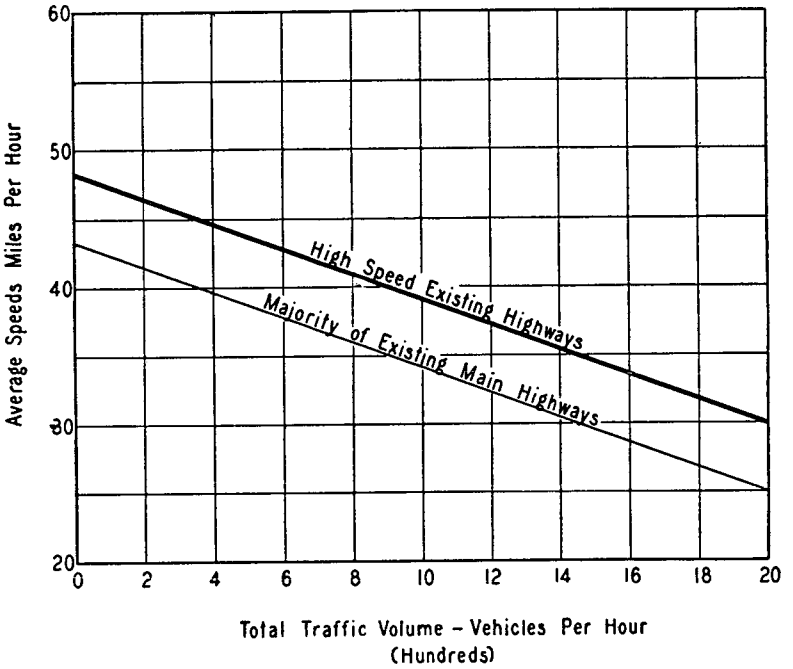


FIGURE V.2
AVERAGE SPEED OF ALL VEHICLES ON LEVEL, TANGENT SECTIONS OF 2-LANE RURAL HIGHWAYS

(Figure 5, page 31, "*Highway Capacity Manual*", Used by Permissions of Bureau of Public Roads, U.S. Department of Commerce.)

Letting D = density in vehicles per mile of roadway, $V = D \cdot S$, so that

$$S = 43 - .009\,V = 43 - .009\,D \cdot S$$

or

$$S = \frac{43}{1 + .009D}$$

By plotting speed against density Figure V.3. is obtained. The graph has very little curvature being nearly a straight line. Hence for practical purposes it may be assumed with slight error that speed varies directly (i. e. lineally) with density. It appears that this may be as nearly correct as the assumption that speed varies directly and lineally with volume.

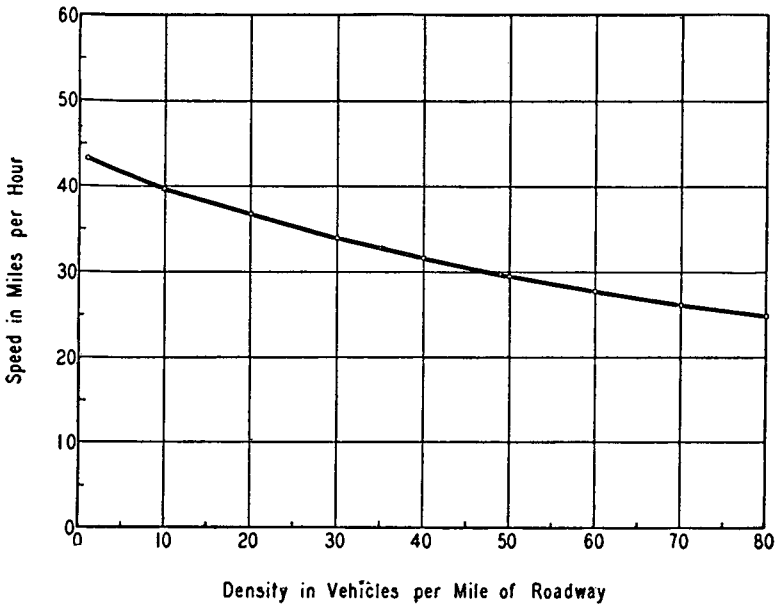


FIGURE V.3
AVERAGE SPEED OF ALL VEHICLES ON LEVEL, TANGENT SECTIONS OF THE MAJORITY OF EXISTING 2-LANE MAIN RURAL HIGHWAYS

Returning to the 1934[4] report it will be noted that in Figure V.1. (taken from page 468 of the report) the point that is marked "free speed" indicates that practically no drop in speed on the two-lane roadway was observed until the volume reached about 400 vehicles per hour. The figures near the curve show the number of groups of 100 vehicles each for which the point marked is the weighted average. The maximum possible volume was not ob-

served directly, but was obtained by assuming that the curve was a straight line. The "free speed" for the curve shown was 43.8 m.p.h. This point is indicated to be about ten units to the right since no noticeable speed drop was observed until the volume reached about 400 vehicles per hour. The maximum possible volume would come at the mid-point of the curve and would equal $\frac{46}{2} \times \frac{195}{2} = 2{,}300$ (approx.) vehicles per hour. That the mid-point of the curve gives the maximum volume is easily proved.

Let S_m = maximum speed and D_m = maximum density, then

$$\text{Slope of curve} = -\frac{S_m}{D_m}$$

Let x = varying values of D, then $V = \left(S - x\frac{S_m}{D_m}\right)x$

$$= \left(Sx - x^2\frac{S_m}{D_m}\right)$$

Differentiating with respect to x

$$\frac{dV}{dx} = S - 2x\frac{S_m}{D_m}$$

For maximum volume $S - 2x\frac{S_m}{D_m} = 0$

whence, $x = \frac{D_m}{2}$ = mid-point of the curve.

If this straight-line relationship holds, then the maximum capacity varies over a small range, since the end points of the line are fixed by the maximum average speed and the minimum spacing which have small variations.

V. 8 *Volume and Speed.* If volume is plotted against speed, the resulting curve is given in Figure V.4. This curve shows that there is a maximum volume and also that there are two speeds that give the same volume. At the lower speed, there is considerable time loss, Figure V.5.

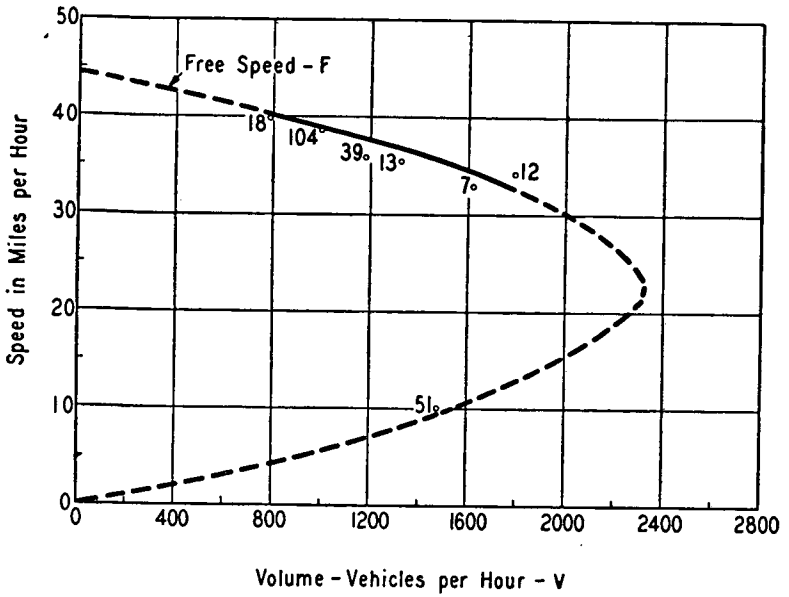


FIGURE V.4
SPEED IN MILES PER HOUR CORRESPONDING TO A GIVEN VOLUME IN VEHICLES PER HOUR ON A 2-LANE HIGHWAY

These curves bring out the fact that capacity needs to be expressed in terms of both volume and speed. At maximum volume there is always a considerable time or speed loss. The maximum volume is evidently not a design volume.

The Capacity Manual gives a great deal of evidence that there are definite relationships between speeds and volumes. This is brought out by numerous curves which show such information as the number of drivers desiring to pass compared to the number that have an opportunity to pass, the total percentage of the time that desired speeds can be maintained, and the point at which drivers become influenced by the presence of vehicles ahead of them. Using the facts set forth in the manual, it is our purpose to see if there is a rational explanation of the interrelationships of the different phases of the behavior of drivers that can be expressed mathematically.

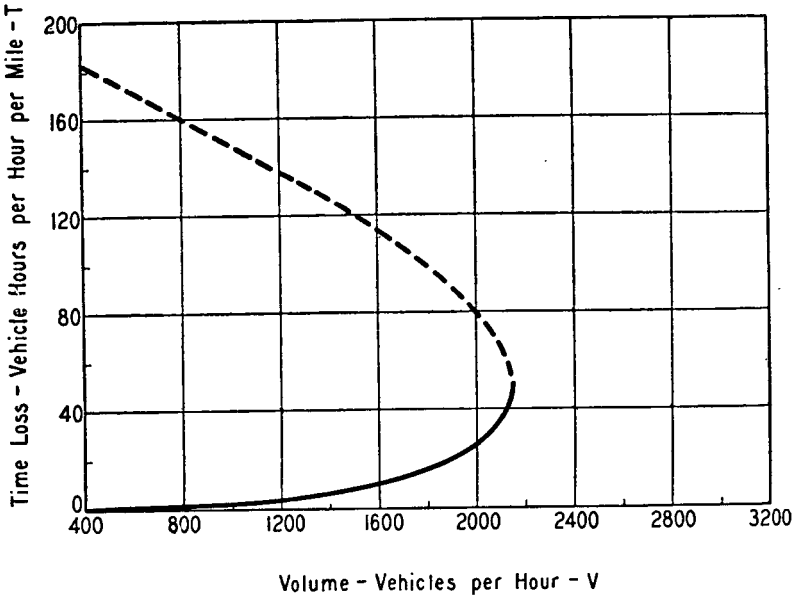


FIGURE V.5

VEHICLE TIME LOSS DUE TO CONGESTION ON A 2-LANE HIGHWAY

V. 9. *The Nature of the Problems of Highway Traffic.* We have discussed some of the elements of the problems of highway capacity, but have said very little about the nature and variability of these elements. It is this variability that makes it difficult to solve the problems involved. If all vehicles traveled at the same speed, or if all people reacted in the same time interval, or if all drivers maintained the same spacing at the same speed, the solutions would be comparatively easy.

There is nothing new about the idea that the behavior pattern of drivers is a stochastic variable. One of the writers found in 1933, as already mentioned, that the minimum spacing depended primarily on reaction-time which psychologists have long recognized as a stochastic variable.[3] Mr. John P. Kinzer assumed in 1934, that the traffic distribution on a roadway followed a "random" or Poisson distribution.[8] In England, Mr. William F. Adams found that free flowing traffic conformed so well to the distribution given

by a random series that it might be described as "normal." That the time spacings between vehicles follow a random series in urban traffic was reaffirmed by a study made in 1944–46.[7]

V. 10. *Spacing as a Random Series.* The assumption that spacing in either time or distance units follows the "random" series furnishes a means of studying the nature of spacing. To satisfy the conditions of the Poisson series, a roadway would have vehicles scattered along it at *random* so that any vehicle would be completely independent of any other vehicle, and equal segments of the road would be equally likely to contain the same number of vehicles. Granting that these conditions exist, the total number of vehicles on a roadway divided by the number of segments of road equals "m" the average number of vehicles per segment. Then, according to the Poisson series, the probability of zero vehicles appearing in a segment is

$$e^{-m}\left(\frac{m^0}{0!}\right)$$

The probability of one vehicle appearing is

$$e^{-m}\left(\frac{m^1}{1!}\right)$$

The probability of two vehicles appearing is

$$e^{-m}\left(\frac{m^2}{2!}\right)$$

and the probability of n vehicles appearing is

$$e^{-m}\left(\frac{m^n}{n!}\right)$$

The sum of all the individual probabilities is

$$e^{-m}\left(\frac{m^0}{0!}+\frac{m^1}{1!}+\frac{m^2}{2!}+\ldots+\frac{m^n}{n!}+\ldots.\right)$$

But

$$e^m = \left(\frac{m^0}{0!} + \frac{m^1}{1!} \cdots\cdots + \frac{m^n}{n!} + \cdots\cdots\right)$$

Therefore,

$$e^{-m} \cdot e^m = e^0 = 1$$

This simply demonstrates what we know, namely that the sum of all probabilities is unity, which means that an event is certain to

Table V.2

FITTING OF POISSON CURVE BY CHI-SQUARE TEST

NUMBER OF VEHICLES APPEARING IN FIVE-MINUTE INTERVALS

Observations Taken on U.S. 20 Near Oaklawn, Illinois. Data Supplied by the U.S. Public Roads Administration.

1	2	3	4	5	6	7
No. of vehicles appearing in a 5-min.-interval (= x)	*Observed frequency* f_o	*Probability from the Poisson Table with* m = 4.75 seconds	*Theoretical frequency* f_t	$f_o - f_t$	$(f_o - f_t)^2$	$\frac{(f_o - f_t)^2}{f_t}$
0	3	.009095	2.983	—.004	.000016	.000001
1	14	.042748	14.021			
2	30	.100457	32.949	—2.949	8.696601	.264
3	41	.157383	51.621	10.261	112.805641	2.185
4	61	.184925	60.655	.345	.110025	.002
5	69	.173830	57.016	11.984	143.616256	2.519
6	46	.136167	44.662	1.338	1.790244	.040
7	31	.091426	29.987	1.013	1.026169	.034
8	22	.053713	17.617	4.383	19.210689	1.090
9	8	.028050	9.200	—5.095	25.959	1.613
10	2	.013184	4.324			
11	0	.005633	1.847			
12	1	.002206	.724			

Chi-square, $\chi^2 = 7.747$

m = 4.75 seconds

Degrees of Freedom = 9 — 2 = 7

happen or not to happen. In this case, it means that any segment is sure to contain zero or more vehicles since this covers all alternatives.

V. 11. *Test of Goodness of Fit of the Poisson Series.* The goodness of fit of the Poisson Series to a set of data may be tested by the Chi-square (χ^2) test. A cumulative Poisson table of probabilities is used to obtain the theoretical frequencies. The data in the illustrative example consist of the numbers of vehicles appearing in five minute intervals on Route U.S. 20 near Oaklawn, Illinois. The volume of flow averaged about 115 vehicles per hour. These data were made available by the Public Roads Administration.

The first two columns in Table V.2. show the observed data. The figures in Column Three are taken from a Poisson table. Column Four is found by multiplying the figures in Column Three by the number of intervals observed (N = 328) to obtain the theoretical frequency. Column Five gives the differences between the observed or actual frequencies and the theoretical. Note that in this column the first two terms and the last four in Column Four have been combined to obtain a minimum actual or theoretical frequency that must be five or more. Column Six gives the square of these differences. The figures in Column Six divided by the theoretical frequency give the values in Column Seven. The sum of these values, 7.747, equals "Chi-square" (χ^2).

The degrees of freedom are equal to the number of classes less 2, i. e., $9 - 2 = 7$. From a Chi-square table of probability levels, it is found that the probability level is about .60 or 60 per cent.

A 5 per cent level is usually taken as sufficient to indicate that there is reason to reject the hypothesis that the data can be represented by the curve. Therefore, the present level of about 60 per cent is taken to be rather conclusive evidence that the data may be represented by the Poisson Curve.

V. 12. *Test of Goodness of Fit of the Poisson Series to the Distribution of Spacings Between Vehicles.* As already mentioned we are also interested in the distribution of the time or distance spacings between successive vehicles. It is these time-gaps on the opposite

Table V.3

Fitting of Poisson Curve by Individual Terms Table
Time Spacing Between Vehicles (Chi-square Test)

Frequency Distribution of Time Spacings Between Vehicles on a Two-Lane Highway (Routes U.S. 50 and 240 in Maryland). Data Furnished by Public U.S. Roads Administration.

1	2	3	4	5	6	7
Class Intervals in Seconds	*Observed Time Spacings* $= f_o$	*Probability from Poisson Table*	*Theoretical Frequency* $= f_t$	$f_o - f_t$	$(f_o - f_t)^2$	$\frac{(f_o - f_t)^2}{f_t}$
0–1	78 } 285	.001360 }	.89 } 6.81	77 } 278		11331.9
1–2	207 }	.008979 }	5.92 }	201 }	77284	
2–3	94	.029629	19.55	74	5476	273.8
3–4	58	.065183	43.02	15	225	5.2
4–5	24	.107553	70.98	47	3619	51.7
5–6	17	.141969	93.70	77	5929	63.8
6–7	23	.156166	103.07	80	6400	62.1
7–8	11	.147243	97.18	86	7396	76.2
8–9	18	.121475	80.17	80	6400	80.0
9–10	10	.089082	58.79	49	2401	41.4
10–11	8	.058794	38.80	31	961	25.3
11–12	5	.035276	23.28	18	324	14.1
12–13	7	.019402	12.81	6	36	3.0
13–14	13	.009851	6.50	7	49	7.1
14–15	8	.004643	3.06	5		
15–16	8	.002043	1.35	7		
16–17	4	.000843	.56	3		
17–18	3	.000327	.22	2		
18–19	1 } 87	.000120 } .008037	.08 } 5.314	9 } 86	6724	1268.7
19–20	5	.000042	.03	4		
20–21	4	.000014	.01	3		
21–22	0	.000004	.003	0		
22 & more	54	.000001	.0006	53		

(Chi-square, $\chi^2 = 13304.3$

m = mean = 6.6 seconds

Degrees of Freedom = 14 − 2 = 12

lane that are used in passing. We shall now check the goodness of fit of the time spacing distribution to the Poisson Curve. The data were taken on Route U.S. 240, Maryland, and were furnished by the Public Roads Administration. The Chi-square test will be used.

According to this method as shown in Table V.3., it is immediately evident that there is a wide discrepancy between the actual and the theoretical frequencies. The probability level is practically zero.

If the distribution of time gaps between vehicles is not a Poisson series, what is it? To determine this, let us re-examine the nature of the Poisson series when applied to spacing distribution.

The probability of the occurrence of a time or distance gap of a given length is the probability that no vehicle will appear in the given interval.

For example, given a volume of 400 vehicles per hour, let it be required to determine the probability "P_0" of a one second interval having no vehicle. The average number of vehicles per second "m" is equal to $^{400}/_{3600} = \frac{1}{9}$; therefore, the probability of a one second interval having no vehicle is equal to

$$e^{-m}\left(\frac{m^0}{0!}\right) = e^{-\frac{1}{9}}\left(\frac{m^0}{0!}\right)$$

$$= e^{-\frac{1}{9}}, \text{ since } \left(\frac{m^0}{0!}\right) = \frac{1}{1} = 1$$

The probability of no vehicle appearing in 2 seconds is $e^{-\frac{2}{9}}$, and in 3 seconds $e^{-\frac{3}{9}}$. In general, the probability P_0 of there being no vehicles in "s" seconds is equal to e^{-m}. This equation is of the general form of

$$y = e^x$$

which may be written

$$\log_e y = x$$

therefore the equation when plotted on semilog-paper becomes a straight line. The exponent, -m, means that the slope of the line is negative.

For plotting on semi-log paper we first arrange the data, as shown in the cumulative Table V.4. where the percentages of spacings equal to or less than a given interval are tabulated.

Table V.4

FITTING OF POISSON CURVE
EXPECTED ERROR METHOD

Class interval in seconds	*Class frequency* (f)	*Cumulated frequency* (f_0)	*Cumulated per cent*	*Expected error or natural uncertainty*	*Expected error in per cent*
0–.9	78	78	10.8	8.28	1.26
1–1.9	207	285	43.2	12.72	1.93
2–2.9	94	379	57.4	12.72	1.93
3–3.9	58	437	66.2	12.15	1.84
4–4.9	24	461	69.8	11.79	1.79
5–5.9	17	478	72.4	11.5	1.74
6–6.9	23	501	75.9	10.9	1.65
7–7.9	11	512	77.6	10.8	1.64
8–9.9	18	530	80.3	10.2	1.55
10–11.9	23	553	83.8	9.4	1.42
12–13.9	20	573	86.8	8.7	1.32
14–15.9	16	589	89.2	8.0	1.21
16–17.9	7	596	90.3	7.5	1.14
18–19.9	6	602	91.2	7.3	1.11
20–21.9	4	606	91.8	7.0	1.06
22–23.9	6	612	92.7	6.7	1.02
24–25.9	6	618	93.6	6.21	.94
26–30.9	10	628	95.1	5.47	.83
31–35.9	11	639	96.8	4.52	.68
36–40.9	8	647	98.0	3.89	.59
41–45.9	6	653	98.9	2.56	.39
46–50.9	1	654	99.1	2.56	.39
51–55.9	4	658	99.7	1.40	.21
56–60.9	0	658	99.7	1.40	.21
61–70.9	1	659	99.8	1.15	.17
71–80.9	1	660	100.	0	0

$$\text{Mean} = \frac{4346.0}{660} = 6.585$$

These percentages are represented by the heavy dots which fall in an irregular line as shown in Fig. V.6. This is to be expected for unless a sample is very large there is always a "natural uncertainty" or difference between the sample values and those of the universe.

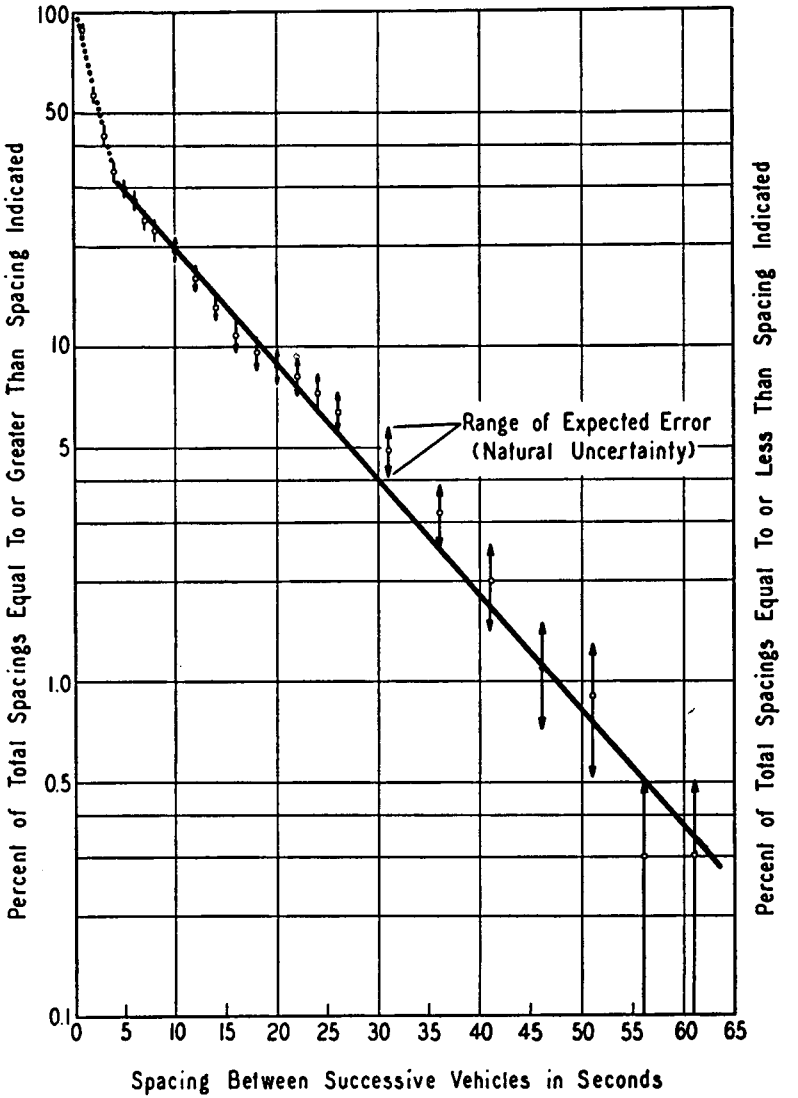


FIGURE V.6

GRAPH SHOWING PERCENTAGE OF VEHICLE SPACINGS AND THE PROBABLE AMOUNTS OF THE "NATURAL UNCERTAINTY" OF THE PLOTTED POINTS

A fair measure of this uncertainty is the standard deviation of a class or sample. The formula for this natural uncertainty is

$$Z = \sqrt{\frac{n}{n-1} f_0 \left(1 - \frac{f_0}{n}\right)}$$

where n equals the total number of happenings recorded, and f_0 equals the accumulated frequency. Since n in the present case is 660, $\frac{n}{n-1}$ is so nearly equal to 1 that it may be omitted and the equation becomes:

$$Z = \sqrt{f_0 \left(1 - \frac{f_0}{n}\right)}$$

An examination of this formula shows that the uncertainty depends upon the size of the sample and not upon the size of the universe. It may seem a little paradoxical that a 20 per cent sample may be no more representative of the universe than a 10 per cent sample. If, however, we recall that the size of the universe may be considered to be infinite, and this is practically true of traffic, then no sample is any nearer than any other to including all the universe. With this in mind it is entirely logical that the size of the universe does not appear in the formula for the measure of uncertainty.

If we could draw a line through the plotted points and stay within the natural uncertainty range we could conclude that the data could be represented by a straight line. But this is not the case as can be seen in Figure V.6., so it must be that the distribution of spacings is not the special case of the Poisson series which may be represented by the curve e^{-m}.

It appears, however, that the data can be closely represented by two straight lines. This implies that there may be two distributions, one for spacings less than about 4 seconds and another for spacings of more than that and that each is "random" in the limited case.

If we take the class intervals equal to 5 seconds in order to smooth the curve we obtain the points shown in Figure V.7. which is approximately a straight line. This indicates that if we are not

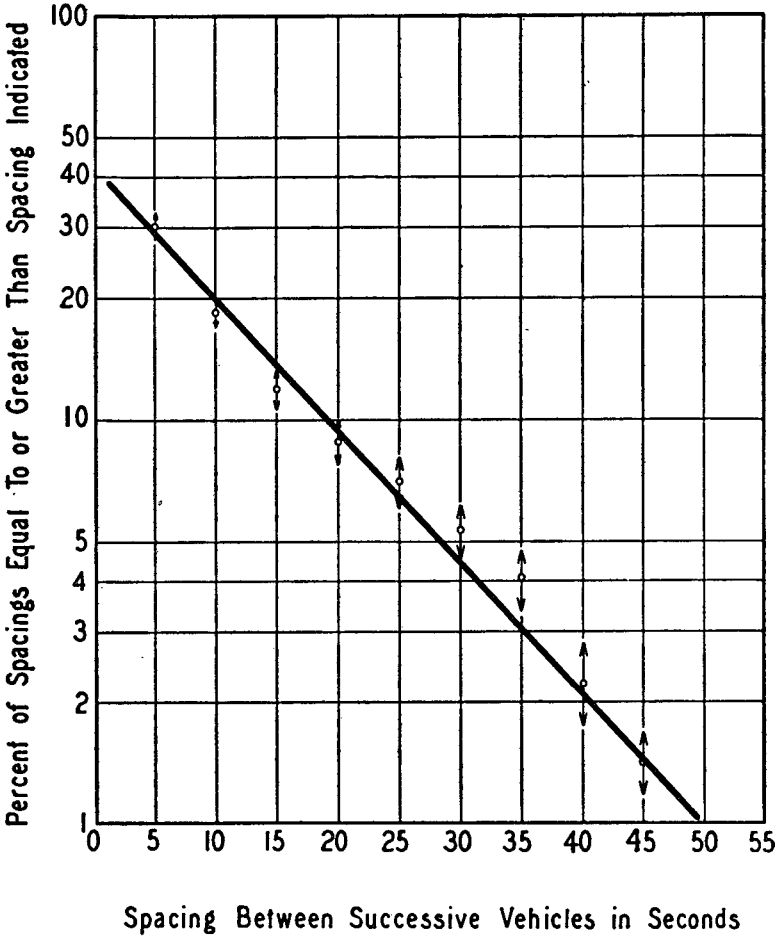


FIGURE V.7

DISTRIBUTION OF SPACINGS BETWEEN SUCCESSIVE VEHICLES: CLASS INTERVALS EQUAL TO 5 SECONDS

concerned with spacings of less than 5 seconds that the straight line represents the distribution of the spacings closely enough for approximate analysis.

V. 13. *Minimum Spacing.* For what is believed to be the first indication that minimum spacing distributions might be different from

those at greater distances, we refer to a study made in Ohio in 1934.[5] The cumulative frequency curve shown in Figure V.8. is plotted from data collected at that time. The spacings, center to center of vehicles, are in feet.

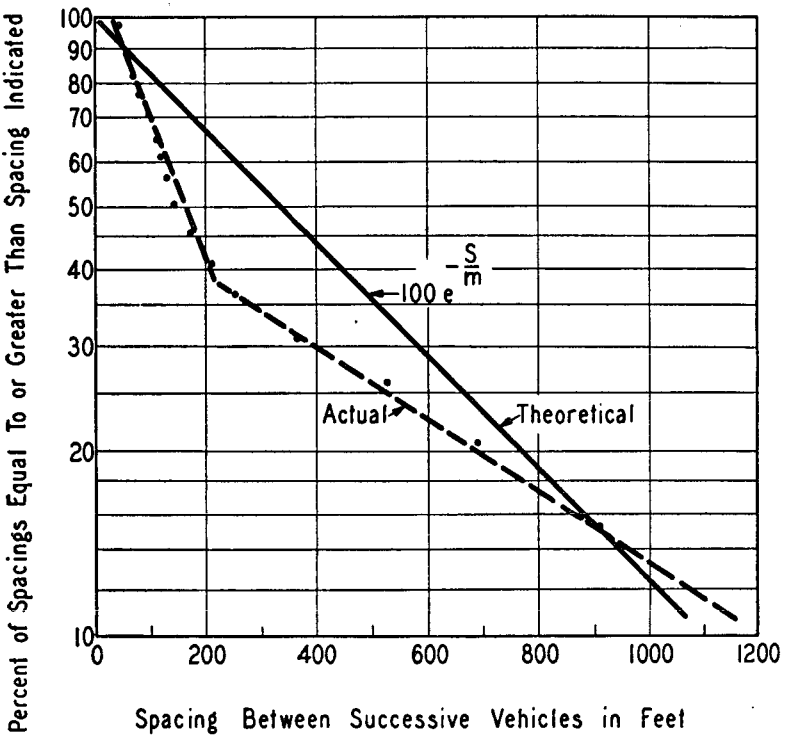


FIGURE V.8

CUMULATIVE FREQUENCY CURVE
OF SPACINGS BETWEEN SUCCESSIVE VEHICLES

It is indicated that the minimum spacing distribution is random and that it extends from about 30 feet to 200 feet. Evidently there are few, if any, spacings below 30 feet, and beyond 200 feet there is another random distribution different from that below 200 feet. This may be interpreted to mean that the distribution at less than 200 feet varies in accordance with the reaction-perception

time of the driver and his judgment of what constitutes a safe distance. Beyond 200 feet, the spacing may be judged to be in accordance with the chance placement of the vehicles on the highway. If the observed results are compared with the theoretical

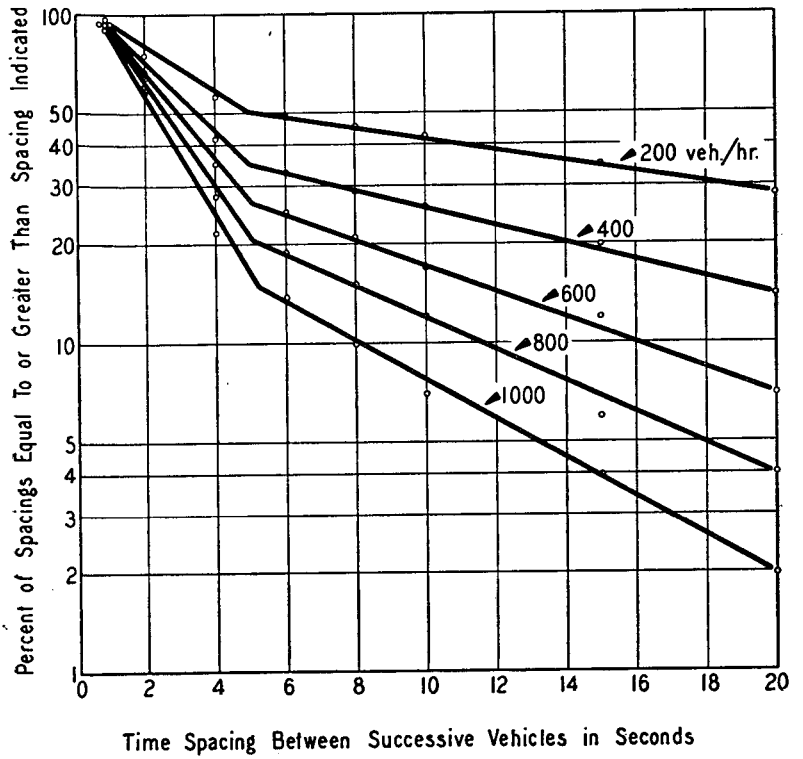


FIGURE V.9

CUMULATIVE FREQUENCY CURVE OF SPACINGS BETWEEN SUCCESSIVE VEHICLES FOR VARIOUS TRAFFIC VOLUMES ON A TYPICAL 2-LANE RURAL HIGHWAY

curve, it is found that the deviations from the random distribution are accounted for by there being:

(a) No spacings below 30 feet.
(b) An excess of spacings between 30 and 200 feet.
(c) A deficit of spacings in excess of 200 feet.

These discrepencies are logical, for the minimum spacing, center to center of vehicles, is limited by the length of the vehicles and because vehicles, closing up behind slower vehicles must wait for an opportunity to pass, create a preponderance of the smaller spacings.

If the spacing of about 200 feet is divided by the average speed of 34.1 miles per hour we obtain about 4 seconds as the limit of the zone of speeds reduced by the presence of other vehicles. These data from two locations, would not be supposed to give a conclusive answer.

For more extensive data, let us turn to Figure 9, page 40 of the *Capacity Manual*. These data replotted as nearly as is possible from the printed curves are shown in Figure V.9. They are in time spacings and the breaks in the curves seem to come between five and six seconds.

Theoretically, if the lines had no breaks there would be no interference, and if all vehicles were restricted there would be no breaks. These conditions were found and reported in the earlier paper referred to. To find the average of the "influenced" spacings we first make the reasonable assumption from the graphs that practically no spacings are under $^1/_2$ second or over 6 seconds, and draw a line between these points as in Figure V.10. This line then represents a random distribution of "influenced" spacings.

The next step is to let $S = m$, where m is the average spacing. Now the expression

$$100\left(e^{\frac{-S}{m}}\right) = 100\,(e^{-1}) = 0.368 = 36.8\%$$

so that the average would be at point 36.8 per cent and would equal about 1.7 seconds. At this average "random" spacing all vehicles would be travelling at a restricted speed due to the closeness of spacing between vehicles.

V.14. *The Minimum Spacing of Four-Lane Traffic*: Traffic on a four-lane highway does not have the same spacing restrictions as a two-lane roadway. Vehicles are free to weave into the adjoining lane. When the curves shown in Figure 10, page 41 of the *Capacity*

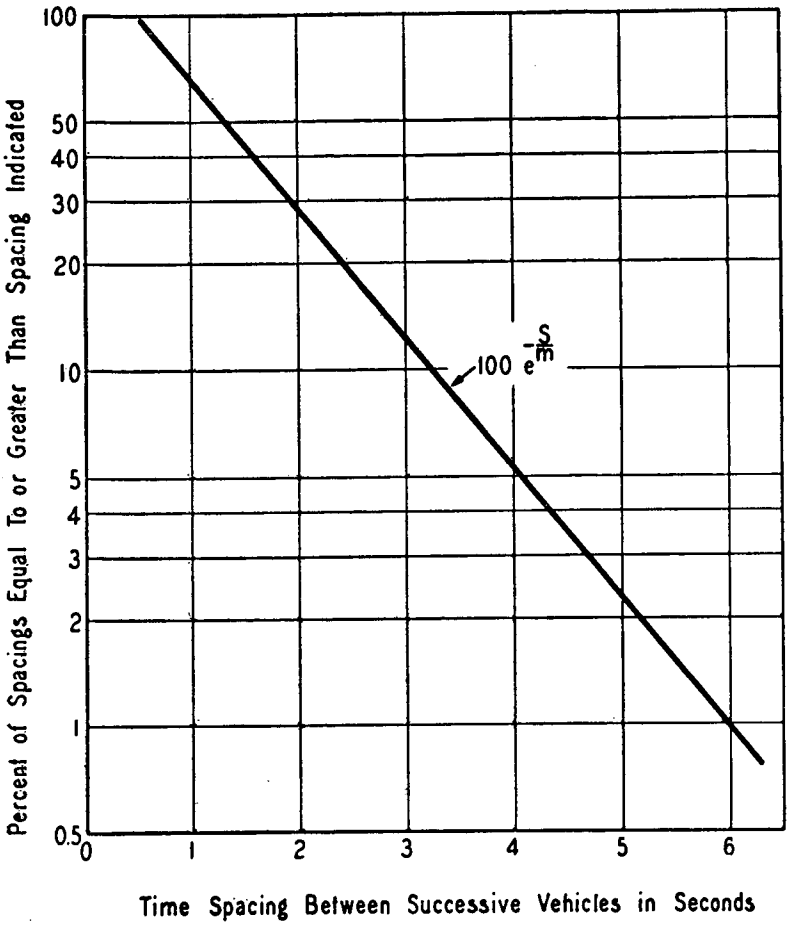


FIGURE V.10
RANDOM DISTRIBUTION OF "INFLUENCED" SPACINGS

Manual are replotted as shown in Figure V.11., the resulting curves show no breaks. The distribution of timespacings is evidently random throughout.

V.15. *Frequency Distribution of Speeds*: Having determined the characteristics of the spacing distributions, the next step is that of determining the nature of the distribution of automobile speeds.

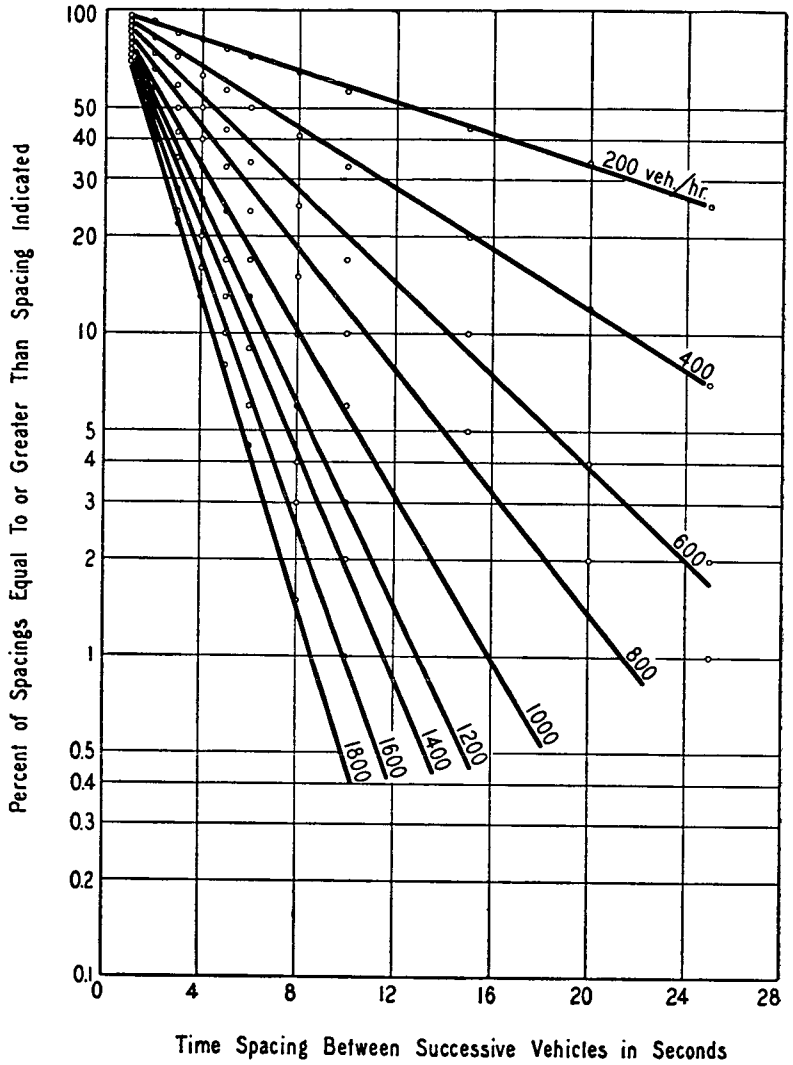


FIGURE V.11

CUMULATIVE FREQUENCY CURVE OF SPACINGS BETWEEN SUCCESSIVE VEHICLES FOR VARIOUS TRAFFIC VOLUMES ON A TYPICAL 4-LANE RURAL HIGHWAY

Table V.5

CALCULATION OF STANDARD DEVIATION
OF DISTRIBUTION OF VEHICLE SPEEDS

1	2	3	4	5
Speed in Miles per hour	*Observed no. of speeds* $= f_0$	*Deviation in class Intervals*	$f_0 d$	$f_0 d^2$
20.6 25.4	5	−4	−20	80
25.6 30.4	7	−3	−21	63
30.6 35.4	19	−2	−38	76
35.6 40.4	23	−1	−23	23
40.6 45.4	13	0	0	0
45.6 50.4	15	1	15	15
50.6 55.4	12	2	24	48
55.6 60.4	5	3	15	45
60.6 65.4	1	4	4	16
			−44	366

Arithmetic Mean $= \overline{X} = 40.8$ miles per hour

$$\sigma = S = 5\sqrt{\frac{\Sigma f_0 (d^2)}{N} - \left(\frac{\Sigma f_0 d}{N}\right)^2}$$

$$= 5.0\sqrt{\frac{366}{100} - \left(\frac{-44}{100}\right)^2}$$

$$= 5.0\sqrt{(3.66 - .1936)}$$

$$= 5\sqrt{(3.4664)}$$

$$= 5\,(1.862) = 9.31$$

$$=$$ standard deviation

Table V. 6.

FITTING OF NORMAL CURVE TO DISTRIBUTION OF VEHICLE SPEEDS CHI-SQUARE METHOD

1	2	3	4	5	6	7	8		9	10	11
Speed in Miles per Hour	*Mid-point of class*	*No. of Speeds = f_o = Observed frequency*	*Deviation of class limit from mean*	*Column 4 in Standard Deviation* $= \frac{(x)}{\sigma}$	*Percent of area between class limit and mean*	*Percent of area in class interval*	*Theoretical frequency = N (%) = f_t*		$(f_o - f_t)$	$(f_o - f_t)^2$	$\frac{(f_o - f_t)^2}{f_t}$
20.6 25.4	23	5	— 20.2	— 2.17	48.50	3.65	3.65	12.29	—.29	.084	.007
25.6 30.4	28	7	— 15.2	— 1.63	44.85	8.64	8.64				
30.6 35.4	33	19	— 10.2	— 1.09	36.21	14.98	14.98		4.02	16.160	1.079
35.6 40.4	38	23	— 5.2	— .56	21.23	20.43	20.43		2.57	6.605	.323
40.6 45.4	43	13	— .2 + 4.6	— .02 + .49	.8 18.75	19.55	19.55		—6.55	42.902	2.194
45.6 50.4	48	15	9.6	1.03	34.85	16.10	16.10		—1.1	1.21	.075
50.6 55.4	53	12	14.6	1.57	44.18	9.33	9.33		2.67	7.129	.764
55.6 60.4	58	5	19.6	2.11	48.26	4.08	4.08	5.41	.59	.348	.064
60.6 65.4	63	1	24.6	2.64	49.59	1.33	1.33				

Average Mean speed = 40.8 miles per hour $\chi^2 = 4.506$ N = 7 classes 7 — 3 = 4 degrees of freedom σ = S = 9.31

It has been found that this distribution closely follows the "normal curve." Again as in the two previous examples of "random" distribution, the usual method of making a test of the goodness of fit is the Chi-Square (χ^2) test. For the sake of simplicity, let us take a small sample of 100 recorded speeds. The area method of fitting a normal curve to the observed distribution will be used. The area included within any number of standard deviations may be obtained from prepared tables of areas of the normal curve. The calculation of the standard deviation is shown in Table V.5.

The steps in the calculation are arranged as shown in Table V.6., with the data in the respective columns consisting of the following:

(1) The speeds in class intervals of 5 miles per hour.
(2) The mid-points of the classes.
(3) The number of speeds recorded, i. e. the frequency f_0.
(4) The deviations of the class limits from the arithmetic mean.
(5) The deviations from the mean in terms of standard deviations. This column is obtained by dividing the numbers in column 4 by the standard deviation.
(6) Per cent of the area between the class limit and the mean. This is obtained from an area table of the normal distribution.
(7) Per cent of area in class interval. This is obtained by subtracting successively the numbers in column 6.
(8) The theoretical frequency f_t is obtained by multiplying the per cent of area in each class interval by the total number of speeds observed. This equals 100 in the present case.
(9) This column gives the difference between the observed frequency f_0 (column 3) and the theoretical frequency f_t (column 8).
(10) This column is obtained by squaring the items in column 9.
(11) The sum of the items in this column equals χ^2. This is the value we use with the Chi-square table.

In using the chi-square table we need to know the degrees of freedom. In fitting a normal distribution three degrees of freedom are lost (or three constraints are imposed) because (1) the total frequency, (2) the arithmetic mean, and (3) the value of the

standard deviation are used in computing the normal frequencies. The possible number of degrees of freedom is equal to the number of class intervals, 7 in this case. Therefore, $7 - 3 = 4$, the degrees of freedom in the given example.

We find from the Chi-square table that the probability level is more than 5 per cent which means that in more than 5 times out of 100 the sample could have come from the universe tested. This level of 5 per cent is taken to mean that there is not sufficient evidence to reject the hypothesis that the data can be represented by a normal curve. In the present case the probability is more than .30 which means that a variation as great as the amount found might occur in 30 cases out of 100 due to chance. Therefore it is not to be considered as significant.

V. 16. *A Graphical Method of Determining Goodness of Fit.* Another means of determining whether the distribution is normal or not is to plot the percentage of speeds at or less than various speeds on arithmetic probability paper. If the distribution is "normal" the observed data will be represented by a straight line. In such a case, due to symmetry the speed given by the intersection of the straight line with the 50 per cent ordinate is the most frequent and average speed, as well as the median. The usual definitions become:

Mean Average Speed = arithmetical mean of all speeds – also called probable or expected speed.

Median Speed = speed such that 50 per cent of the speeds are greater, and 50 per cent less.

Modal Speed = the most frequently occurring speed.

The data utilized are the numbers of cars with speeds equal to or less than a given series of equally spaced values. The same data will be used as in the first illustration. It is shown in Table V.7.

The points listed in Table V.7. are plotted in Figure V.12. It will be seen that they fall in rather irregular fashion, and that at first glance, the position of the 63.5 mile per hour point appears to preclude the possibility of drawing a satisfactory straight line.

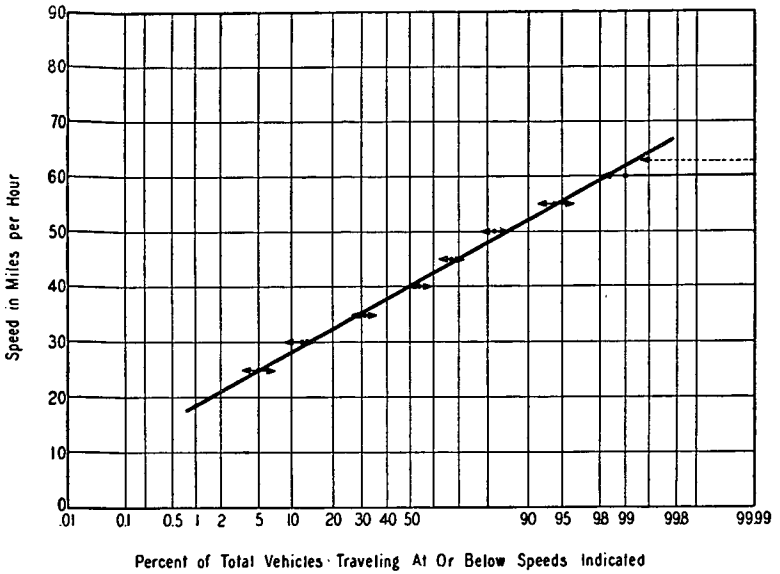


FIGURE V.12

GRAPH SHOWING PERCENTAGE OF VEHICLES TRAVELING ABOVE AND BELOW VARIOUS SPEEDS AND THE PROBABLE AMOUNTS OF THE "NATURAL UNCERTAINTY" OF THE PLOTTED POINTS

Table V.7

Speed in Miles Per Hour	*Cumulated Frequency* (f)	*Percent Equal to or Slower*	*Natural Uncertainty in Percent*
20.5	0	0	0.0
25.5	5	5	2.18
30.5	12	12	3.24
35.5	31	31	4.62
40.5	54	54	4.97
45.5	67	67	4.70
50.5	82	82	3.84
55.5	94	94	2.37
60.5	99	99	0.99
63.5	100	100	0.0
65.5	100	100	0.0

First, however, it is important to consider the probable amounts of the "natural uncertainty". Recall that the natural uncertainty $Z = \sqrt{f_0\left(1 - \frac{f_0}{n}\right)}$. This natural uncertainty is given for each frequency in the last column of the table.

If the percentage of cars travelling slower than a given speed or equal to it is plotted against speed, the points will fall in an irregular line. This is to be expected, particularly when the number of cars represented in one diagram is only 100. If counts are made a number of times under precisely the same conditions of traffic, the percentage traveling faster than, say 40 miles per hour, will never be exactly the same, except by chance. There will be a certain dispersion around the average value for several groups of 100 cars. This we have already referred to in article V.12. as a "natural uncertainty".

Through each plotted point, a horizontal line is drawn representing the allowed $\pm$ range in the value of f_0. It is then permissible to draw a smoothed curve in such a way that it passes through all the horizontal lines, attempting to draw it so that the sum of the deviations from the actually counted values shall be equal.

In the present case, a straight line satisfies all but the 63.5 mile per hour point. In the preceeding formula, f_0 should really be the mean number of cars with velocity equal to or less than the given amount, found from a great number of sets of 100 cars under the same traffic conditions. In such cases, it is fair to suppose that an occasional car traveling faster than 63.5 miles per hour would be found. Then the actual percentage slower than 63.5 would be slightly less than 100. If, for example, it were 99.5, the natural uncertainty would then be $\pm$ 0.7, and the point and the dotted line would give the result. In this case, it is evident that the straight line can be passed through all the horizontal lines. This means principally, that the points given by the higher speeds are too erratic and sensitive to accidental fluctuations to be given much weight in drawing of the curve. Probably all points for percentages less than 2 and greater than 98 should be ignored in drawing the curve.

That the "normal" dispersion pattern describes the speed range is demonstrated if we replot some of the speed curves shown in Figure 5 of the *Capacity Manual.* These curves plotted on arithmetic probability paper are very nearly straight lines as shown in Figure V.13., where the distributions for traffic volumes of 600, 1200, and 1800 vehicles per hour are given.

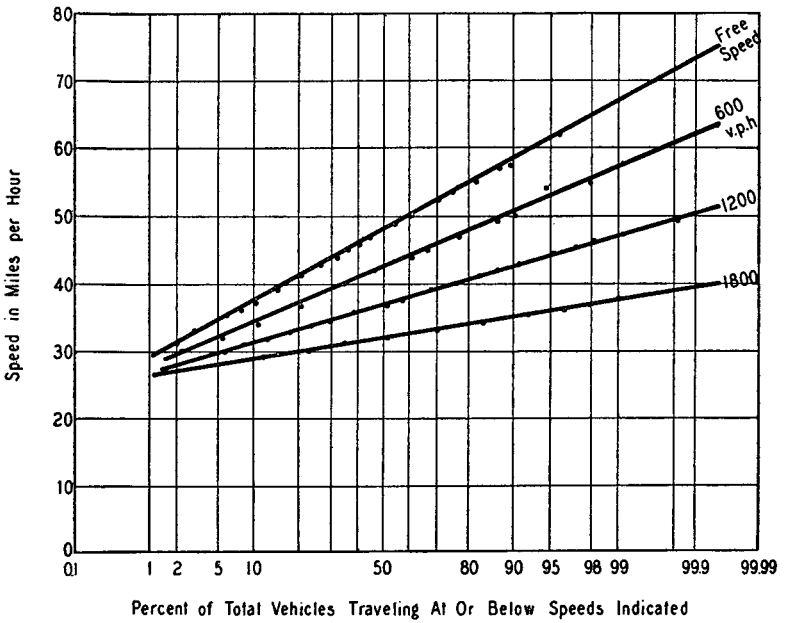


FIGURE V.13

TYPICAL SPEED DISTRIBUTIONS AT VARIOUS TRAFFIC VOLUMES ON LEVEL, TANGENT SECTIONS OF 2-LANE, HIGH-SPEED EXISTING HIGHWAYS

Judging from these examples it may be assumed that a straight line will satisfy the data and that the "smoothed" values read from the curve may be used in analysis.

V. 17. *Estimating Speeds and Volumes.* Having determined the free speed distribution on a highway, it is possible to estimate the speed at greater traffic volumes.

The first step is to find the average difference in speed between the vehicles being passed and the passing vehicles. The rate at which the faster vehicles are overtaking the slower ones can be found from a speed distribution curve.[a] Such a curve is shown in Figure V.14. as replotted from Figure 4, page 30, of the *Capacity*

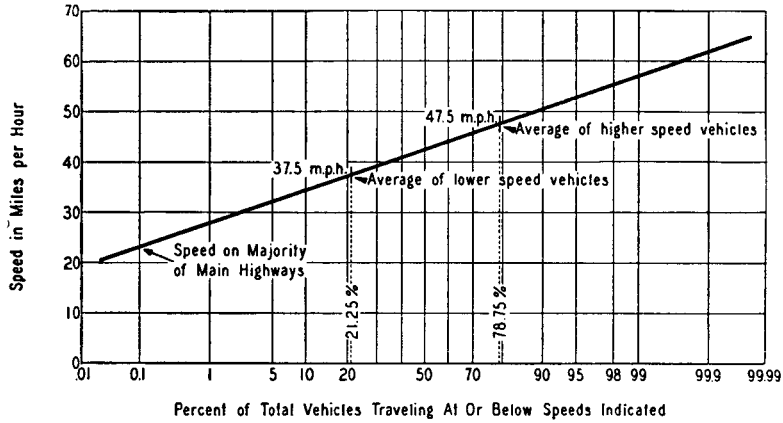


FIGURE V.14

FREQUENCY DISTRIBUTION OF TRAVEL SPEEDS OF FREE MOVING VEHICLES ON LEVEL, TANGENT SECTIONS OF THE MAJORITY OF EXISTING 2-LANE MAIN RURAL HIGHWAYS

Manual. It is evident that there are just as many vehicles traveling above the average (or 50 percentile speed) as below it. The average speed differential is the difference between the average speed of the 50 per cent faster vehicles and the 50 per cent slower vehicles. The average of the 50 per cent faster vehicles comes at the 78.75 percentile, and the average of the 50 per cent slower vehicles comes at the 21.25 percentile.[b]

(a) In a study of passing made in 1935[6], it was found that vehicles in the act of passing other slower vehicles were traveling 9 to 10 miles per hour faster. The *Capacity Manual* gives 9.6 miles as the average passing speed differential. (Footnote continued on p. 183).

(b) This can be proved as follows: Let Figure V. 15 represent the same curve as Figure V. 14., but plotted on linear cross section paper.

The average speed of the faster vehicles equals 47.5 miles per hour and the average for the slower ones is 37.5 miles per hour, so that the average difference is 10 miles per hour.

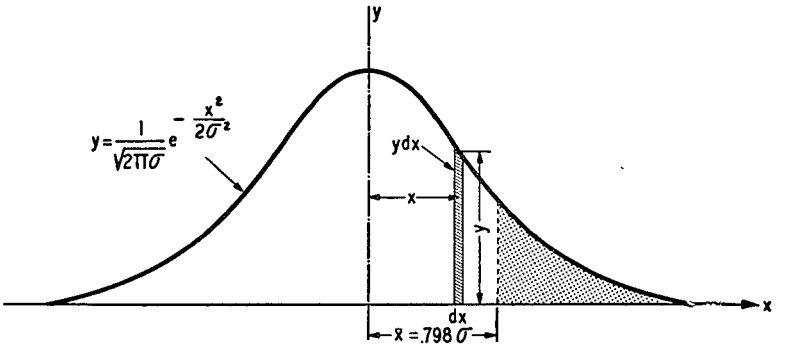


FIGURE V.15

DETERMINATION OF THE MEAN ABSCISSA OF THE UPPER HALF OF THE NORMAL DISTRIBUTION CURVE AND THE AREA TO THE RIGHT OF THIS ABSCISSA

Required: To find (1) the mean abscissa of the upper half of the normal distribution curve, and (2) the area to the right of this abscissa.

$$\bar{X} = \frac{\int_0^\infty x\,y\,dx}{\int_0^\infty y\,dx} = 2\int_0^\infty x\,y\,dx$$

$$= \frac{2}{\sqrt{2\,\pi}\,\sigma}\int_0^\infty x e^{-\frac{x^2}{2\sigma^2}}\,dx$$

$$= \sqrt{\frac{2}{\pi}}\,\sigma, \text{ which is about } = .798\,\sigma.$$

From a table of areas under the normal curve, the area to the right of .798 σ is .2125, or 21.25 per cent of the total area. In other words, 21.25% of the speeds will exceed the average of all the speeds higher than the average speed. Similarly, because of symmetry, 21.25% of the speeds less than the average will be less than the average of all the speeds lower than the average speed.

Having found the average speed differential we next find the percentage of spaces either large enough or too small to permit passing.

Assume for example that a two lane road is carrying 800 vehicles per hour and that the distribution of time spaces is random with the average spacing $m = \frac{3600}{400} = 9$ seconds, (since there are 400 vehicles passing a point every hour in one direction or every

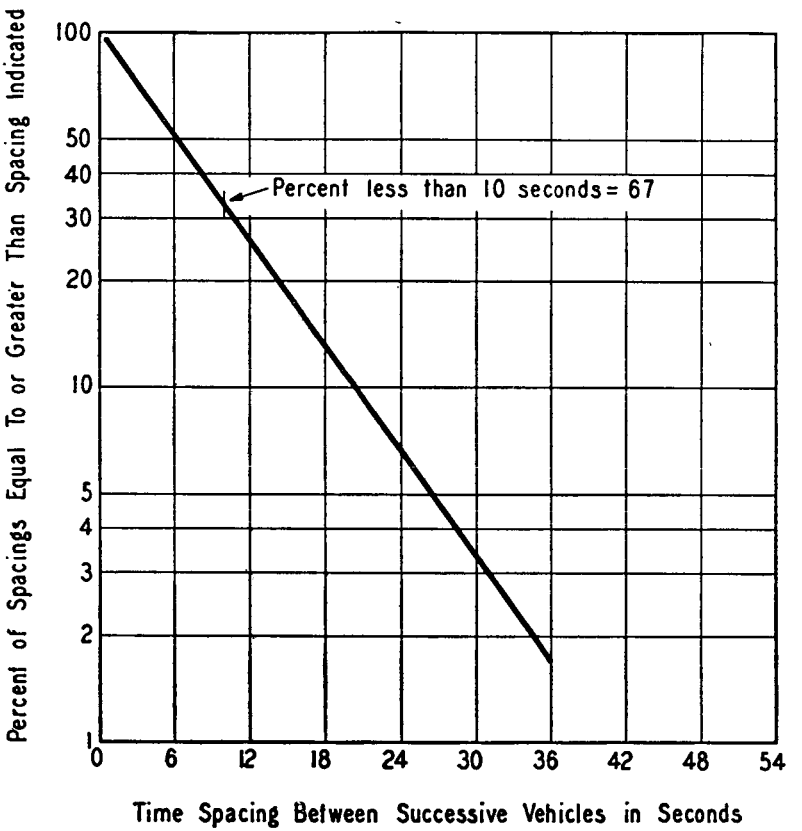


FIGURE V.16

CUMULATIVE DISTRIBUTION OF TIME SPACES ASSUMED FOR 2-LANE ROAD CARRYING 800 VEHICLES PER HOUR

3600 seconds) and that the minimum spacing is $^1/_2$ second. The curve for the distribution is shown in Figure V.16.

With 10 seconds as the average time required for passing we find from curve V.16. that 67 per cent of the spaces are too small for passing. This means that 67 per cent of the time a driver on this highway could not pass because of vehicles on the opposite lane.

This concept becomes clear if we keep in mind that at any instant the chance of there being a space of less than 10 seconds of free space on the opposite lane is equal to the percentage of the total spaces that are less than 10 seconds. In this sense the size of the time-gap has nothing to do with the chance of its being opposite the driver at any particular instant. It is only the frequency of the occurrence of the space that determines the probability of its happening in so far as passing is concerned. This reasoning becomes clearer if we remember that a space even if large is usually used for only one passing. For example 6 time spaces might occupy 50 seconds with one equal to 10 seconds to permit one passing or one of the spaces might be 25 seconds and still permit only one passing during the 50 seconds. (See Article V.23 for mathematical solution.)

If a driver is not to be retarded, he must every time he approaches a vehicle ahead, immediately pass the leading vehicle. If his speed is on the average 10 miles an hour faster, then that per cent of the time he cannot pass is the per cent of the 10 miles per hour difference that he must lose. In the present instance he would lose 67 per cent of 10 miles per hour or 6.7 miles per hour. Subtracting this from the 43 miles per hour average speed gives 36.3 miles per hour as the estimated average speed if the volume is 800 vehicles per hour for two lanes. This very nearly equals the observed speed of 36 miles per hour as shown in the lower curve, Figure 5, page 31, of the Capacity Manual. This result would indicate that this method of estimating is accurate enough to give good design figures. As a further check let us estimate the speed for 1200 vehicles per hour for two lanes. From the curve shown in Figure V.17. we find that vehicles are prevented from passing for 83 per cent of the time.

The speed drop is thus 83 per cent of 10 miles per hour = 8.3 miles per hour. Subtracting this from 43 = 34.7. This is more than

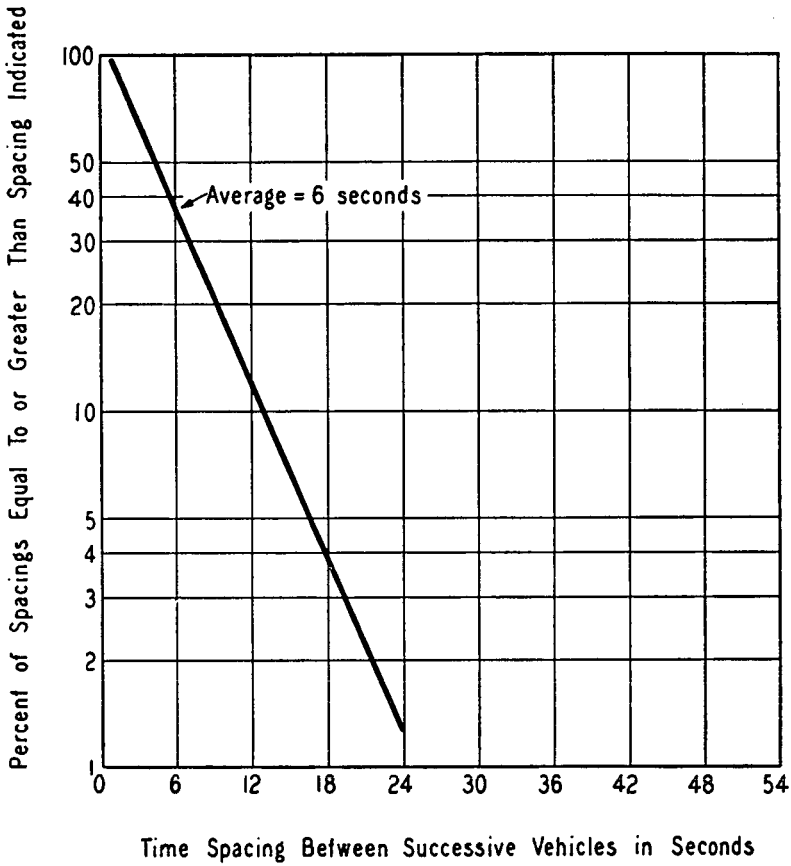


FIGURE V.17

CUMULATIVE DISTRIBUTION OF TIME SPACES ASSUMED FOR 2-LANE ROAD CARRYING 1200 VEHICLES PER HOUR

the observed results of about 32 miles per hour shown in Figure 5, page 31, of the *Manual*.

This lack of agreement needs to be examined to see if there is an explanation. According to the theory just advanced the speed drop due to inability to pass cannot exceed the average speed differential. How can we account for a speed drop greater than this ? The logical conclusion is that a further speed drop is not due to an inability to

pass but to some other cause. If we recall that there is a speed drop directly proportional to spacing the reason for the further speed loss becomes clear. With a volume of 1200 vehicles per hour, a high percentage of vehicles are traveling in the six second zone of mutual interference and are slowed because they are too close together rather than because of an inability to pass.

V. 18. *Estimate of Size Gap Required for Weaving.* It is impossible to estimate the speed drop for a given increase in volume on a four-lane road without knowing the time-gap required for weaving. But since the speed drop has been measured, it is possible, by reversing the method just explained, to estimate the time-gap for weaving.

From Figure 46, page 122, of the Capacity Manual, we find that at 1700 vehicles per hour, the distribution between lanes is equal. The speed on both lanes at this point should be the same. Referring to Figure 7, page 33, of the Capacity Manual, the speed at a flow of 1700 vehicles per hour is about 41 miles per hour. This is a drop of 7 miles per hour. Since the average speed differential is 8.8 miles per hour, in order for a speed decrease of 7 miles per hour to take place, on the average each car driver would be retarded $\frac{7}{8.8} = 79.5$ per cent of the time. This means that 79.5 per cent of the spaces on the adjoining lane are too small to permit weaving. From Figure 10, page 41, of the Capacity Manual, we find that the intersection of the 1700 vehicles per hour abscissa and the 79.5 per cent ordinate gives 3 seconds as about the time-gap required for weaving. This time-gap compares very closely indeed with the average weaving gap of 3 seconds as found by Wynn and Gourlay[10].

V. 19. *Physical Features of Highway: Effect on Traffic Flow.* Having discussed the interrelationships of the characteristics of flow, uninterrupted except by other traffic, the next step is to find what happens if the flow is slowed or interrupted by physical features of the highway. Let us first direct our attention to a location where passing is prohibited. This occurs in mountainous or hilly country where grades or restricted sight distances prevent passing.

For this problem assume that the average speed differential is

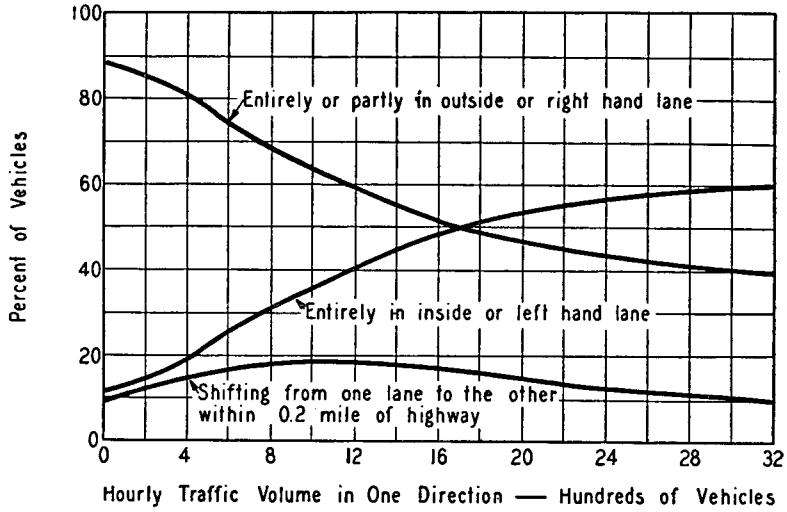


FIGURE V.18

DISTRIBUTION OF VEHICLES BETWEEN TRAFFIC LANES ON A 4-LANE HIGHWAY DURING VARIOUS HOURLY TRAFFIC VOLUMES

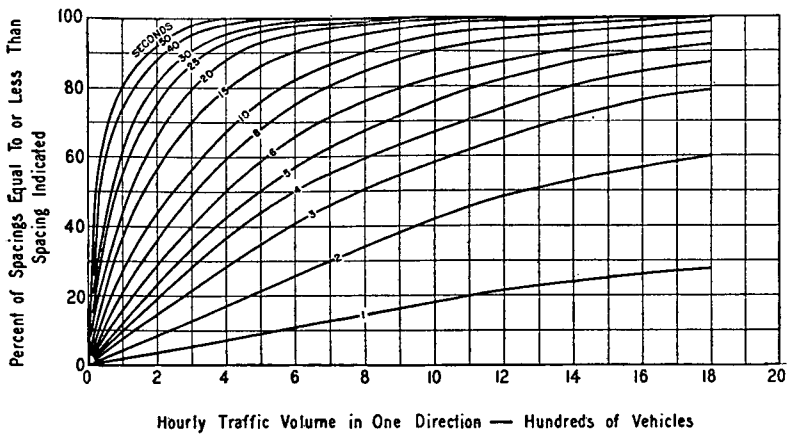


FIGURE V.19

FREQUENCY DISTRIBUTION OF TIME SPACING BETWEEN SUCCESSIVE VEHICLES TRAVELING IN THE SAME DIRECTION, AT VARIOUS TRAFFIC VOLUMES ON A TYPICAL 4-LANE RURAL HIGHWAY

(Figure 10, page 41, and Figure 46, page 122, "*Highway Capacity Manual*," Used by permission of Bureau of Public Roads, U.S. Department of Commerce.)

9 miles per hour and that it is required to estimate the time loss due to a stretch of highway where passing cannot take place for one half of the time. Let us further assume that the volume is 600 vehicles per hour. Reasoning as before, that a driver in order not to lose speed must be able to pass as soon as he approaches behind a slower vehicle, we conclude that for one half of the time he must sacrifice the speed differential between his own speed and that of the slower vehicle. Thus if the average speed differential is 9 miles per hour the speed loss in this case would be $\frac{1}{2} \times 9 = 4.5$ miles per hour. To this loss must be added the loss due to an inability to pass because of vehicles on the opposite lane. Proceding as before, for a volume of 600 vehicles per hour we find 17 per cent of the spaces are greater than the 10 seconds required for passing. This means that for 83 per cent of the time that there is sufficient sight distance to pass, the passing maneuver is prevented by traffic on the opposite lane. The additional speed loss is $0.83 \times 4.5 = 3.75$. Therefore, the total speed loss is equal to $4.5 + 3.75 = 8.25$ miles per hour.

V. 20. *Crossing Streams of Traffic.* The capacity of a highway or street is limited by delays at intersections. The basic condition, but not the simplest to analyze, may be thought of as the intersecting of 2 two-lane roads without any traffic control[7]. Each vehicle under such a condition crosses during a gap in the opposing stream of vehicles. The average minimum acceptable time gap has been measured and found to range from 4.6 to 6 seconds depending upon the type of intersection with the average being 4.8 seconds[11]. Mr. Raff calls this "minimum acceptable time-gap" a critical lag and correctly defines it as the size lag which has the property that the number of accepted lags shorter than L, the critical lag, is the same as the number of rejected lags longer than L. In other words, the acceptable time gap is just as likely to be accepted as it is to be rejected. The probability that it will be accepted is thus equal to $\frac{1}{2}$.

The chances of any single vehicle being delayed at an intersection can be deduced in the same manner as the delay in passing by saying that the chance of crossing depends upon the probability

of there being a time-gap of sufficient size at the instant the vehicle approaches the crossing. This probability depends upon the relative frequency of gaps and not upon their size. Thus if 75 per cent of the gaps are as large or larger than required for crossing, then the chance of being able to cross without delay is 75 per cent, and the chance of being delayed is 25 per cent. With this reasoning, and recalling the exponential law of distribution of time-gaps, the probability of being delayed would be

$$(1 - e^{-m}) = (1 - e^{-m}) \times 100 \text{ in per cent}$$

The probability of not being delayed would equal

$$e^{-m} = (e^{-m}) \times 100 \text{ in per cent}$$

where m is the average size of time-gap on the street being crossed.

This reasoning applies to single or "first-in-line" vehicles for a next-in-line vehicle has to wait for the first vehicle to clear and hence is delayed a longer time, or looking at it in a different way, has a greater chance of being delayed. This question of added delay will be considered later in Art. V.25. For an illustration let the traffic on the main highway be 400 vehicles per hour. The fact that it is moving in two directions is immaterial. For our purpose it may be considered to all be in one direction. The average spacing between vehicles on the main highway will be $\frac{3600}{400} = 9$ seconds. Since there are practically no spacings below $\frac{1}{2}$ second the distribution of spacings will be approximately that shown in Figure V.16. Recall that the average is at point .368 on the per cent ordinate. This curve shows that 52 per cent of the spaces are greater than 6 seconds and 48 per cent smaller.

V.21. *Mathematical Determination of Vehicle Delay Time.* The problem of determining the proportion of time that a vehicle is delayed may be approached by a more rigorous mathematical analysis. This problem along with other related problems has been solved by Mr. W. F. Adams in examples worked out in connection with his paper, "Road Traffic Considered as a Random Series."[12]

The proportion of time occupied by intervals greater than t seconds, according to Mr. Adams, is

$$e^{-Nt}(Nt + 1) \qquad \text{V.21.1.}$$

wherein N equals vehicles per second. The proof is as follows:

Consider the intervals of lengths lying between t and $t + dt$, and for the moment assume we are dealing with a period of one hour.

In one hour the expected number of intervals greater than t is,

$$Te^{-Nt}$$

T = vehicles per hour. This is basically the same as the formula, $100\, e^{-\frac{s}{M}}$, but with different notation.

Similarly, the expected number of intervals greater than $t + dt$ is

$$Te^{-N(t+dt)} = Te^{-(Nt+Ndt)}$$

$$= Te^{-Nt}\, e^{-Ndt} \text{ by the rule for addition of indices.}$$

The number of intervals of lengths between t and $t + dt$ is

$$= Te^{-Nt} - Te^{-Nt}e^{-Ndt} = Te^{-Nt}(1 - e^{-Ndt})$$

Expanding e^{-Ndt} in terms of Ndt,

$$= Te^{-Nt}(1 - 1 + Ndt - N^2dt^2/2! + N^3dt^3/3! \ldots.)$$

$$= Te^{-Nt}Ndt. \text{ Omitting terms in } dt^2 \text{ and higher powers,}$$

$$= TNe^{-Nt}dt$$

To the first order of small quantities, the length of all such intervals may be taken as t.

The time occupied by these intervals is therefore

$$TNte^{-Nt}dt \text{ seconds}$$

The time occupied by all intervals greater than t during one hour is found by integrating this expression between limits t and infinity,

$$= TN \int_t^\infty te^{-Nt}dt$$

Integrating by parts, $\int udv = uv - \int vdu$

Put $u = t$, $du = dt$, and $dv = e^{-Nt}dt$ so that

$$v = \int e^{-Nt}dt = -e^{-Nt}/N$$

The above expression then becomes

$$TN\left[-te^{-Nt}/N+\int e^{-Nt}dt/N\right]_t^\infty = TN\left[-te^{-Nt}/N-e^{-Nt}/N^2\right]_t^\infty$$

Both terms are zero when t is infinite, so that the number of seconds occupied by intervals over t seconds during one hour becomes

$$TN(te^{-Nt}/N+e^{-Nt}/N^2) = 3600\,N^2(te^{-Nt}/N+e^{-Nt}/N^2)$$
$$= 3600\,e^{-Nt}(Nt+1)$$

Now the total time considered is 3600 seconds, so that the proportion of time occupied by intervals over t seconds is

$$e^{-Nt}(Nt+1)$$

Conversely, the proportion of time occupied by intervals less than t is

$$1-e^{-Nt}(Nt+1) \qquad \text{V.21.2.}$$

V. 22. *Graphical Method of Determining Proportion of Time Occupied by Time-Gaps of Given Size.* The time occupied by time-gaps larger (or smaller) than any given value may be determined graphically. This is possible because we know that the average size gap in any range is always at .368 or the 36.8 percentile point of the range.

For the purpose of demonstration let it be required to find the proportion of time occupied by time-gaps larger than 6 seconds in a stream of traffic of 600 vehicles per hour. The average space is equal to $\frac{3600}{600} = 6$ seconds. This average is at the 36.8 percentile point so we may construct the curve $100\,e^{-\frac{S}{m}}$ which we have already discussed by selecting several values for S to get values for $\frac{S}{m}$ $(m = 6)$ to give points on the curve. The curve is shown in Figure V.20.

The average spacing is 6 seconds at 36.8 percentile point. The average for the spacings greater than 6 seconds is at the point 36.8 per cent of 36.8 per cent or 13.5 per cent. The corresponding spacing

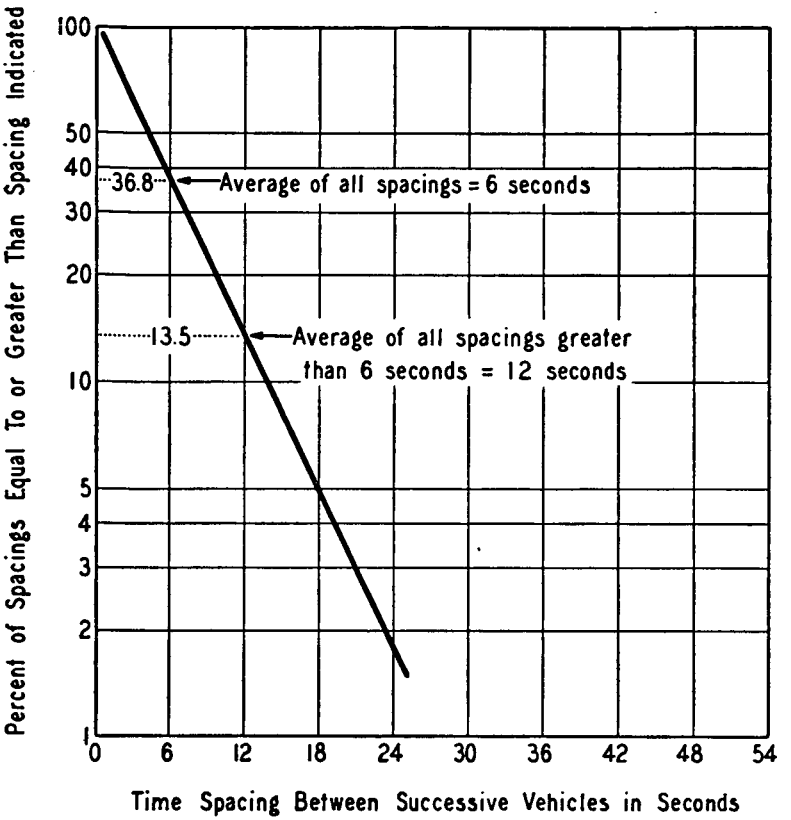


FIGURE V.20

CUMULATIVE DISTRIBUTION OF TIME SPACES ASSUMED FOR 2-LANE ROAD CARRYING 600 VEHICLES PER HOUR

is 12 seconds. Thus, the average of all spacings is 6 seconds and the average for the spacings above 6 seconds is 12 seconds. Therefore, the proportion of time occupied by spacings greater than 6 seconds is equal to

$$\frac{36.8 \text{ (per cent)} \times 12}{100 \text{ (per cent)} \times 6} = .736$$

$$= 73.6 \text{ per cent}$$

Using the formula $e^{-Nt}(Nt + 1)$; $N = \frac{1}{6}$, $t = 6$:

$$e^{-Nt}(Nt + 1) = e^{-1}(1 + 1) = .368 \times 2$$
$$= .736 = 73.6\%$$

V. 23. *The Average Length of All Intervals.* The average length of all intervals greater than t seconds is equal to the total time greater than t seconds divided by the number of intervals greater than t seconds, i. e.,

$$\frac{e^{-Nt}(Nt + 1)}{N e^{-Nt}} = \left(\frac{1}{N} + t\right) \text{seconds} \qquad \text{V.23.1.}$$

Conversely, the average length of interval less than t seconds is equal to the total time occupied by intervals less than t seconds divided by the number of intervals of less than t seconds, i. e.,

$$\frac{1 - e^{Nt}(Nt + 1)}{N(1 - e^{-Nt})}$$
$$= \frac{1 - Nt\,e^{-Nt} - e^{-Nt}}{N(1 - e^{-Nt})}$$
$$= \frac{1 - e^{-Nt}}{N(1 - e^{-Nt})} - \frac{Nte^{-Nt}}{N(1 - e^{-Nt})}$$
$$= \frac{1}{N} - \frac{t\,e^{-Nt}}{1 - e^{-Nt}} \qquad \text{V.23.2.}$$

Having determined the average length of intervals of less than t seconds it still remains to be found how much delay these intervals cause. The following solution is given by Mr. Adams:
Solution:

When any pedestrian or driver arrives, he may find

(a) that no vehicle arrives during the next t seconds. The probability of this is e^{-Nt} and in this case his waiting time is zero.
(b) that a vehicle arrives during the first t seconds, but none arrives in the t seconds following the arrival of the first vehicle. The probability of this is $(1 - e^{-Nt})\,e^{-Nt}$ and the waiting time is *one* interval.

(c) that the first two intervals after his arrival are each less than t seconds, but the third is greater than t. The probability is $(1-e^{-Nt})^2 e^{-Nt}$ and he has to wait for *two* intervals each less than t seconds.

In similar manner it may be shown that the probability of any driver or pedestrian having to wait for n intervals each less than t seconds is

$$(1-e^{-Nt})^n e^{-Nt}$$

The Expectation[a] of intervals for which the driver or pedestrian has to wait is given by the series

$$0\,e^{-Nt} + 1\,(1-e^{-Nt})\,e^{-Nt} + 2\,(1-e^{-Nt})^2\,e^{-Nt}\ldots$$
$$= e^{-Nt}\left\{1\,(1-e^{-Nt}) + 2\,(1-e^{-Nt})^2 + 3\,(1-e^{-Nt})^3\ldots\right\}$$

Summing the series in brackets to infinity[b] the expected number of intervals becomes

$$\frac{e^{-Nt}\,(1-e^{-Nt})}{(e^{-Nt})^2}$$
$$= \frac{1-e^{-Nt}}{e^{-Nt}} \qquad \text{V.23.3.}$$

The average length of the intervals of less than t seconds as already found is

$$\frac{1}{N} - \frac{te^{-Nt}}{1-e^{-Nt}} \text{ seconds.}$$

The average waiting time will be the product of the expected number of intervals and the average length of interval

$$= \frac{1-e^{-Nt}}{Ne^{-Nt}} - \frac{te^{-Nt}(1-e^{-Nt})}{e^{-Nt}(1-e^{-Nt})}$$
$$= \frac{1}{Ne^{-Nt}} - \frac{1}{N} - t \qquad \text{V.23.4.}$$

This is the average delay to all drivers or pedestrians, whether each one is delayed or not. However, a proportion e^{-Nt} of them find that the first vehicle does not arrive during the t seconds following their own arrival, so that this proportion of them is not delayed at all.

(a) The 'Expectation' of an event which may at each trial take any one of a number of possible values is found by multiplying each of the possible

The proportion delayed is therefore

$$(1 - e^{-Nt})$$

and the average waiting time of those who suffer delay is

$$\frac{1/Ne^{-Nt} - 1/N - t}{1 - e^{-Nt}}$$

$$= \frac{1}{Ne^{-Nt}} \cdot \frac{(1 - e^{-Nt})}{(1 - e^{-Nt})} - \frac{t}{(1 - e^{-Nt})}$$

$$= \frac{1}{Ne^{-Nt}} - \frac{t}{1 - e^{-Nt}} \quad \text{V.23.6.}$$

Mr. Warren S. Quimby[13] using the formula in a modified form, gives the delay as

$$\text{Delay} = \frac{3600}{ve^{-\frac{vt}{3600}}} - \frac{t}{1 - e^{-\frac{vt}{3600}}} \quad \text{V.23.7.}$$

wherein t = acceptable time gap in seconds
v = number of vehicles per lane per hour
e = base of Napierian logarithms = 2.71828.
3600 = number of seconds in one hour.

These delays are for a single vehicle approaching the intersections. Mr. Quimby gives a comparison of the theoretical delay with the observed delay in the following table:

values by the probability of its occurrence and summing the resultant products. It represents the average value to be expected from a large number of trials (Cf. Footnote b.)

(b) Put $(1 - e^{-Nt}) = a$ and note that a, being a probability, must be less than 1.

The series then becomes

$$a + 2a^2 + 3a^3 + 4a^4 + \ldots . na^n + \ldots .$$

The sum to infinity of this series (see Hall and Knight's "Higher Algebra" Chap. V., section 60, example 1) is

$$a/(1-a)^2 = (1 - e^{-Nt})/(e^{-Nt})^2$$

Table V.8

COMPARISON OF THEORETICAL AND FIELD DELAYS TO FIRST VEHICLE IN LINE

Sample	A	B	C	D	E	F
Theoretical delay, seconds	6.60	7.10	6.91	6.95	7.04	4.05
Actual delay, seconds	6.4	6.2	6.8	8.0	8.7	4.4

For determining the percentage of vehicles delayed, Mr. Quimby gives the following formula:

$$\text{Per cent delayed} = 1 - e^{-vt/3600} + (1 - e^{-vt/3600})T,$$

wherein the terms are as already defined with the exception of T which is the probability of a vehicle arriving in any given time interval.

Mr. Quimby states that this formula includes a consideration of both main and side street volumes and this is affected by a change in the volume on either street.

The following table compares the actual with the theoretical delay:

Table V.9

COMPARISON OF THEORETICAL AND FIELD OBSERVATIONS OF TOTAL TRAFFIC DELAYED

Sample	A	B	C	D	E	F
Main street volume	568	635	606	608	627	200
Side street volume	110	115	116	123	191	181
Per cent delayed – theory	55.3	60.7	58.7	59.3	65.9	16.0
Per cent delayed – actual	53.8	55.0	56.5	59.2	63.0	14.6

Another researcher to use a rational approach to this same problem is Mr. Morton S. Raff[11].

All cars are not "first-in-line" for often several vehicles are blocked so that there is a second, a third and so on, position car. He states that the percentage of vehicles delayed as given by the formula

$$P = 100\,(1 - e^{-NL})$$

is too small. This formula will again be recognized as the same one as just discussed but with a different notation. That is $NL = Nt$. In this formula N = number of vehicles on main street and L = the "lag." In order to take account of this sluggishness, Mr. Raff modifies the formula and arrives at the following:

$$P = 100\left\{1 - \frac{e^{-2.5N_s}e^{-2NL}}{1 - e^{-2.5N_s}(1 - e^{-NL})}\right\}$$

where

P = Percentage of side cars delayed
N = Main Street volume, in cars per second
N_s = Side-street volume, in cars per second
L = Critical lag in seconds
e = Base of natural logarithm

Mr. Raff states an examination shows that:

1. The limit of P, as N_s approaches zero, is $100(1 - e^{-NL})$, which is the theoretical formula. In other words, if there are no side-street cars, there is no sluggishness effect.
2. P always exceeds $100(1 - e^{-NL})$, except when N_s equals zero. In other words, the sluggishness effect delays more cars than would be delayed if it did not exist.
3. P is always less than 100 per cent, for any finite volume.
4. The partial derivatives of P with respect to N, N_s, and L are all positive. This means that an increase in either of the two volumes or the critical lag causes an increase in the percentage of cars delayed, as given by this formula.[11]

The coefficient of N_s has been found from observed delays to give values close to actual experimental results. For the theoretical development of the formula see Mr. Raff's book.

V. 24. *The Signalized Intersection.* The signalized intersection presents a problem that is different from that where there is no control or only a stop sign. The periods for crossing are at fixed intervals rather than at random as are the openings in an opposing stream of traffic. Since traffic is naturally distributed haphazardly, it follows that any fixed time signal causes unnecessary delay. The minimum delay follows the shortest timing interval that

will permit all the waiting vehicles to clear. This fact is easily comprehended if we think of a very long timing such as a 30 minute red followed by a 30 minute green signal. During the 30 minute green interval on one street there would be no delay but on the other street all traffic appearing at the intersection during the long interval would be blocked. The average wait would thus be about 15 minutes. Obviously, as the timing is decreased, the average waiting time decreases until such time as the traffic fails to clear during each signal change.

The two fundamental problems in signal control therefore are (1) finding the shortest timing that will not cause excessive failures to clear the waiting traffic and (2) determining the delay caused by the fixed timing.

Perhaps the method of determining the chances of signal failures to clear traffic may most easily be explained by means of an illustrative solution.[a]

Let it be required to find the probability of the cycle failure for 395 vehicles per hour on each lane with a 20 second green and a 20 second red signal cycle. Since observations have shown that usually slightly more than 20 seconds are required after the light changes to green for seven vehicles to enter the intersection, it will be assumed that the cycle will fail whenever seven or more vehicles appear in 40 seconds.

The average number of vehicles appearing in 40 sec. $= \frac{40 \times 395}{3600}$

$= 4 \cdot 4 = m$. With this value of m, the probability of seven or more vehicles appearing in 40 sec. (found from table) equals 15.63 per cent. Therefore, the traffic signal will fail to clear the waiting traffic 15.63 per cent of the time.

If it is desired to reduce the per cent of failures to say 5 per cent, it is only necessary to try a longer cycle. Two or three trials will usually give a result sufficiently close. The method is one of cut and try.

(a) This treatment of the signalized intersection is abstracted from: "Application of Statistical Sampling Methods to Traffic Performance at Urban Intersections" by Bruce D. Greenshields, (Proceedings of the Twenty-Sixth Annual Meeting), The Highway Research Board, December, 1946, pp. 377–389.

For a second trial, let us try a 25 second green — 25 second red cycle. The average number of vehicles appearing during the cycle of 50 seconds is $\frac{50 \times 395}{3600} = 5.5$ m. Since 10 vehicles will cause a failure, the percentage of the time that 10 or more will appear is read from the Poisson Table as .0537 or 5.37 per cent.

This is nearly the desired answer and serves to illustrate the procedure. If a more accurate result is wanted, another trial could be made.

Any signal failure will affect the chances of a succeeding failure since there will be vehicles left over from the first cycle. In the present example with a 20–20 signal, the second signal will fail if:

1. Seven vehicles arrive during the first and six or more during the second cycle.
2. Eight vehicles arrive during the first and five or more during the second cycle.
3. Nine vehicles arrive during the first and four or more during the second cycle.
4. Ten vehicles arrive during the first and three or more during the second cycle.
5. Eleven vehicles arrive during the first and two or more during the second cycle.
6. Twelve vehicles arrive during the first and one or more during the second cycle.

If the probabilities of the arrivals of the vehicles, as found in the Poisson tables, are multiplied together and added to give the total probability, the result is as follows:

1. $.0778 \times .2800 = .02178$
2. $.0428 \times .4488 = .01921$
3. $.0209 \times .6405 = .01338$
4. $.0092 \times .8149 = .00750$
5. $.0037 \times .9337 = .00345$
6. $.0013 \times .9877 = .00128$

$.06660$

This means that two signals will fail in succession 6.66 per cent of the time. In order to have three successive failures, there would need to be:

Thirteen vehicles in the first two cycles and six or more in the third,

Fourteen vehicles in the first two cycles and five or more in the third,

Fifteen vehicles in the first two cycles and four or more in the third, etc.

with the added condition that there be seven or more in the first cycle. While it is possible as just shown to compute the probabilities for these, it is cumbrous. Therefore a much less tedious method that gives results that agree closely with the more exact procedure will now be described.

In the example just given the two cycles would fail in succession if 13 or more vehicles appeared during the two cycles, provided that seven or more appeared in the first cycle.

The average number appearing in two cycles (80 secs.) equals $\frac{80 \times 395}{3600} = 8.8 = m$.

The probability of 13 or more appearing in the two cycles is .1102 as found in the Poisson tables (4 places is considered sufficient).

The average flow for the two failing cycles is not eight, the average flow on the roadway, but "13 or more vehicles". If it were known just how many vehicles "13 or more" amounts to it would be possible with this value of m to determine the probability of seven or more vehicles appearing in the first cycle. The next step is to find the mean value of "13 or more". Finding the arithmetical average requires extensive multiplication, but the mean value can be found very quickly. From the Poisson table it is found that the probability of:

13 or more vehicles appearing equals	0.1102
14 or more vehicles appearing equals	.0642
15 or more vehicles appearing equals	.0353
16 or more vehicles appearing equals	.0184

17 or more vehicles appearing equals .0091
18 or more vehicles appearing equals .0043
19 or more vehicles appearing equals .0019
20 or more vehicles appearing equals .0008

The mean of .1102 (the probability of 13 or more vehicles appearing) is .0551. According to the Poisson table above the number of vehicles corresponding to .0550 falls between 14 and 15. The values from the table above are plotted on semi-log paper.

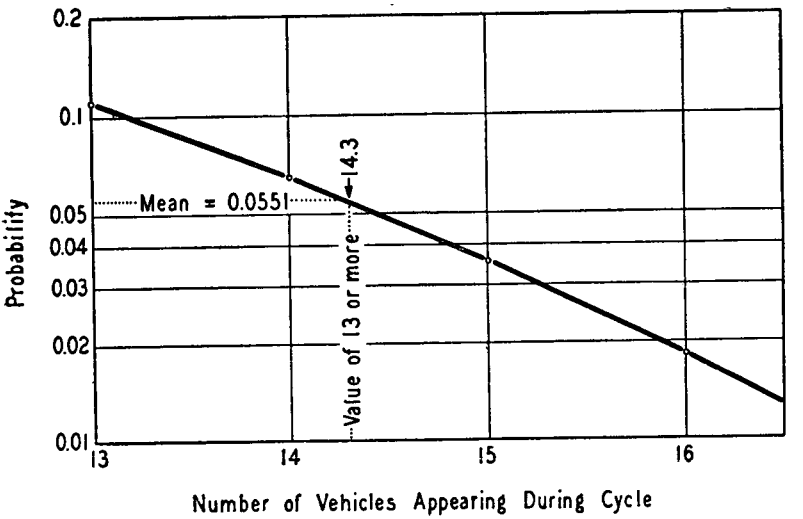


FIGURE V.21

PROBABILITIES ACCORDING TO POISSON DISTRIBUTION OF VARIOUS NUMBERS OF VEHICLES APPEARING AT AN INTERSECTION DURING ONE SIGNAL CYCLE

Note that the points fall on a nearly straight line. This fact makes it possible to interpolate between 14 and 15. The number of vehicles shown on the abscissa corresponding to 0.0551 is equal to approximately 14.3 which is the mean of "13 or more" for the two cycles or approximately 7.15 for one cycle. With this new m the probability of seven or more vehicles appearing in the first cycle is equal to 0.5939.

The probability of the two cycles failing is equal to the probability of there being 13 or more in the two cycles multiplied by the probability of there being seven or more in the first cycle or 0.1102 $\times$ .5939 = 0.0654. This may be compared with the correct value of .0666.

The probability of three cycles failing in succession would be equal to the probability of 19 or more vehicles appearing in three cycles times the probability of 13 or more in two cycles (with m equal to $\frac{19}{3}$), times the probability of seven or more in the first cycle.

V. 25. *Calculating Delay at Signalized Intersections.* It is possible to calculate the delay at a signalized intersection by first finding the probability of retarding 1, 2, 3....n vehicles, and then computing the average delay for the first, second, third, etc. vehicles in line. The theoretical method of doing this is explained in "Traffic Performance at Urban Street Intersections",[7] pages 91–94, but the procedure is too tedious to be practical. A method that is practical is described in this same reference pages 95–97, and 100.

V. 26. *Practical Method for Determining Number of Vehicles Retarded at the Signalized Intersection*: Before determining the delay per light cycle, it is necessary to ascertain the number of vehicles retarded. The proportion of vehicles retarded is greater than the proportion of the red signal to the entire cycle, since each retarded vehicle in effect increases the blocking period. The exact extent to which this occurs has been measured.

For the first vehicle to arrive at the intersection the potential blocking period is equal to the red interval R of the signal, though it may not experience the full potential if it arrives after the beginning of the red interval. The second vehicle, if it is not stopped, may not follow closer on the average than 1.7 seconds behind the first vehicle which enters 3.8 seconds after the light changes to green. The blocking period for the second vehicle therefore is

$$R + 3.8 + 1.7 = R + 5.5 \text{ seconds.}$$

The second vehicle enters 3.1 seconds after the first, so that the potential blocking period for the third vehicle becomes

$$R + 3.8 + 3.1 + 1.7 = R + 8.6 \text{ seconds.}$$

Similarly the potential blocking period for the fourth vehicle equals $R + 3.8 + 3.1 + 2.7 + 1.7 = R + 11.3$ seconds

In general, the potential blocking period is obtained by adding to the signal interval the additional delay interval caused by the preceding vehicles plus 1.7 seconds.

The additional blocking periods created when various number of vehicles are retarded is shown in Figure V.22 taken from page 96 of Traffic Performance at Urban Street Intersections.[7]

As an illustrative example, let it be required to find the average number of vehicles retarded for a traffic volume of 228 vehicles per hour on a single lane with the signal set for 30 second *go* and 20 second *stop*. The average number of vehicles arriving during the 20 second red period is 1.27 vehicles $[(20 \times 228)/3600]$. (This might be approximately one for each of three cycles and two for the fourth cycle.) As explained, these 1.27 vehicles tend to increase the effective length of the red signal. Reference to Figure V. 22. shows that 1.27 vehicles increase the blocking period by about 6.4 seconds. The blocking period may now be considered to be 26.4 seconds (20 + 6.4). A 26.4 second blocking period, however, will retard about 1.67 vehicles, $[(26.4 \times 228)/3600]$.

The increase of the blocking period due to 1.67 vehicles is 7.7 seconds and the blocking period is now estimated to be 27.7 seconds. During the 27.7 seconds of blocking period 1.75 vehicles will be retarded to increase the estimate of the blocking period to 27.95 seconds. By further successive approximation, the number of vehicles retarded can be obtained with any degree of accuracy desired. This information may be shown in tabular form:

Table V.10. Average Number of Vehicles Stopped with 228 Vehicles per Hour per Lane and 20 Second Red Period

	Length of Blocking Period	*Average No. of Vehicles Retarded*
1st Approximation	20 seconds	1.27
2nd ,,	26.4 ,,	1.67
3rd ,,	27.7 ,,	1.75
4th ,,	27.95 ,,	1.77
5th ,,	28 ,,	1.77

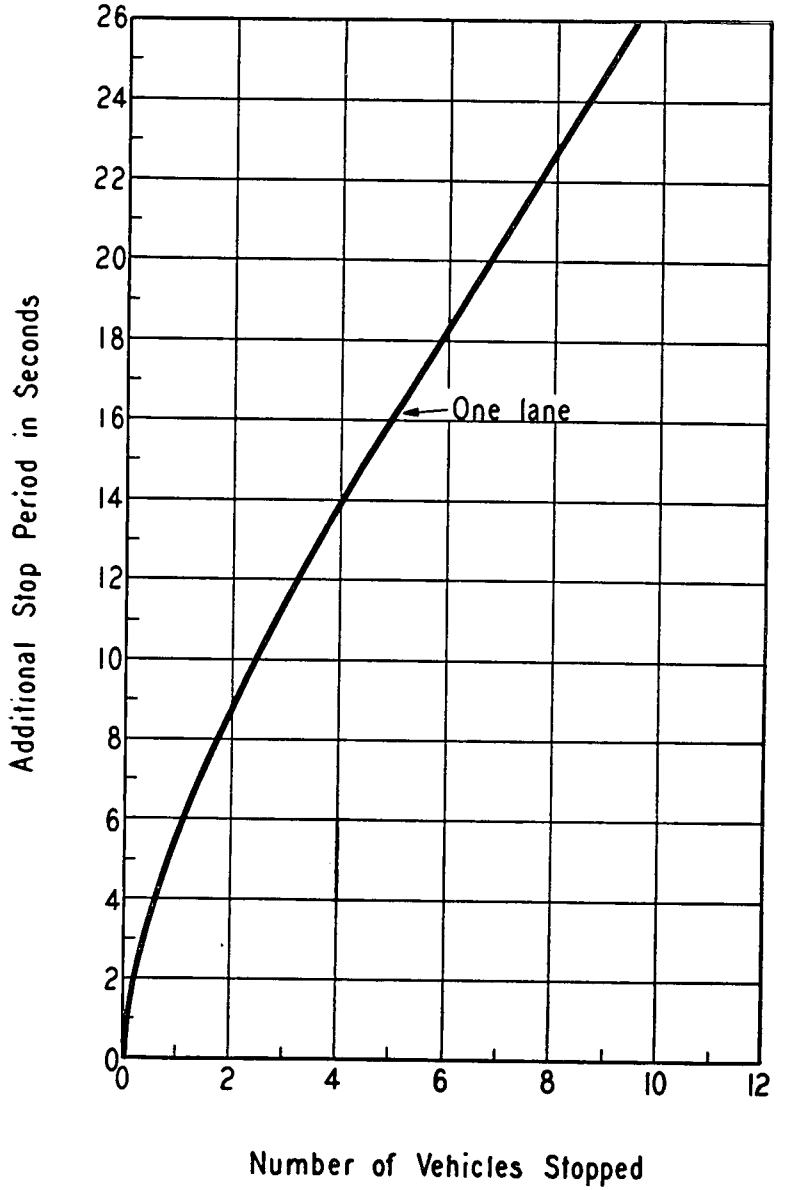


FIGURE V.22. ADDITIONAL BLOCKING PERIODS CREATED WHEN VARIOUS NUMBERS OF VEHICLES ARE RETARDED

For this particular example it seems sufficiently accurate to use an average of 1.77 vehicles per red signal. This shows that with a volume of 228 vehicles per hour per lane a 20 second red interval becomes, in effect, a 28 second blocking period.

V.27. *The Average Arrival Method of Determining Delay.* A practical method of calculating the time loss for a given number of vehicles stopped is based upon an assumption as to the arrival time of the first vehicle. The method may be illustrated as follows:

Let the *red* interval be 30 seconds. It is assumed that the first vehicle will arrive on the average at the mid-point, wait 15 seconds, and it will lose 3.8 seconds in entering the intersection. To this is added another two seconds lost in accelerating to a speed of 15 miles an hour, giving a total loss of 20.8 seconds. (The acceleration loss would be greater for higher speeds). The total loss (using symbols) is

$$\frac{R}{2} + 3.8 + a$$

wherein R equals the *red* interval and a the acceleration loss for a given normal traveling speed. The second vehicle arrives on the average at the mid-point of the stop period of R + 5.5, and leaves at R + 6.9. The time loss is equal to

$$R + 6.9 - \frac{(R + 5.5)}{2} + 1 = 20.15 \text{ seconds}$$

wherein 1 is a the acceleration loss.
The loss for the third vehicle is:

$$R + 9.6 - \frac{(R + 3.8 + 3.1 + 1.7)}{2} + a$$

$$= 39.6 - \frac{(30 + 3.8 + 3.1 + 1.7)}{2} + 1 = 21.3 \text{ seconds}$$

The loss for the fourth vehicle is:

$$R + 12 - \frac{(R + 9.6 + 1.7)}{2} = 21.35 \text{ seconds.}$$

No acceleration loss is added for the fourth vehicle since it has reached normal speed by the time it enters the intersection.

By following this method the delay for any number of vehicles retarded may be calculated, but it is only the method that is of interest to us here. According to the reference just mentioned the observed delay agrees very closely with that calculated. The delay occurring in traffic with various proportions of trucks, street cars, and other types of vehicles needs to be observed to obtain more accurate and representative field constants.

V. 28. *Rare Events (Accidents).* There are many events in traffic that are comparatively rare. This is particularly true of certain types of accidents. Taken as a whole, traffic accidents exact a high toll in lives and property but the average driver is rarely involved in a serious mishap. Problems involving rare events may be analyzed by the Poisson distribution which is also known as the law of small chances.

One study that made use of the law was conducted by Dr. H.M. Johnson[14]. He examined the accident histories of 29,531 Connecti-

Table V.11

ACTUAL AND EXPECTED DISTRIBUTION OF ACCIDENTS, INCLUDING CASUALTIES AND PROPERTY DAMAGE EXCEEDING $25, REPORTED TO THE COMMISSIONER OF MOTOR VEHICLES OF CONNECTICUT, 1931–36, IN A LICENSED DRIVER SAMPLE SELECTED AT RANDOM.

Accidents per operator during experience	*Operators having these accidents*		
	Actual number	*Expected number*	*Difference*
0	23,881	23,234	647
1	4,503	5,572	—1,069
2	936	668	268
3	160	53	107
4	33	4	47
5	14		
6	3		
7	1		
Totals	29,531	29,531	0

Note: The probability that the differences between the actual and expected distributions is due to chance $= 1.6\,(10)^{-161}$, which is insignificant.

cut drivers selected at random, each of whom had been licensed for the period 1931–1936.

Among these 29,531 drivers there accrued 7,082 accidents which involved 5,650 operators, Mr. Johnson found that the accidents were not distributed among the drivers according to the law of chances for which the sole parameter is the rate per operator. He, therefore, concluded that some operators were accident prone for some reason that could only be determined experimentally.

The table shows the actual accidents, the expected number as calculated from the Poisson distribution and the difference between the theoretical and the actual number.

It may be noted that there are more accident-free drivers than accounted for by the laws of chance and also more repeaters with a corresponding deficiency of drivers having a moderate accident rate.

Mr. Johnson found among other things, that drivers who were under 16–20 years old at the beginning of the experience and under 22–27 years old at its close had 1.47 times as many of the non-personal accidents as they would have if the distribution of accidents were independent of age. That this difference is not accidental, according to Mr. Johnson, is evidenced by the fact that the probability of the independency-hypothesis being true is less than 10^{-24}.

The significance of Mr. Johnson's report is that it demonstrates the use of the Poisson distribution in studying rare events. Suppose that one wishes to know whether a driver having 3 accidents in 6 years is an accident-prone driver. According to Mr. Johnson's figures the average for all drivers is

$$\frac{7082}{29531} = .2398 = .24 \text{ accidents} = \mathbf{m}.$$

With this value of m we find from a Poisson distribution table that the probability of a driver having 3 accidents is .0018 or .18 per cent. This means that the chances are 100 to .18 or approximately 550 to 1 against an average driver's having 3 accidents. We may conclude, therefore, that a driver who has this many mishaps is a bad risk.

V. 29. *Rare Events (Accidents at Intersections).* Washington, D. C. has a total of 7,683 intersections open to traffic. During the year 1950 there were 6,211 accidents at intersections. Suppose it is desired to know how many accidents at an intersection make it accident prone.

The average number of accidents $= \frac{6211}{7683} = .8 = \text{m}$. According to the Poisson distribution, the probabilities of accidents occurring at an intersection are as follows:

Table V.12

Number of Accidents	*Probability*
2	.0438
3	.0383
4	.0077
5	.0012
3 or more	.0474
4 or more	.0091
5 or more	.0014

Suppose that it is decided that when the odds are 20 to 1 that the accidents occurring are not due to chance alone, an intersection is to be considered accident prone. According to the table, 3 or more accidents will occur due to chance 4.74 per cent of the time. This ratio of one to .0474 is over 20 to 1, hence an intersection having over 3 accidents would be considered unduly hazardous.

Records are not available as to the distribution of intersections having less than 5 accidents, but of those with five or more it is possible to compare the actual occurrence of accidents with the number expected to occur according to the Poisson distribution. See Table V. 13.

This procedure is presented to illustrate a method of approach and not as a suggested analysis, for obviously the records should be much more complete. Clearly the volume of traffic is one of the most important, if not the most important, factor.

V. 30. *Size of Sample to Determine Average Number of Car Passengers.* In making a traffic survey it is required to know the average number of persons per car. The problem is to determine the size

Table V.13. NUMBER OF INTERSECTIONS IN WASHINGTON, D.C. AT WHICH 5 OR MORE ACCIDENTS OCCURRED IN 1950

Number of Intersections having Accidents	*Number of Accidents Per Intersection*	*Total Number of Accidents*	*Number of Intersections Expected to have Number of Accidents Shown in Col. 2*
85	5	425	27
68	6	408	40
76	7	532	50
55	8	440	55
22	9	198	54
32	10	320	47
12	11	132	38
10	12	120	28
7	13	91	19
5	14	70	12
9	15	135	7
4	16	64	4
4	17	68	2
4	18	72	1
3	19	57	Less than 1
5	20	100	
2	21	42	
1	22	22	
1	23	23	
1	27	27	
1	28	28	
1	32	32	
1	37	37	
1	45	45	
1	64	64	
1	86	86	
412		3638	

Note: In this case, $m = \frac{3638}{412} = 8.8$. The last column, *Number of Intersections Expected to have Number of accidents shown in Column* 2, can be obtained by multiplying the probabilities of occurrence taken directly from "*Poisson Exponential Binomial Limits*,"[9] by 412, the total number of intersections. It may also be obtained from Appendix Table No. VI, page 226. This table gives the probability of *x or more* events occurring during a given interval, when *m*, the average number of events per interval is known. In using Table VI, the probability that x, a specific number of events will occur, is equal to the difference between the probabilities of *x or more* and *(x + 1) or more* events occurring. In the above table, the pure chance probability of 5 accidents occurring at an intersection is the difference in probability of 5 or more and 6 or more accidents occurring. Multiplying this difference by the total number of intersections gives the number of intersections expected to have 5 accidents. Referring again to Table VI, 0.872 (the probability that 6 or more accidents will take place) subtracted from 0.938 (the probability that 5 or more accidents will take place) leaves 0.066 or 6.6%. Multiplying 412 by 6.6% gives 27, the number of intersections that may be expected to have 5 accidents.

of sample to give a 95 per cent assurance that the mean value will not be in error more than 0.1.

Suppose that the following typical occupancy count has been made:

Occupants (x)	Number of Observations (f)
1	15
2	10
3	4
4	2
5	1
Mean $= \bar{X} = 1.9$	$N = 32$

The standard deviation s is first calculated and found to be 1.054.

From formula IV.7.3.

$$\frac{N-1}{t^2} = \frac{s^2}{\epsilon^2} = \frac{(1.054)^2}{(.1)^2} = \frac{1.11}{.01} = 111$$

From Appendix Table 3, Ratio of Degrees of Freedom to (t^2), we find that with a probability level of 5 per cent (95 per cent assurance) that for $N - 1 = 400$, that $\frac{s^2}{\epsilon^2} = 103.069$ and for $N - 1 = 500$, $\frac{s^2}{\epsilon^2} = 128.836$. Since 111 lies between these two values we conclude that the size of sample required is between 400 and 500, and if we wish to be conservative we take the higher value. Also it would have been better to have taken a larger (preliminary) sample to obtain the trial standard deviation.

V. 31. *Size of Sample Required in Speed Study.* It is desired to know the average speed on each block within one mile per hour on a street with 60 intersections. It is also desired that there be a 95 per cent assurance as to the result. It is assumed that the speed will vary with the volume of traffic, the weather, the amount of parking, and perhaps other conditions. The problem is to find the required size of sample and, having determined this, to recommend a method of making the observations that will yield a truly random sample.

The logical procedure is to take a random sample of about 100 observations in order to obtain an estimated standard deviation to be used in determining the size of sample. Suppose from this sample that it is found that the speed range is from 5 to 40 miles per hour and that the standard deviation, s, equals 4.5 miles per hour.

We use the t-distribution to find the size of sample. From formula IV.7.3.

$$\frac{N-1}{t^2} = \frac{s^2}{\epsilon^2}$$

we find the ratio of $N - 1$ to t^2 by inserting the values for s and ϵ. The standard deviation s in the present example, as found from the preliminary sample, is 4.5 miles per hour and the allowable error is one mile per hour.

$$\text{Hence,} \frac{N-1}{t^2} = \frac{s^2}{\epsilon^2} = \frac{(4.5)^2}{1^2} = \frac{20.25}{1} = 20.25$$

From a table of ratio of degrees of freedom to t^2 we find that with a probability level of 5 per cent that a ratio of $\frac{N-1}{t^2} = 20.202$ corresponds to $N - 1 = 80$ and 22.727 corresponds to $N - 1 = 90$. Therefore, we conclude that N, the size of sample, lies between 81 and 91. To be on the safe side, we may say that a sample of 100 observations will give us at least a 95 per cent assurance that the average speed will be obtained within $\pm$ 1 mile per hour. If a 99 per cent assurance is desired the size of sample according to the table would be between 100 and 200.

The next phase of the problem is that of getting a truly random sample. Obviously taking all the speeds on a day of light traffic would give a biased result. Clearly there must be some knowledge of the relative duration of the various conditions that influence speeds. Increasing the size of the sample so that observations might be distributed over a greater number of hours of the day, more days of the week and more months of the year would assure a better estimate of the speed. Increasing the size of sample to 200 should give sufficient coverage.

Since the speed is desired for each block it is necessary that observations be taken in each block. Some accurate mechanical device that is free from human errors is always preferable. This, however, would require either 60 recording devices or a rotation of a lesser number. Since they would give "spot" checks they would also need to be rotated to different positions in the blocks.

Another way would be to have an observer's car "float" with the traffic. The observer as well as recording speed could also note pertinent information such as the amount of parking. Manual recording could be supplemented or replaced by some mechanical device such as taking a picture of the conditions in each block and including in the picture a clock to show the time of reaching each intersection. The cost of such pictures taken on 16 mm film would be negligible.

The particular method to be employed in this or any other problem involving the collection and analysis of data should be selected by the engineer in charge after he has made a preliminary study of both the nature of the data and the reliability and cost of the various possible methods of conducting the field study. Statistics is merely an aid to the engineer and not a substitute for experience and judgment.

REFERENCES, CHAPTER V

[1] "*Highway Capacity Manual*," Committee on Highway Capacity, Department of Traffic and Operation, Highway Research Board, Washington, D.C., 1950.

[2] Hess, Dr. Victor F., "*The Capacity of a Highway*," Traffic Engineering, Institute of Traffic Engineers, New Haven, Connecticut, August 1950, page 420.

[3] Greenshields, Bruce D., "*The Photographic Method of Studying Traffic Behavior*," Proceedings, Highway Research Board, Washington, D.C., 1933.

[4] Ibid., "*A Study of Traffic Capacity*," Proceedings, Highway Research Board, Washington, D.C., 1933.

[5] Ibid., "*Initial Traffic Interferences*," Presented for discussion at the 16th Annual Meeting of the Highway Research Board, November 19, 1936, Washington, D.C., 9 pages mimeo and the comments by W. F. Adams.

[6] Ibid., "*Distance and Time Required to Overtake and Pass Cars*," Proceedings, Highway Research Board, Washington, D.C., 1935, pages 332–342.

[7] Ibid., Schapiro, Donald; Ericksen, Elroy L. "*Traffic Performance at Urban Street Intersections*," Yale University, Bureau of Highway Traffic, New Haven, Connecticut, 1947.

[8] "*Digest of the Application of Theory of Probability to Problems of Highway Traffic*," Proceedings, Institute of Traffic Engineers, New Haven, Connecticut, 1934, pages 118–123.

[9] Molina, E. C., "*Poisson Exponential Binomial Limits*," (Table) D. Van Nostrand Co., New York, 1942.

[10] Wynn, Houston F.; Gourlay, Stewart M.; and Strickland, Richard, I., "*Study of Weaving and Merging Traffic*," Technical Report No. 4, Yale University, Bureau of Highway Traffic, New Haven, Connecticut.

[11] Raff, Morton S., and Hart, Jack W., "*A Volume Warrant for Urban Stop Signs*," Eno Foundation for Highway Traffic Control, Inc., Saugatuck, Connecticut, 1950.

[12] Adams, W. F., "*Road Traffic Considered as a Random Series*," Journal of the Institute of Civil Engineers, London, 1936.

[13] Quimby, Warren S., "*Behavior Patterns for Merging Traffic*," Student Thesis, Yale University, Bureau of Highway Traffic, New Haven, Connecticut, 1949, page 40.

[14] Johnson, Dr. H. M., "*The Detection of Accident-Prone Drivers*," Proceedings, Highway Research Board, Washington, D.C., 1937, pages 444–454.

APPENDIX

APPENDIX Table I

Areas Under the Normal Probability Curve

From the Mean to Distances $\frac{x}{\sigma}$ from the Mean, Expressed as Decimal Fractions of the Total Area 1.0000

The proportional part of the curve included between an ordinate erected at the mean and an ordinate erected at any given value on the X axis can be read from the table by determining x (the deviation of the given value from the mean) and computing $\frac{x}{\sigma}$. Thus if $\bar{X}$ = \$25.00, σ = \$4.00, and it is desired to ascertain the proportion of the area under the curve between ordinates erected at the mean and at \$20.00; x = \$5.00 and $\frac{x}{\sigma} = \frac{\$5.00}{\$4.00}$ = 1.25. From the table it is found that .3944, or 39.44 per cent, of the entire area is included.

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0159	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3718	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4083	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4430	.4441
1.6	.4452	.4463	.4474	.4485	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4758	.4762	.4767
2.0	.4773	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4865	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4980	.4980	.4981
2.9	.4981	.4982	.4983	.4984	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.49865	.4987	.4987	.4988	.4988	.4988	.4989	.4989	.4989	.4990
3.1	.49903	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993129									
3.3	.4995166									
3.4	.4996631									
3.5	.4997674									
3.6	.4998409									
3.7	.4998922									
3.8	.4999277									
3.9	.4999519									
4.0	.4999683									
4.5	.4999966									
5.0	.4999997133									

Used by permission of Houghton Mifflin Company, publishers of Rugg's "*Statistical Methods Applied to Education*".

APPENDIX Table II

Table of Values of t

For Given Degrees of Freedom (n) and at Specified Levels of Significance (P)

In the use of this table it is to be remembered that a level of significance refers to both tails of the distribution. Thus, the .02 level ($P = .02$) includes .01 of the area of the curve in each tail. It is to be observed that this table is set up in a different form from the table of normal curve areas, Appendix Table I. The table of normal curve areas showed values of $\frac{x}{\sigma}$ in the margins and proportionate areas from $\overline{X}$ to $\frac{x}{\sigma}$ (one direction only) in the body. A tail of the normal distribution is obtained by subtracting this value from .5000. Doubling the resulting figure yields the level of significance. The t table, on the other hand, shows n (degrees of freedom) in the stub, t in the body, and P (the level of significance) in the caption. The last row of the t table, for $N = \infty$, shows t values as obtained from the normal curve.

n	Level of Significance (P)												
	.9	.8	.7	.6	.5	.4	.3	.2	.1	.05	.02	.01	.001
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.126	.255	.388	.529	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	.126	.254	.387	.527	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	.126	.254	.386	.526	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	.126	.253	.385	.524	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Appendix Table II is reprinted from Fisher and Yates: "*Statistical Tables for Biological, Agricultural, and Medical Research*", published by Oliver and Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

APPENDIX

Table III

Ratio of Degrees of Freedom to $(t)^2$

Degrees of Freedom	*Probability Level*		
	5%	2%	1%
1	0.006	0.001	0.0002
2	0.108	0.041	0.020
3	0.296	0.145	0.088
4	0.519	0.285	0.189
5	0.756	0.442	0.308
6	1.002	0.607	0.437
7	1.252	0.778	0.572
8	1.504	0.954	0.711
9	1.759	1.131	0.852
10	2.015	1.309	0.996
11	2.271	1.489	1.140
12	2.527	1.670	1.286
13	2.786	1.851	1.433
14	3.043	2.033	1.580
15	3.303	2.216	1.727
16	3.560	2.398	1.875
17	3.818	2.580	2.024
18	4.078	2.764	2.173
19	4.337	2.947	2.321
20	4.596	3.130	2.471
21	4.854	3.312	2.620
22	5.115	3.498	2.768
23	5.373	3.680	2.919
24	5.634	3.865	3.068
25	5.891	4.048	3.219
26	6.151	4.231	3.367
27	6.412	4.415	3.516
28	6.676	4.601	3.668
29	6.934	4.784	3.818
30	7.195	4.969	3.967
40	9.803	6.813	5.447
60	15.000	10.504	8.480
120	30.596	21.582	17.523

APPENDIX Table IV

Values of χ^2

For Given Degrees of Freedom (n) and for Specified Values of P

n	Value of P														
	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.00
1	.000157	.000628	.00393	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341	16.268
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.465
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.517
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.302
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703

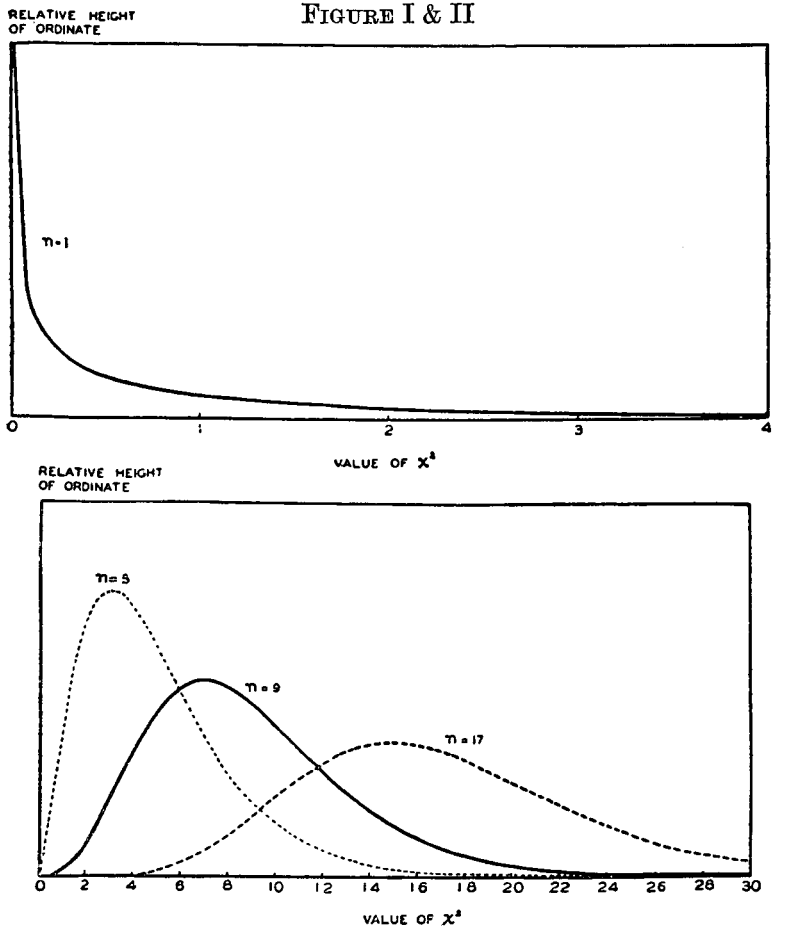
For large values of n compute $\sqrt{2\chi^2}$, the distribution of which is approximately normal around a mean of $\sqrt{2n-1}$ with $\sigma = 1$. P is the ratio of one tail of the normal distribution to the area under the entire curve.

A detailed table of the probability of various values of χ^2 for one degree of freedom is given in G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, 11th edition, pp. 534–535, Charles Griffin and Co., London, 1937.

Appendix Table IV is reprinted from Fisher and Yates: "*Statistical Tables for Biological, Agricultural, and Medical Research*", published by Oliver and Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

APPENDIX

FIGURE I & II



Distribution of χ^2 for $n = 1$, $n = 5$, $n = 9$, and $n = 17$. The maximum ordinate is at $\chi^2 = n - 2$ except when $n = 1$. When $n = 1$, the maximum ordinate is at $\chi^2 = 0$. When $n = 1$, there is 4.55 per cent of the curve beyond $\chi^2 = 4$. Beyond $\chi^2 = 30$ there is .0015 of one per cent of the curve when $n = 5$; .0439 of one per cent of the curve when $n = 9$; 2.6345 per cent of the curve when $n = 17$. The two charts have been drawn to different scales. If the vertical axis of the upper chart is expanded to approximately 20 times its length and the horizontal axis is contracted to about one-eighth of its length, the curves will be roughly comparable as to area.

APPEN-

5 % and 1 % Points for Distribution of F.

n_2	n_1 degrees of freedom (for greater mean square)											
	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
	4,052	**4,999**	**5,403**	**5,625**	**5,764**	**5,859**	**5,928**	**5,981**	**6,022**	**6,056**	**6,082**	**6,106**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.34**	**99.36**	**99.38**	**99.40**	**99.41**	**99.42**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.27**	**10.15**	**10.05**	**9.96**	**9.89**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.59**	**3.52**	**3.45**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**

The function, $F = e$ with exponent $2z$, is computed in part from Fisher's table VI (7). Ad-

DIX Table V

(5 % in Roman Type, 1 % in Bold Face Type).

n_1 degrees of freedom (for greater mean square)												n_2
14	16	20	24	30	40	50	75	100	200	500	∞	
245	246	248	249	250	251	252	253	253	254	254	254	1
6,142	**6,169**	**6,208**	**6,234**	**6,258**	**6,286**	**6,302**	**6,323**	**6.334**	**6,352**	**6,361**	**6,366**	
19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50	2
99.43	**99.44**	**99.45**	**99.46**	**99.47**	**99.48**	**99.48**	**99.49**	**99.49**	**99.49**	**99.50**	**99.50**	
8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53	3
26.92	**26.83**	**26.69**	**26.60**	**26.50**	**26.41**	**26.35**	**26.27**	**26.23**	**26.18**	**26.14**	**26.12**	
5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63	4
14.24	**14.15**	**14.02**	**13.93**	**13.83**	**13.74**	**13.69**	**13.61**	**13.57**	**13.52**	**13.48**	**13.46**	
4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36	5
9.77	**9.68**	**9.55**	**9.47**	**9.38**	**9.29**	**9.24**	**9.17**	**9.13**	**9.07**	**9.04**	**9.02**	
3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67	6
7.60	**7.52**	**7.39**	**7.31**	**7.23**	**7.14**	**7.09**	**7.02**	**6.99**	**6.94**	**6.90**	**6.88**	
3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24		7
6.35	**6.27**	**6.15**	**6.07**	**5.98**	**5.90**	**5.85**	**5.78**	**5.75**	**5.70**	**5.67**	**5.65**	
3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93	8
5.56	**5.48**	**5.36**	**5.28**	**5.20**	**5.11**	**5.06**	**5.00**	**4.96**	**4.91**	**4.88**	**4.86**	
3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71	9
5.00	**4.92**	**4.80**	**4.73**	**4.64**	**4.56**	**4.51**	**4.45**	**4.41**	**4.36**	**4.33**	**4.31**	
2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54	10
4.60	**4.52**	**4.41**	**4.33**	**4.25**	**4.17**	**4.12**	**4.05**	**4.01**	**3.96**	**3.93**	**3.91**	
2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40	11
4.29	**4.21**	**4.10**	**4.02**	**3.94**	**3.86**	**3.80**	**3.74**	**3.70**	**3.66**	**3.62**	**3.60**	
2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30	12
4.05	**3.98**	**3.86**	**3.78**	**3.70**	**3.61**	**3.56**	**3.49**	**3.46**	**3.41**	**3.38**	**3.36**	
2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21	13
3.85	**3.78**	**3.67**	**3.59**	**3.51**	**3.42**	**3.37**	**3.30**	**3.27**	**3.21**	**3.18**	**3.16**	
2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13	14
3.70	**3.62**	**3.51**	**3.43**	**3.34**	**3.26**	**3.21**	**3.14**	**3.11**	**3.06**	**3.02**	**3.00**	
2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07	15
3.56	**3.48**	**3.36**	**3.29**	**3.20**	**3.12**	**3.07**	**3.00**	**2.97**	**2.92**	**2.89**	**2.87**	
2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01	16
3.45	**3.37**	**3.25**	**3.18**	**3.10**	**3.01**	**2.96**	**2.89**	**2.86**	**2.80**	**2.77**	**2.75**	
2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96	17
3.35	**3.27**	**3.16**	**3.08**	**3.00**	**2.92**	**2.86**	**2.79**	**2.76**	**2.70**	**2.67**	**2.65**	
2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92	18
3.27	**3.19**	**3.07**	**3.00**	**2.91**	**2.83**	**2.78**	**2.71**	**2.68**	**2.62**	**2.59**	**2.57**	
2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88	19
3.19	**3.12**	**3.00**	**2.92**	**2.84**	**2.76**	**2.70**	**2.63**	**2.60**	**2.54**	**2.51**	**2.49**	
2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84	20
3.13	**3.05**	**2.94**	**2.86**	**2.77**	**2.69**	**2.63**	**2.56**	**2.53**	**2.47**	**2.44**	**2.42**	
2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81	21
3.07	**2.99**	**2.88**	**2.80**	**2.72**	**2.63**	**2.58**	**2.51**	**2.47**	**2.42**	**2.38**	**2.36**	
2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78	22
3.02	**2.94**	**2.83**	**2.75**	**2.67**	**2.58**	**2.53**	**2.46**	**2.42**	**2.37**	**2.33**	**2.31**	
2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76	23
2.97	**2.89**	**2.78**	**2.70**	**2.62**	**2.53**	**2.48**	**2.41**	**2.37**	**2.32**	**2.28**	**2.26**	
2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73	24
2.93	**2.85**	**2.74**	**2.66**	**2.58**	**2.49**	**2.44**	**2.36**	**2.33**	**2.27**	**2.23**	**2.21**	
2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71	25
2.89	**2.81**	**2.70**	**2.62**	**2.54**	**2.45**	**2.40**	**2.32**	**2.29**	**2.23**	**2.19**	**2.17**	
2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69	26
2.86	**2.77**	**2.66**	**2.58**	**2.50**	**2.41**	**2.36**	**2.28**	**2.25**	**2.19**	**2.15**	**2.13**	

ditional entries are by interpolation, mostly graphical.

"*Statistical Methods, 4th Edition*".

APPENDIX

5 % and 1 % Points for the Distribution of F.

n_2	n_1 degrees of freedom (for greater mean square)											
	1	2	3	4	5	6	7	8	9	10	11	12
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07
	7.50	**5.34**	**4.46**	**3.97**	**3.66**	**3.42**	**3.25**	**3.12**	**3.01**	**2.94**	**2.86**	**2.80**
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05
	7.44	**5.29**	**4.42**	**3.93**	**3.61**	**3.38**	**3.21**	**3.08**	**2.97**	**2.89**	**2.82**	**2.76**
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03
	7.39	**5.25**	**4.38**	**3.89**	**3.58**	**3.35**	**3.18**	**3.04**	**2.94**	**2.86**	**2.78**	**2.72**
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**	**2,75**	**2.69**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**	**2.73**	**2.66**
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99
	7.27	**5.15**	**4.29**	**3.80**	**3.49**	**3.26**	**3.10**	**2.96**	**2.86**	**2.77**	**2.70**	**2.64**
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98
	7.24	**5.12**	**4.26**	**3.78**	**3.46**	**3.24**	**3.07**	**2.94**	**2.84**	**2.75**	**2.68**	**2.62**
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**	**2.66**	**2.60**
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96
	7.19	**5.08**	**4.22**	**3.74**	**3.42**	**3.20**	**3.04**	**2.90**	**2.80**	**2.71**	**2.64**	**2.58**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.18**	**3.02**	**2.88**	**2.78**	**2.70**	**2.62**	**2.56**
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93
	7.12	**5.01**	**4.16**	**3.68**	**3.37**	**3.15**	**2.98**	**2.85**	**2.75**	**2.66**	**2.59**	**2.53**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**	**2.56**	**2.50**
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90
	7.04	**4.95**	**4.10**	**3.62**	**3.31**	**3.09**	**2.93**	**2.79**	**2.70**	**2.61**	**2.54**	**2.47**
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.79**	**2.65**	**2.56**	**2.47**	**2.40**	**2.33**
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82
	6.81	**4.75**	**3.91**	**3.44**	**3.14**	**2.92**	**2.76**	**2.62**	**2.53**	**2.44**	**2.37**	**2.30**
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.90**	**2.73**	**2.60**	**2.50**	**2.41**	**2.34**	**2.28**
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78
	6.70	**4.66**	**3.83**	**3.36**	**3.06**	**2.85**	**2.69**	**2.55**	**2.46**	**2.37**	**2.29**	**2.23**
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76
	6.66	**4.62**	**3.80**	**3.34**	**3.04**	**2.82**	**2.66**	**2.53**	**2.43**	**2.34**	**2.26**	**2.20**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75
	6.64	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**

Table V *(Continued)*

(5 % in Roman Type, 1 % in Bold Face Type).

n_1 degrees of freedom (for greater) mean square)												n_2
14	16	20	24	30	40	50	75	100	200	500	∞	
2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67	27
2.83	**2.74**	**2.63**	**2.55**	**2.47**	**2.38**	**2.33**	**2.25**	**2.21**	**2.16**	**2.12**	**2.10**	
2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65	28
2.80	**2.71**	**2.60**	**2.52**	**2.44**	**2.35**	**2.30**	**2.22**	**2.18**	**2.13**	**2.09**	**2.06**	
2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64	29
2.77	**2.68**	**2.57**	**2.49**	**2.41**	**2.32**	**2.27**	**2.19**	**2.15**	**2.10**	**2.06**	**2.03**	
2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62	30
2.74	**2.66**	**2.55**	**2.47**	**2.38**	**2.29**	**2.24**	**2.16**	**2.13**	**2.07**	**2.03**	**2.01**	
2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59	32
2.70	**2.62**	**2.51**	**2.42**	**2.34**	**2.25**	**2.20**	**2.12**	**2.08**	**2.02**	**1.98**	**1.96**	
2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57	34
2.66	**2.58**	**2.47**	**2.38**	**2.30**	**2.21**	**2.15**	**2.08**	**2.04**	**1.98**	**1.94**	**1.91**	
1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55	36
2.62	**2.54**	**2.43**	**2.35**	**2.26**	**2.17**	**2.12**	**2.04**	**2.00**	**1.94**	**1.90**	**1.87**	
1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53	38
2.59	**2.51**	**2.40**	**2.32**	**2.22**	**2.14**	**2.08**	**2.00**	**1.97**	**1.90**	**1.86**	**1.84**	
1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51	40
2.56	**2.49**	**2.37**	**2.29**	**2.20**	**2.11**	**2.05**	**1.97**	**1.94**	**1.88**	**1.84**	**1.81**	
1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49	42
2.54	**2.46**	**2.35**	**2.26**	**2.17**	**2.08**	**2.02**	**1.94**	**1.91**	**1.85**	**1.80**	**1.78**	
1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48	44
2.52	**2.44**	**2.32**	**2.24**	**2.15**	**2.06**	**2.00**	**1.92**	**1.88**	**1.82**	**1.78**	**1.75**	
1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46	46
2.50	**2.42**	**2.30**	**2.22**	**2.13**	**2.04**	**1.98**	**1.90**	**1.86**	**1.80**	**1.76**	**1.72**	
1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45	48
2.48	**2.40**	**2.28**	**2.20**	**2.11**	**2.02**	**1.96**	**1.88**	**1.84**	**1.78**	**1.73**	**1.70**	
1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44	50
2.46	**2.39**	**2.26**	**2.18**	**2.10**	**2.00**	**1.94**	**1.86**	**1.82**	**1.76**	**1.71**	**1.68**	
1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41	55
2.43	**2.35**	**2.23**	**2.15**	**2.06**	**1.96**	**1.90**	**1.82**	**1.78**	**1.71**	**1.66**	**1.64**	
1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39	60
2.40	**2.32**	**2.20**	**2.12**	**2.03**	**1.92**	**1.87**	**1.79**	**1.74**	**1.68**	**1.63**	**1.60**	
1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37	65
2.37	**2.30**	**2.18**	**2.09**	**2.00**	**1.90**	**1.84**	**1.76**	**1.71**	**1.64**	**1.60**	**1.56**	
1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35	70
2.25	**2.28**	**2.15**	**2.07**	**1.98**	**1.88**	**1.82**	**1.74**	**1.69**	**1.62**	**1.56**	**1.53**	
1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32	80
2.32	**2.24**	**2.11**	**2.03**	**1.94**	**1.84**	**1.78**	**1.70**	**1.65**	**1.57**	**1.52**	**1.49**	
1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28	100
2.26	**2.19**	**2.06**	**1.98**	**1.89**	**1.79**	**1.73**	**1.64**	**1.59**	**1.51**	**1.46**	**1.43**	
1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25	125
2.23	**2.15**	**2.03**	**1.94**	**1.85**	**1.75**	**1.68**	**1.59**	**1.54**	**1.46**	**1.40**	**1.37**	
1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22	150
2.20	**2.12**	**2.00**	**1.91**	**1.83**	**1.72**	**1.66**	**1.56**	**1.51**	**1.43**	**1.37**	**1.33**	
1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19	200
2.17	**2.09**	**1.97**	**1.88**	**1.79**	**1.69**	**1.62**	**1.53**	**1.48**	**1.39**	**1.33**	**1.28**	
1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13	400
2.12	**2.04**	**1.92**	**1.84**	**1.74**	**1.64**	**1.57**	**1.47**	**1.42**	**1.32**	**1.24**	**1.19**	
1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08	1000
2.09	**2.01**	**1.89**	**1.81**	**1.71**	**1.61**	**1.54**	**1.44**	**1.38**	**1.28**	**1.19**	**1.11**	
1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00	∞
2.07	**1.99**	**1.87**	**1.79**	**1.69**	**1.59**	**1.52**	**1.41**	**1.36**	**1.25**	**1.15**	**1.00**	

APPENDIX Table VI

Poisson Tables

Construction of the Table Giving the Probability of x or More Events Happening in a Given Interval if 'm', the Average Number of Events per Interval is Known – The probability that 'x' Events will Happen in a given time or space segment is equal to

$$P_n = \frac{e^{-m}(m^x)}{x!}.$$

where x refers to any value of 'n'.

The value of this expression for various values of 'm' and 'x' is readily available in standard Poisson tables.

Thus P_n may be found for any given values of 'x' and 'm'. For example, if $m = 4$ and $x = 0$.

$$P_0 = \frac{e^{-m}(m^x)}{x!} = \frac{e^{-4}(4^0)}{0!} = 0.018$$

If $m = 4$ and $x = 1$

$$P_1 = \frac{e^{-m}(m^x)}{x!} = \frac{e^{-4}(4^1)}{1!} = \frac{0.0183\ (4)}{1} = 0.073$$

If $m = 4$ and $x = 2$

$$P_2 = \frac{e^{-4}(4^2)}{2!} = \frac{.0183\ (16)}{2} = 0.147$$

If $m = 4$ and $x = 3$

$$P_3 = \frac{e^{-4}(4^3)}{3!} = \frac{0.0183\ (64)}{6} = 0.195$$

This procedure can of course, be continued.

The probability of getting three or less is the sum of the probability of getting 0, 1, 2 or 3 and therefore is equal $0.018 + 0.073 + 0.147 + 0.195 = 0.433 = 43.3$ in 100 or 43.3 per cent. The probability of getting four or more is 56.7 out of 100 or 56.7 per cent. This follows from the fact that the total probability of getting all possible numbers is one or 100 per cent. This is the procedure followed in the calculation of the tables. Therefore, the values given in the tables are

$$1 - e^{-m}\left(\frac{m^0}{0!} + \frac{m^1}{1!} + \frac{m^2}{2!} + \ldots + \frac{m^{(x-1)}}{(x-1)!}\right)$$

IF "m", THE AVERAGE NUMBER OF EVENTS PER INTERVAL, IS KNOWN, THEN THE PROBABILITY OF "x" OR MORE HAPPENING IN THIS INTERVAL MAY BE READ FROM THIS TABLE

m \ x	1	2	3	4	5	6	7	8	9	10	11
.1	.095	.005									
.2	.181	.018	.001								
.3	.259	.037	.004								
.4	.330	.062	.008	.001							
.5	.393	.090	.014	.002							
.6	.451	.122	.023	.003							
.7	.503	.156	.034	.006	.001						
.8	.551	.191	.047	.009	.001						
.9	.593	.228	.063	.013	.002						
1.0	.632	.264	.080	.018	.004	.001					
1.1	.667	.301	.100	.026	.005	.001					
1.2	.699	.337	.121	.034	.008	.002					
1.3	.727	.373	.143	.043	.011	.002					
1.4	.753	.408	.167	.054	.014	.003	.001				
1.5	.777	.442	.191	.066	.019	.004	.001				
1.6	.798	.475	.217	.079	.024	.006	.001				
1.7	.817	.507	.243	.093	.030	.008	.002				
1.8	.835	.537	.269	.109	.036	.010	.003	.001			
1.9	.850	.566	.296	.125	.044	.013	.003	.001			
2.0	.865	.594	.323	.143	.053	.017	.005	.001			
2.1	.878	.620	.350	.161	.062	.020	.006	.001			
2.2	.889	.645	.377	.181	.072	.025	.007	.002			
2.3	.900	.669	.404	.201	.084	.030	.009	.003	.001		
2.4	.909	.692	.430	.221	.096	.036	.012	.003	.001		
2.5	.918	.713	.456	.242	.109	.042	.014	.004	.001		
2.6	.926	.733	.482	.264	.123	.049	.017	.005	.001		
2.7	.933	.751	.506	.286	.137	.057	.021	.007	.002	.001	
2.8	.939	.769	.531	.308	.152	.065	.024	.008	.002	.001	
2.9	.945	.785	.554	.330	.168	.074	.029	.010	.003	.001	
3.0	.950	.801	.577	.353	.185	.084	.034	.012	.004	.001	
3.1	.955	.815	.599	.375	.202	.094	.039	.014	.005	.001	
3.2	.959	.829	.620	.397	.219	.105	.045	.017	.006	.002	
3.3	.963	.841	.641	.420	.237	.117	.051	.020	.007	.002	.001
3.4	.967	.853	.660	.442	.256	.129	.058	.023	.008	.003	.001
3.5	.970	.864	.679	.463	.275	.142	.065	.027	.010	.003	.001

IF "m", THE AVERAGE NUMBER OF EVENTS PER INTERVAL, IS KNOWN, THEN THE PROBABILITY OF "x" OR MORE HAPPENING IN THIS INTERVAL MAY BE READ FROM THIS TABLE

m \ x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
3.6	.973	.874	.697	.485	.294	.156	.073	.031	.012	.004	.001						
3.7	.975	.884	.715	.506	.313	.170	.082	.035	.014	.005	.002						
3.8	.978	.893	.731	.527	.332	.184	.091	.040	.016	.006	.002	.001					
3.9	.980	.901	.747	.547	.352	.199	.101	.045	.019	.007	.002	.001					
4.0	.982	.908	.762	.567	.371	.215	.111	.051	.021	.008	.003	.001					
4.1	.983	.915	.776	.586	.391	.231	.121	.057	.024	.010	.003	.001					
4.2	.985	.922	.790	.605	.410	.247	.133	.064	.028	.011	.004	.001					
4.3	.986	.928	.803	.623	.430	.263	.144	.071	.032	.013	.005	.002	.001				
4.4	.988	.934	.815	.641	.449	.280	.156	.079	.036	.015	.006	.002	.001				
4.5	.989	.939	.826	.658	.468	.297	.169	.087	.040	.017	.007	.002	.001				
4.6	.990	.944	.837	.674	.487	.314	.182	.095	.045	.020	.008	.003	.001				
4.7	.991	.948	.848	.690	.505	.332	.195	.104	.050	.022	.009	.003	.001				
4.8	.992	.952	.857	.706	.524	.349	.209	.113	.056	.025	.010	.004	.001				
4.9	.993	.956	.867	.721	.542	.366	.233	.123	.062	.028	.012	.005	.002	.001			
5.0	.993	.960	.875	.735	.560	.384	.238	.133	.068	.032	.014	.005	.002	.001			
5.1	.994	.963	.884	.749	.577	.402	.253	.144	.075	.036	.016	.006	.002	.001			
5.2	.994	.966	.891	.762	.594	.419	.268	.155	.082	.040	.018	.007	.003	.001			
5.3	.995	.969	.898	.775	.610	.437	.283	.167	.089	.044	.020	.008	.003	.001			
5.4	.995	.971	.905	.787	.627	.454	.298	.178	.097	.049	.023	.010	.004	.001			
5.5	.996	.973	.912	.798	.642	.471	.314	.191	.106	.054	.025	.011	.004	.002	.001		
5.6	.996	.976	.918	.809	.658	.488	.330	.203	.114	.059	.028	.012	.005	.002	.001		
5.7	.997	.978	.923	.820	.673	.505	.346	.216	.123	.065	.031	.014	.006	.002	.001		
5.8	.997	.979	.928	.830	.687	.522	.362	.229	.133	.071	.035	.016	.007	.003	.001		
5.9	.997	.981	.933	.840	.701	.538	.378	.242	.143	.077	.039	.018	.008	.003	.001		
6.0	.998	.983	.938	.849	.715	.554	.394	.256	.153	.084	.043	.020	.009	.004	.001	.001	
6.1	.998	.984	.942	.857	.728	.570	.410	.270	.163	.091	.047	.022	.010	.004	.002	.001	
6.2	.998	.985	.946	.866	.741	.586	.426	.284	.174	.098	.051	.025	.011	.005	.002	.001	
6.3	.998	.987	.950	.874	.753	.601	.442	.298	.185	.106	.056	.028	.013	.005	.002	.001	
6.4	.998	.988	.954	.881	.765	.616	.458	.313	.197	.114	.061	.031	.014	.006	.003	.001	
6.5	.998	.989	.957	.888	.776	.631	.473	.327	.208	.123	.067	.034	.016	.007	.003	.001	
6.6	.999	.990	.960	.895	.787	.645	.489	.342	.220	.131	.073	.037	.018	.008	.003	.001	.001
6.7	.999	.991	.963	.901	.798	.659	.505	.357	.233	.140	.079	.041	.020	.009	.004	.002	.001
6.8	.999	.991	.966	.907	.808	.673	.520	.372	.245	.150	.085	.045	.022	.010	.004	.002	.001
6.9	.999	.992	.968	.913	.818	.686	.535	.386	.258	.160	.092	.049	.024	.011	.005	.002	.001
7.0	.999	.993	.970	.918	.827	.699	.550	.401	.271	.170	.099	.053	.027	.013	.006	.002	.001

IF "m", THE AVERAGE NUMBER OF EVENTS PER INTERVAL, IS KNOWN, THEN THE PROBABILITY OF "x" OR MORE HAPPENING IN THIS INTERVAL MAY BE READ FROM THIS TABLE

m \ x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
7.1	.999	.993	.973	.923	.836	.712	.565	.416	.284	.180	.106	.058	.030	.014	.006	.003	.001						
7.2	.999	.994	.975	.928	.844	.724	.580	.431	.297	.190	.113	.063	.033	.016	.007	.003	.001						
7.3	.999	.994	.976	.933	.853	.736	.594	.446	.311	.201	.121	.068	.036	.018	.008	.004	.001	.001					
7.4	.999	.995	.978	.937	.860	.747	.608	.461	.324	.212	.129	.074	.039	.020	.009	.004	.002	.001					
7.5	.999	.995	.980	.941	.868	.759	.622	.475	.338	.224	.138	.079	.043	.022	.010	.005	.002	.001					
7.6	.999	.996	.981	.945	.875	.769	.635	.490	.352	.235	.146	.085	.046	.024	.011	.005	.002	.001					
7.7	1.00	.996	.983	.948	.882	.780	.649	.504	.366	.247	.155	.091	.050	.026	.013	.006	.003	.001					
7.8	1.00	.996	.984	.952	.888	.790	.662	.519	.380	.259	.165	.098	.055	.029	.014	.007	.003	.001					
7.9	1.00	.997	.985	.955	.894	.799	.674	.533	.393	.271	.174	.105	.059	.031	.016	.007	.003	.001	.001				
8.0	1.00	.997	.986	.958	.900	.809	.687	.547	.407	.283	.184	.112	.064	.034	.017	.008	.004	.002	.001				
8.1	1.00	.997	.987	.960	.906	.818	.699	.561	.421	.296	.194	.119	.069	.037	.019	.009	.004	.002	.001				
8.2	1.00	.997	.988	.963	.911	.826	.710	.575	.435	.308	.204	.127	.074	.040	.021	.010	.005	.002	.001				
8.3	1.00	.998	.989	.965	.916	.835	.722	.588	.449	.321	.215	.135	.079	.044	.023	.011	.005	.002	.001				
8.4	1.00	.998	.990	.968	.921	.843	.733	.601	.463	.334	.226	.143	.085	.048	.025	.013	.006	.003	.001				
8.5	1.00	.998	.991	.970	.926	.850	.744	.614	.477	.347	.237	.151	.091	.051	.027	.014	.007	.003	.001	.001			
8.6	1.00	.998	.991	.972	.930	.858	.754	.627	.491	.360	.248	.160	.097	.055	.030	.015	.007	.003	.001	.001			
8.7	1.00	.998	.992	.974	.934	.865	.765	.640	.504	.373	.259	.169	.103	.060	.033	.017	.008	.004	.002	.001			
8.8	1.00	.999	.993	.976	.938	.872	.774	.652	.518	.386	.271	.178	.110	.064	.035	.018	.009	.004	.002	.001			
8.9	1.00	.999	.993	.977	.942	.878	.784	.664	.531	.399	.282	.187	.117	.069	.038	.020	.010	.005	.002	.001			
9.0	1.00	.999	.994	.979	.945	.884	.793	.676	.544	.413	.294	.197	.124	.074	.041	.022	.011	.005	.002	.001			
9.1	1.00	.999	.994	.980	.948	.890	.802	.688	.557	.426	.306	.207	.132	.079	.045	.024	.012	.006	.003	.001	.001		
9.2	1.00	.999	.995	.982	.951	.896	.811	.699	.570	.439	.318	.217	.139	.084	.085	.026	.013	.007	.003	.001	.001		
9.3	1.00	.999	.995	.983	.954	.901	.819	.710	.583	.452	.330	.227	.147	.090	.052	.028	.015	.007	.003	.002	.001		
9.4	1.00	.999	.995	.984	.957	.907	.827	.721	.596	.465	.342	.237	.155	.096	.056	.031	.016	.008	.004	.002	.001		
9.5	1.00	.999	.996	.985	.960	.911	.835	.731	.608	.478	.355	.248	.164	.102	.060	.033	.018	.009	.004	.002	.001		
9.6	1.00	.999	.996	.986	.962	.916	.843	.742	.620	.491	.367	.259	.172	.108	.064	.036	.019	.010	.005	.002	.001		
9.7	1.00	.999	.996	.987	.965	.921	.850	.752	.632	.504	.379	.270	.181	.115	.069	.039	.021	.011	.005	.002	.001		
9.8	1.00	.999	.997	.988	.967	.925	.857	.761	.644	.517	.392	.281	.190	.121	.073	.042	.023	.012	.006	.003	.001	.001	
9.9	1.00	.999	.997	.989	.969	.929	.863	.771	.656	.529	.404	.292	.199	.128	.078	.045	.025	.013	.007	.003	.001	.001	
10.0	1.00	1.00	.997	.990	.971	.933	.870	.780	.667	.542	.417	.303	.208	.136	.083	.049	.027	.014	.007	.003	.002	.001	
10.1	1.00	1.00	.997	.990	.973	.937	.876	.789	.678	.555	.429	.315	.218	.143	.089	.052	.029	.016	.008	.004	.002	.001	
10.2	1.00	1.00	.998	.991	.974	.940	.882	.797	.689	.567	.442	.326	.228	.151	.094	.056	.032	.017	.009	.004	.002	.001	
10.3	1.00	1.00	.998	.992	.976	.943	.888	.806	.700	.579	.454	.338	.238	.158	.100	.060	.034	.019	.010	.005	.002	.001	
10.4	1.00	1.00	.998	.992	.977	.947	.893	.814	.710	.591	.467	.350	.248	.166	.106	.064	.037	.020	.011	.005	.003	.001	.001
10.5	1.00	1.00	.998	.993	.979	.950	.898	.821	.721	.603	.479	.361	.258	.175	.112	.068	.040	.033	.012	.006	.003	.001	.001

IF "*m*", THE AVERAGE NUMBER OF EVENTS PER INTERVAL, IS KNOWN, THEN THE PROBABILITY OF "*x*" OR MORE HAPPENING IN THIS INTERVAL MAY BE READ FROM THIS TABLE

m \ x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
10.6	1.00	1.00	.998	.993	.980	.952	.903	.829	.731	.615	.492	.373	.268	.183	.118	.073	.043	.024	.013	.006	.003	.001	.001					
10.7	1.00	1.00	.998	.994	.982	.955	.908	.836	.740	.626	.504	.385	.279	.192	.125	.077	.046	.026	.014	.007	.003	.002	.001					
10.8	1.00	1.00	.999	.994	.983	.958	.913	.843	.750	.637	.516	.397	.290	.201	.132	.082	.049	.028	.015	.008	.004	.002	.001					
10.9	1.00	1.00	.999	.995	.984	.960	.917	.850	.759	.649	.528	.409	.300	.210	.137	.087	.052	.030	.016	.008	.004	.002	.001					
11.0	1.00	1.00	.999	.995	.985	.962	.921	.857	.768	.659	.540	.421	.311	.219	.146	.093	.056	.032	.018	.009	.005	.002	.001					
11.1	1.00	1.00	.999	.995	.986	.965	.925	.863	.777	.670	.552	.433	.322	.228	.153	.098	.060	.035	.019	.010	.005	.003	.001	.001				
11.2	1.00	1.00	.999	.996	.987	.967	.929	.869	.785	.681	.564	.445	.333	.238	.161	.104	.064	.037	.021	.011	.006	.003	.001	.001				
11.3	1.00	1.00	.999	.996	.988	.969	.933	.875	.794	.691	.575	.456	.345	.247	.169	.109	.068	.040	.022	.012	.006	.003	.001	.001				
11.4	1.00	1.00	.999	.996	.988	.971	.936	.881	.802	.701	.587	.468	.356	.257	.177	.115	.072	.043	.024	.013	.007	.003	.002	.001				
11.5	1.00	1.00	.999	.997	.989	.972	.940	.886	.809	.711	.598	.480	.367	.267	.185	.122	.076	.046	.026	.014	.008	.004	.002	.001				
11.6	1.00	1.00	.999	.997	.990	.974	.943	.892	.817	.721	.609	.492	.378	.277	.193	.128	.081	.049	.028	.016	.008	.004	.002	.001				
11.7	1.00	1.00	.999	.997	.991	.975	.946	.897	.824	.730	.621	.504	.390	.287	.202	.135	.086	.052	.030	.017	.009	.005	.002	.001				
11.8	1.00	1.00	.999	.997	.991	.977	.949	.901	.831	.740	.631	.515	.401	.297	.210	.141	.091	.056	.033	.018	.010	.005	.002	.001	.001			
11.9	1.00	1.00	.999	.998	.992	.978	.952	.906	.838	.749	.642	.527	.413	.308	.219	.148	.096	.059	.035	.020	.011	.006	.003	.001	.001			
12.0	1.00	1.00	.999	.998	.992	.980	.954	.910	.845	.758	.653	.538	.424	.318	.228	.156	.101	.063	.037	.021	.012	.006	.003	.001	.001			
12.1	1.00	1.00	1.00	.998	.993	.981	.957	.915	.851	.766	.663	.550	.435	.329	.237	.163	.107	.067	.040	.023	.013	.007	.003	.002	.001			
12.2	1.00	1.00	1.00	.998	.993	.982	.959	.919	.858	.775	.673	.561	.447	.340	.246	.170	.113	.071	.043	.025	.014	.007	.004	.002	.001			
12.3	1.00	1.00	1.00	.998	.994	.983	.961	.923	.864	.783	.683	.572	.458	.350	.256	.178	.118	.075	.046	.027	.015	.008	.004	.002	.001			
12.4	1.00	1.00	1.00	.998	.994	.984	.963	.927	.869	.791	.693	.583	.470	.361	.265	.186	.124	.080	.049	.029	.016	.009	.004	.002	.001			
12.5	1.00	1.00	1.00	.998	.995	.985	.965	.930	.875	.799	.703	.594	.481	.372	.275	.194	.131	.084	.052	.031	.017	.009	.005	.002	.001	.001		
12.6	1.00	1.00	1.00	.999	.995	.986	.967	.934	.880	.806	.712	.605	.492	.383	.285	.202	.137	.089	.055	.033	.019	.010	.005	.003	.001	.001		
12.7	1.00	1.00	1.00	.999	.995	.987	.969	.937	.886	.813	.722	.616	.504	.394	.295	.210	.144	.094	.059	.035	.020	.011	.006	.003	.001	.001		
12.8	1.00	1.00	1.00	.999	.996	.988	.971	.940	.891	.821	.731	.626	.515	.405	.305	.219	.150	.099	.062	.037	.022	.012	.006	.003	.002	.001		
12.9	1.00	1.00	1.00	.999	.996	.989	.973	.943	.896	.827	.740	.637	.526	.416	.315	.228	.157	.104	.066	.040	.023	.013	.007	.004	.002	.001		
13.0	1.00	1.00	1.00	.999	.996	.989	.974	.946	.900	.834	.748	.647	.537	.427	.325	.236	.165	.110	.070	.043	.025	.014	.008	.004	.002	.001		
13.1	1.00	1.00	1.00	.999	.997	.990	.976	.949	.905	.841	.757	.657	.548	.438	.335	.245	.172	.115	.074	.045	.027	.015	.008	.004	.002	.001	.001	
13.2	1.00	1.00	1.00	.999	.997	.991	.977	.951	.909	.847	.765	.667	.559	.449	.345	.254	.179	.121	.078	.048	.029	.016	.009	.005	.002	.001	.001	
13.3	1.00	1.00	1.00	.999	.997	.991	.978	.954	.913	.853	.773	.677	.569	.460	.356	.264	.187	.127	.082	.051	.031	.018	.010	.005	.003	.001	.001	
13.4	1.00	1.00	1.00	.999	.997	.992	.980	.956	.917	.859	.781	.686	.580	.471	.366	.273	.195	.133	.087	.055	.033	.019	.011	.006	.003	.001	.001	
13.5	1.00	1.00	1.00	.999	.997	.992	.981	.959	.921	.865	.789	.696	.591	.482	.377	.282	.202	.139	.092	.058	.035	.020	.011	.006	.003	.002	.001	
13.6	1.00	1.00	1.00	.999	.998	.993	.982	.961	.925	.870	.796	.705	.601	.493	.387	.292	.211	.146	.096	.061	.037	.022	.012	.007	.004	.002	.001	
13.7	1.00	1.00	1.00	.999	.998	.993	.983	.963	.928	.876	.804	.714	.611	.503	.398	.301	.219	.152	.101	.065	.040	.024	.013	.007	.004	.002	.001	
13.8	1.00	1.00	1.00	.999	.998	.994	.984	.965	.932	.881	.811	.723	.622	.514	.408	.311	.227	.159	.107	.069	.042	.025	.014	.008	.004	.002	.001	.001
13.9	1.00	1.00	1.00	.999	.998	.994	.985	.967	.935	.886	.818	.731	.632	.525	.419	.321	.235	.166	.112	.072	.045	.027	.016	.009	.005	.002	.001	.001
14.0	1.00	1.00	1.00	1.00	.998	.994	.986	.968	.938	.891	.824	.740	.642	.536	.430	.331	.244	.173	.117	.076	.048	.029	.017	.009	.005	.003	.001	.001

IF "m", THE AVERAGE NUMBER OF EVENTS PER INTERVAL, IS KNOWN, THEN THE PROBABILITY OF "x" OR MORE HAPPENING IN THIS INTERVAL MAY BE READ FROM THIS TABLE

m \ x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
14.1	1.00	1.00	1.00	1.00	.998	.995	.987	.970	.941	.895	.831	.748	.651	.546	.440	.341	.253	.180	.123	.081	.051	.031	.018	.010	.005	.003	.001	.001	
14.2	1.00	1.00	1.00	1.00	.998	.995	.987	.972	.944	.900	.837	.756	.661	.557	.451	.351	.262	.187	.129	.085	.054	.033	.019	.011	.006	.003	.002	.001	
14.3	1.00	1.00	1.00	1.00	.999	.995	.988	.973	.947	.904	.843	.764	.670	.567	.461	.361	.271	.195	.135	.089	.057	.035	.021	.012	.006	.003	.002	.001	
14.4	1.00	1.00	1.00	1.00	.999	.996	.989	.975	.949	.908	.849	.772	.680	.577	.472	.371	.280	.203	.141	.094	.060	.037	.022	.013	.007	.004	.002	.001	
14.5	1.00	1.00	1.00	1.00	.999	.996	.990	.976	.952	.912	.855	.780	.689	.587	.482	.381	.289	.210	.147	.099	.064	.040	.024	.014	.008	.004	.002	.001	.001
14.6	1.00	1.00	1.00	1.00	.999	.996	.990	.977	.954	.916	.861	.787	.698	.598	.493	.391	.298	.218	.153	.104	.067	.042	.025	.015	.008	.004	.002	.001	.001
14.7	1.00	1.00	1.00	1.00	.999	.997	.991	.979	.956	.920	.866	.795	.707	.608	.503	.401	.307	.226	.160	.109	.071	.045	.027	.016	.009	.005	.003	.001	.001
14.8	1.00	1.00	1.00	1.00	.999	.997	.991	.980	.958	.923	.871	.802	.715	.617	.514	.411	.317	.234	.167	.114	.075	.047	.029	.017	.010	.005	.003	.001	.001
14.9	1.00	1.00	1.00	1.00	.999	.997	.992	.981	.961	.927	.877	.809	.724	.627	.524	.422	.326	.243	.174	.119	.079	.050	.031	.018	.010	.006	.003	.002	.001
15.0	1.00	1.00	1.00	1.00	.999	.997	.992	.982	.963	.930	.882	.815	.732	.638	.534	.432	.336	.251	.181	.125	.083	.053	.033	.019	.011	.006	.003	.002	.001

INDEX

α	Alpha
β	Beta
γ	Gamma
δ	Delta
ε	Epsilon
ζ	Zeta
η	Eta
θ	Theta
ι	Iota
$\varkappa$	Kappa
λ	Lambda
μ	Mu
ν	Nu
ξ	Xi
o	Omicron
π	Pi
ϱ	Rho
σ or ς	Sigma
τ	Tau
υ	Upsilon
φ	Phi
χ	Chi
ψ	Psi
ω	Omega