ELEMENTARY SAMPLING FOR TRAFFIC ENGINEERS

DAVID F. VOTAW, JR.
The Mitre Corporation, Bedford, Massachusetts

HERBERT S. LEVINSON
Wilbur Smith and Associates, New Haven, Connecticut
Department of City Planning, Yale University, New Haven, Connecticut

THE ENO FOUNDATION FOR HIGHWAY TRAFFIC CONTROL
SAUGATUCK, CONNECTICUT, 1962

Eno Foundation publications are provided
through an endowment by William P. Eno.

Foreword

In its continuing effort to broaden the scope of approach toward solving the growing, intricate problems arising from increasing use of the automobile, the Eno Foundation welcomes the opportunity to publish and distribute this book.

ENO FOUNDATION

Preface

In traffic engineering there is an extensive use of quantitative data in the planning, design and operation of transportation facilities. Frequently there is an urgent need for drawing conclusions and making decisions on the basis of this data. As a result, traffic engineers are often faced by difficult statistical problems—i.e., problems in the collection, analysis and interpretation of data. Moreover, these problems become increasingly complex as traffic engineering technology advances. Many traffic engineers would therefore be helped by having a really modern account of the statistical approach to traffic engineering problems. Our aim in writing this book is to meet their needs.

To avoid "riding off in all directions at once," we decided to focus attention on one area of statistics—namely, sampling—and to show, in depth, how it relates to traffic engineering. Sampling was selected since it is broadly applicable to traffic engineering and since its essential ideas can be grasped readily. *The theme of this book is that sampling is a powerful tool for the planning and design of traffic engineering studies and the analysis of data obtained from them.*

The concepts that underlie sampling are presented and illus-

trated in Chapter 1 which places special emphasis on the concepts of *population* and *sample*. The methods of classical statistical inference are introduced and applied in the subsequent chapters. *Point* and *interval estimation* of population characteristics are treated in Chapters 2 and 3, respectively; *significance testing* is treated in Chapter 4; Chapter 5 gives traffic engineering applications of the ideas presented in the first four chapters. The scope of the book includes sampling from any of a broad class of populations; however, the main emphasis throughout is on sampling from binomial, Poisson, and normal populations.

It was necessary to assume that most readers would not be familiar with statistics. As a result the book is quite elementary in a statistical sense, and many potential applications of sampling to traffic engineering had to be omitted. For example, techniques of correlation, regression, and the analysis of variance have not been treated, and relatively little discussion of sample surveys has been included. The mathematical level of the book is also elementary; in fact, a knowledge of calculus is not necessary for understanding the subject matter.

We are indebted to A. M. Mood and the McGraw-Hill Book Company for permission to use Appendix Table 1, to Professor E. S. Pearson, editor of *Biometrika*, and the Biometrika Office for permission to include Figure 1 and Appendix Table 2; to Sir Ronald A. Fisher and Frank Yates, and Messrs. Oliver and Boyd, Ltd., Edinburgh, for permission to include Appendix Table 3; to the Rand Corporation and The Free Press for permission to include Appendix Table 4; and to W. E. Ricker and the editor of the *Journal of the American Statistical Association* for permission to include Table VI.

We wish to express our appreciation to Mr. William R. McGrath of the City of New Haven for his review of the manuscript, and to Mr. Matthew J. Huber of the Yale University Bureau of Highway Traffic for helpful discussion of certain points regarding short counts. We also wish to express appreciation to the Eno Foundation for counsel and support.

Winchester, Massachusetts
New Haven, Connecticut
March, 1962

D. F. V., JR.
H. S. L.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

There is an important relation between statistics and traffic engineering. Statistics is the science that deals with general principles regarding the collection, analysis, and interpretation of data. Traffic engineering deals with specific applications of those principles in fundamental traffic research and in everyday studies—for example, studies of traffic volumes, origins and destinations, travel modes and patterns, speed and delay, and parking.

From a statistical point of view the traffic engineer's data can often be regarded as forming a *sample* from a larger *population.* In terms of this point of view, the traffic engineer's objective in collecting the data is to draw conclusions or make decisions about the population. Sampling methods and concepts are directly relevant to the planning and design of his studies and the analysis of data obtained from them. The use of these methods and concepts can make important contributions to the effectiveness and efficiency of his work.

The purpose of this book is to help traffic engineers in the use of sampling. Accordingly, emphasis is placed on techniques and applications rather than theory and derivation. Most of the techniques considered pertain to estimation and significance testing. Almost all of the illustrative examples are closely related to traffic engineering practice.

1.1. Preliminary Remarks

Information about a population may be obtained by sampling from the population. The information usually becomes more accurate as the "size" of the sample increases. In fact, the information is completely accurate if every element of the population is included in the sample. On the other hand, costs will often increase as the "size" of the sample increases. Accordingly, in designing a sampling procedure the traffic engineer will frequently wish to consider both *cost* and *accuracy* of information.

The relation between "size" of the sample and accuracy of information can be indicated with regard to coin tossing. If a

"true" coin is tossed a number of times (e.g., 10 times), the observed proportion of Heads will usually not be exactly equal to 1/2, which is the "expected" proportion of Heads (since the single-toss probability of Heads equals 1/2). The set of observed tosses can be regarded as a sample from a (hypothetical) population of tosses, and the observed proportion of Heads can be regarded as an estimate of the expected proportion. The number of tosses is the sample *size*. Accuracy of information in the sample is represented here by the accuracy of the estimate.

Table I: Effect of Sample Size on Accuracy of Estimating the Probability of Heads in Coin Tossing

Number of Tosses (Sample Size)	*Range Within Which Observed Proportion of Heads Will Fall 95 percent of the Time**	*Maximum Per Cent Deviation†*
10	0.20–0.80	± 60
20	0.30–0.70	± 40
30	0.34–0.66	± 32
40	0.35–0.65	± 30
50	0.36–0.64	± 28
100	0.40–0.60	± 20
250	0.44–0.56	± 12
1000	0.47–0.53	± 6

*More precisely, if the single-toss probability of Heads equals 1/2 then the probability is 0.95 (approximately) that the observed proportion of Heads will fall *inside* the range, and the probability is 0.05 (approximately) that the proportion will fall *outside* the range.

†"Maximum Per Cent Deviation" here means the percent deviation of the limits of the range from 1/2. For example, the first entry is ± 60 since (0.80 − 0.50) / (0.50) = + 60 percent and (0.20 − 0.50) / (0.50) = − 60 percent.

The relation between sample size and accuracy of the estimate is indicated in Table I. If the coin is tossed 10 times, then the observed proportion of Heads is likely to be within 60 percent of the expected proportion (namely 1/2). On the other hand, if the coin is tossed 1000 times, the observed proportion is likely to be within 6 percent of the expected proportion. It is clear that the maximum percent deviation decreases as the sample size increases.

It is noteworthy that as the sample size increases a hundredfold (from 10 to 1000), the decrease in maximum percent deviation is

only tenfold (from $\pm$ 60 to $\pm$ 6). This is a typical feature of estimation by means of sampling. An increase in sample size does not yield a proportional increase in accuracy of estimation.

1.2. Some Important Definitions and Concepts

1.2.a. Populations and Samples. A *population* (or universe) is a class or set of objects. The set may be finite or infinite. A *sample* from the population is a set of objects "drawn" from the population. To "draw" an object requires only that the object be observed; it is not required that the object be removed from the population. Further details on populations and samples are given in Section 1 of the Appendix.

Table II: Examples of Populations and Samples Considered in Traffic Engineering

Subject of Study	*Population*	*Sample*
Daily traffic at a given location	Set of all vehicle passages past the location in 24 hours	Set of all observed vehicle passages past the location in the 24-hour period
Mode of travel of people entering a store	Set of all people entering store	Set of people entering store who are interviewed
Spot speeds at a given location	Set of speeds of all vehicles passing the location	Set of speeds observed
Home interview origin-destination study (origins and destinations of trips in survey area)	Set of all dwelling units in survey area	Set of dwelling units where interviews are obtained
Trip origins of vehicles passing a given location	Set of trip origins of all vehicles passing the location	Set of recorded trip origins of vehicles passing the location

Illustrative examples of samples and populations considered in traffic engineering are given in Table II. For example, in a study of vehicle speed at a given location (on a given day) the population would consist of the speeds of all vehicles at the time they passed

the location; and the sample would consist of the *observed* speeds of vehicles at the time they passed the location. Similarly, 10,000 dwelling units would represent a 10 percent sample from a population of 100,000 dwelling units.

It will be evident from subsequent chapters of this book that populations can be specified in such a way that their elements are numerical. Samples from such populations also consist of numerical elements. Unless otherwise indicated, the elements of a sample will be represented by numerical values, say $x_1, x_2, \ldots, x_n$. The number, n, of elements is called the *sample size.*

Certain functions of a sample are particularly useful. Well-known examples of such functions are the *sample mean* and the *sample variance.* These and other special functions of samples are described in Section 2 of the Appendix.

Most of the sampling considered in this book is *random sampling.* Uusually the elements $x_1, x_2, \ldots, x_n$ of a random sample will be regarded as observed values of a *random variable* whose *probability distribution* involves important characteristics of the population being sampled.* It is appropriate to say that the objective of random sampling is to obtain information about the probability distribution of the random variable involved.

1.2.b. Random Variables and Probability Distributions. Examples of random variables are: (1) the number that comes up when a die is tossed; (2) the proportion of Heads obtained in a given number of tosses of a coin. (Example (2) is involved in the situation treated in Table I.) Each random variable considered in this book is characterized by a *frequency function,* which specifies the variable's probability distribution. The probability distribution can also be specified by a *cumulative distribution function.* Section 3 of the Appendix gives definitions of a random variable, a probability distribution, a cumulative distribution function, a frequency function, and other closely related terms.

It should be remarked here that probability theory is the source of such concepts as random variable and frequency function; how-

*For further remarks on random sampling see Section 1 of the Appendix. For a discussion of random sampling techniques and the selection of a value of a random variable see Section 5.2.

Table III: Some Important Frequency Functions

Name of Distribution	*Frequency Function*	*Parameters*	*Examples of Applications*
Binomial	$f(x) = C_x^n p^x (1-p)^{n-x}$ (C_x^n is a binomial coefficient*)	p = probability of a given category	Populations having *two* categories† (e.g., "Successes" and "Failures," local vehicles and non-local vehicles, etc.).
Poisson	$f(x) = \frac{e^{-m} m^x}{x!}$	m = expected value (mean)	Random arrival of cars in a parking garage entrance.
Normal	$f(x) = \frac{1}{\sigma\sqrt{(2\pi)}} e^{-(x-u)^2/2\sigma^2}$	u = expected value (mean) σ^2 = variance	Studies of spot speeds, reaction times, etc.
Standard Normal	$f(x) = \frac{1}{\sqrt{(2\pi)}} e^{-x^2/2}$	$u = 0$ $\sigma^2 = 1$	Calculations involving normal distributions.
Hypergeometric	$f(x) = \frac{C_x^{N_1} C_{n-x}^{N-N_1}}{C_n^N}$, N = Population size. N_1 = No. of "successes" in population n = Sample size. x = No. of "Successes" in sample. ($C_x^{N_1}$, $C_{n-x}^{N-N_1}$, C_n^N are binomial coefficients*)	N_1/N = Proportion of "Successes" in population	Finite binomial population.
Multinomial	$f(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1!\, n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$ $(n_1 + n_2 + \ldots + n_k = n)$. The parameters are $p_1, p_2, \ldots, p_k$, where p_i = probability of drawing category i $(p_1 + p_2 + \ldots + p_k = 1)$.		Populations having k categories. $(k \geq 2)$.

*The definition of a binomial coefficient is given below the first formula in Section 4 of the Appendix.
†In effect the populations are assumed to be infinite (see Section 2.2).

ever, a treatment of probability theory is beyond the scope of this book. An excellent account of the subject is given by Feller.*

1.2.c. Important Frequency Functions. Several frequency functions important in traffic engineering are presented in Table III and are described in further detail in Section 4 of the Appendix. They are the *binomial*, *Poisson*, *normal*, *standard normal*, *hypergeometric*, and *multinomial* frequency functions. In almost every case only a small number of parameters† is involved. For example, the binomial and Poisson frequency functions are each characterized by one parameter and the normal frequency function is characterized by two. The parameter in the binomial is represented by p, and the parameter in the Poisson is represented by m; the parameters in the normal frequency function are represented by u and σ^2.

Typical traffic engineering applications of these frequency functions are also shown in Table III. For example, the distribution of the number of cars arriving at a garage often follows a Poisson frequency function (see Gerlough); and the distribution of spot speeds usually follows a normal frequency function.

*The Bibliography of this book (pp. 122-123) gives a complete reference to the book by Feller and to all other literature referred to in the text.

†A population average (or mean) is an example of a parameter contained in many frequency functions. A parameter of a frequency function is, of course, a parameter of the probability distribution specified by the frequency function. Similarly, a parameter of a cumulative distribution function is a parameter of the probability distribution specified by the cumulative distribution function.

Chapter 2: Point Estimation

As indicated in Chapter 1, the objective of sampling is to obtain information about the probability distribution associated with a group of observations. Often it is desirable to use this information to *estimate* some characteristic of the distribution—for example, a parameter such as the *mean** of the distribution.

There are two basic types of statistical estimation—namely, point estimation and interval estimation. The subject matter dealt with in this chapter is mainly the point estimation of parameters of binomial, Poisson, and normal distributions, and of *percent points** of distributions. Interval estimation of such quantities is considered in Chapter 3.

It should be remarked that sometimes a characteristic of a distribution will be referred to as a characteristic of a population. Accordingly, a parameter of a distribution will be referred to as a population parameter. More specifically, the mean of a distribution will be referred to as a population mean. When a population is finite, the population mean is, of course, the average of the elements of the population.

2.1. Definitions and Notation

When a single number obtained from a sample is used as an estimate of a population parameter, it is called a *point estimate* of the parameter. An analytic discussion of this and related concepts is given below.

Let θ represent a parameter of the distribution of a random variable, say X. The value of θ is assumed to be unknown to the investigator, and his aim is to estimate the value of θ on basis of a sample $x_1, x_2, \ldots, x_n$. To accomplish his purpose the investigator makes use of a function, say $\theta\ (x_1, x_2, \ldots, x_n)$, of the sample. For example, if θ is the mean of the distribution, then in many cases a suitable function to use is $\hat{\theta}\ (x_1, x_2, \ldots, x_n) = (x_1 + x_2 + \ldots + x_n) / n$, which is the average of the sample (or *sample mean*). A

*Definitions of the *mean* of a distribution and a *percent point* of a distribution are given in Section 3 of the Appendix.

function $\hat{\theta}(x_1, x_2, \ldots, x_n)$ is called an *estimator* of θ. The value of $\hat{\theta}(x_1, x_2, \ldots, x_n)$ in any given sample is called a *point estimate* of θ. For convenience, the function $\hat{\theta}(x_1, x_2, \ldots, x_n)$ will be represented simply by $\hat{\theta}$.

The ideas presented above will now be illustrated. Suppose that one wishes to estimate the average trip length from home to place of work for heads of households in a certain community. Suppose further that a random sample of heads of household is obtained, and that the average distance for those in the sample turns out to be 2.7 miles. In this illustration θ is represented by the (unknown) population average, $\hat{\theta}$ is represented by the sample average, and the point estimate of θ is represented by the observed value of the sample average—namely 2.7 miles.

In advance of drawing a sample an estimator, $\hat{\theta}$, is a random variable. $\hat{\theta}$ is said to be an *unbiassed* estimator of θ if its expected value (mean) equals the value of θ. Exact or approximate unbiassedness is a desirable property of an estimator. The estimators used in this book have that and other desirable properties. Whenever possible, the estimators used herein are *maximum-likelihood* estimators. For a discussion of the principle of maximum-likelihood estimation and the properties of good estimators, see Mood (pp. 147–161).

Notation for Parameters of Certain Distributions And Estimates of the Parameters

Distribution	*Parameter(s)*	*Estimate* of Parameter*
Binomial	p	$\hat{p}$
Poisson	m	$\hat{m}$
Normal	u	$\hat{u}$
	σ^2	$\hat{\sigma}^2$

In Sections 2.2, 2.3, and 2.4 estimation of parameters of the binomial, Poisson, and normal distributions is considered. The notation used in these sections is indicated above.

*The same symbol is used for the estimate and the estimator. The estimator is a function and the estimate is a value of the function.

2.2. Estimation of the Parameter of a Binomial Distribution*

A population having only two kinds of elements is called a binomial population. Examples of the elements of such populations are: "Successes" and "Failures"; Heads and Tails; Commercial Vehicles, Non-commercial Vehicles; Vehicles Turning Left (on entering a certain intersection), Vehicles Not Turning Left (on entering the intersection); Local Vehicles, Non-local Vehicles; Trips Between Two Given Zones, Trips Not Between Those Two Zones.

Let p be the probability that an element drawn at random from the population is a "success." (p may be considered as the "proportion" of successes in the population, and $1-p$ as the "proportion" of failures in the population.) Let a success be represented by 1 and a failure by 0. A sample of size n from the population consists of n values, say $x_1, \ldots, x_n$, where each x equals 1 or 0. An estimator, say $\hat{p}$, for p is

$$\hat{p} = \frac{x_1 + \ldots + x_n}{n} = \frac{\Sigma x_i}{n}. \tag{2:1}$$

Thus $\hat{p}$ represents the fraction of cases which were successes—i.e., the relative frequency of successes in the sample. It is also of interest that $\hat{p}$ in (2:1) is the sample average (mean). The quantity $n\hat{p}$ has a binomial distribution (see Table III and Section 4 of the Appendix).

Example. Suppose there is need for an estimate of the proportion of local vehicles in the traffic flow on a major street in a certain community. Let p represent this proportion. Suppose further that 50 vehicles in the traffic are observed, and that 35 of them are local and 15 are non-local. Regarding these 50 cases as a sample from a binomial population with parameter p, compute a point estimate of p. Substituting in equation (2:1) one obtains the following result:

$$\hat{p} = \frac{35(1) + 15(0)}{50} = \frac{35}{50} = 0.7, \tag{2:2}$$

*It is assumed here that the population is infinitely large. Sampling from a *finite* binomial population is discussed in Section 3.2.b and in the Appendix.

which is the point estimate of p. (Note that $n=50$. Of the 50 x's exactly 35 are 1's and 15 are 0's.)

2.3. Estimation of the Parameter of a Poisson Distribution

The Poisson distribution is specified by the frequency function

$$f(x) = \frac{e^{-m}m^x}{x!} \quad (x = 0, 1, \ldots). \tag{2:3}$$

The parameter m is the mean of the distribution. Let $x_1, \ldots, x_n$ be a sample of n values of a random variable having the Poisson distribution. An estimator, say $\hat{m}$, for the parameter is:

$$\hat{m} = \frac{x_1 + \ldots + x_n}{n} = \frac{\Sigma x_i}{n} \tag{2:4}$$

which is simply the sample mean, $\bar{x}$. (It is of interest that $n\hat{m}$ (which equals the *sample sum* $S(x)$) has a Poisson distribution with parameter nm).

Example. In ten one-minute intervals during a certain period of the day the numbers of cars observed passing a given point on a street were, respectively, 2, 0, 6, 5, 1, 5, 3, 0, 3, 6. Assuming that these ten observations are a sample of values of a Poisson variable, compute a point estimate of the parameter m (m here would be the theoretical average number of cars *per minute*). Substituting in (2.4) one finds that the estimate is

$$\hat{m} = \frac{2 + \ldots + 6}{10} = \frac{31}{10} = 3.1.$$

2.4. Estimation of the Parameters of a Normal Distribution

Unlike the binomial and Poisson distributions, which have only one parameter, the normal distribution has two. These parameters are the mean of the distribution, denoted by u, and the variance of the distribution, denoted by σ^2 (see Table III). Let $\hat{u}$ and $\hat{\sigma}^2$ represent estimators of these two parameters, based on a sample of size n. The sample mean ($\bar{x}$) and sample variance (S_x^2) are good estimators of u and σ^2, respectively. Accordingly, $\hat{u}$ and $\hat{\sigma}^2$ are chosen as follows:

$$\hat{u} = \frac{x_1 + \ldots + x_n}{n} = \frac{\Sigma x_i}{n} = \bar{x} \text{ (the sample mean)},$$

$$\hat{\sigma}^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} = \frac{\Sigma(x_i - \bar{x})^2}{n} = S_x^2 \tag{2:5}$$

(the sample variance).*

$\hat{\sigma} = S_x$ is a point estimate of the standard deviation, σ.

Example. Assume that the speeds of 15 vehicles passing a certain observation point are as follows (in miles per hour); 41, 53, 48, 46, 39, 50, 49, 52, 38, 42, 55, 44, 55, 51, 47. Regarding these observations as a sample from a normal distribution, compute point estimates of the parameters of the distribution. Substituting in (2:5) one finds that the estimates of u and σ^2 are:

$$\hat{u} = 47.33,$$

$$\hat{\sigma}^2 = 28.89,$$

The estimate of σ is $\hat{\sigma} = \sqrt{28.89} = 5.37$.

2.5. Estimation of Percent Points (Percentiles)

Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be an arrangement of sample elements $x_1, x_2, \ldots, x_n$ in increasing order of magnitude. For example, $x_{(1)}$ is the least value in the sample and $x_{(n)}$ is the largest value. The quantities $x_{(1)}, \ldots, x_{(n)}$ are called *order statistics* (see Section 2 of the Appendix). It will be assumed here that the distribution function $F(x)$ is continuous and increasing for every value of x. (This implies that the probability equals 1 that no two x_i are equal.) $F(x)$ need not be a normal distribution; in fact, it can be "very" unsymmetrical.

An order statistic can be regarded as an "estimator" of a percent point of a distribution $F(x)$ (see Section 3 of the Appendix). An example of a percent point of $F(x)$ is the *median*, which is the 50 percent point (denoted by $x_{.50}$). This is a value of x such that half

*The expected value of $\hat{\sigma}^2$ in equation (2.5) is $(1 - \frac{1}{n})\,\sigma^2$. Thus $\hat{\sigma}^2$ is slightly biassed, although the bias decreases as n increases. In statistical work the following unbiassed estimator of σ^2 is sometimes used: $\Sigma(x_i - \bar{x})^2/(n - 1)$.

the population is less than or equal to the value, and half is greater. Other percent points can be described in a similar way. For example, the 85 percent point, $x_{.85}$, is a value of x such that 85 percent of the population is less than or equal to this value and 15 percent is greater. Although a percent point is an important characteristic of a distribution, it is not called a parameter of the distribution.

The order statistic $x_{(r)}$ $(1 \leqq r \leqq n)$ in a sample of size n is an estimator of the $100r / (n+1)$ percent point of $F(x)$.* For example, when $n = 29$ the order statistic $x_{(15)}$ is an estimator of the 50 percent point (i.e., the median) of $F(x)$. (Note that $100r / (n+1) = 50$ when $n = 29$ and $r = 15$.) It is also of interest that in a sample of size 29 the order statistic $x_{(15)}$ is the *sample median* (see Section 2 of the Appendix). In this illustration the value of the sample median would be a point estimate of the median of $F(x)$. This result can be generalized: the value of the sample median is always a point estimate of the median of $F(x)$.

The estimation considered in this section does not require that the functional form of the distribution be known. For that reason it is termed *non-parametric* or *distribution-free* estimation.

Example. Suppose that it is desired to estimate the 85 percent point of a distribution $F(x)$ and that a sample of size 100 is available. A suitable estimate here might be $x_{(86)}$ since $86 / 101$ closely approximates 0.85. (In fact, $86 / 101 = 0.851$.)

Incidentally, when the graph of $F(x)$ is nearly linear between $x_{85/101}$ and $x_{86/101}$, the function $0.15\, x_{(85)} + 0.85\, x_{(86)}$ is a satisfactory estimator of $x_{.85}$.

*Note that when no two elements of the sample are equal, there are $(n + 1)$ intervals between $-\infty$, the order statistics, and $+\infty$.

Chapter 3: Interval Estimation

A point estimate of the value of a quantity becomes more meaningful when it is accompanied by an indication of the possible error of the estimate. One way of indicating the possible error is to specify a range (interval) that is likely to include the value of the quantity. Such an interval is referred to as an *interval estimate* of the value of the quantity. For example, on the basis of a survey one might say there is good reason to believe that the average length of trips in a community is between 4 and 6 miles. The interval in this illustration is, of course, specified by its endpoints —namely, 4 and 6 (miles).

In statistical work a sample is used to form an interval estimate which is called a *confidence interval.* A general description of the procedure is given in Section 3.1 below.

This chapter gives confidence intervals for parameters of binomial, Poisson, and normal distributions. Confidence intervals for population means and for percent points are also given.

3.1. Definitions and Notation

Let θ be a parameter of the distribution of a random variable, X, and let $x_1, \ldots, x_n$ be a random sample of n values of X. Let θ' and θ'' be two functions of the sample; thus $\theta' = \theta'(x_1, \ldots, x_n)$ and $\theta'' = \theta''(x_1, \ldots, x_n)$. Assume furthermore that for any sample, θ' is less than θ''. The quantities θ' and θ'' are represented in many of the formulas in this chapter; for instance, in (3.6) the quantities u' and u'' are examples of θ' and θ'', respectively.

In *advance* of drawing a sample both θ' and θ'' are random variables. Let it be assumed that the following equation is true, irrespective of the value of θ:

$$\Pr(\theta' < \theta < \theta'') = \lambda, \tag{3:1}$$

where λ is a preassigned probability.*

Formula (3:1) states: the probability that the random interval (θ', θ'') covers θ is equal to λ. The random interval (θ', θ'') can

*The notation "Pr ( )" means "the probability that" (see Section 3 of the Appendix.)

be regarded as an "interval estimator" of θ. When a particular sample has been obtained, the particular interval thereby obtained is termed a 100λ *percent confidence interval* for θ. λ is called the *confidence coefficient* associated with the confidence interval. The particular values, say θ_0' and θ_0'', assumed by θ' and θ'', respectively, are termed lower and upper 100λ *percent confidence limits* for θ.

When θ' and θ'' are such that the probability on the left in equation (3:1) is always not less than λ (but not always equal to λ), (θ', θ'') is said to be a *conservative* 100λ *percent confidence interval* for θ and the confidence limits are said to be conservative. Conservative confidence intervals are of practical importance since it may be possible to calculate such intervals when "exact" confidence intervals cannot be calculated.

Confidence coefficients such as 0.68, 0.95 and 0.99 are often used in statistical work. Since a confidence coefficient of 0.95 is usually satisfactory in traffic engineering studies, this value is used frequently in this book. It should be added, however, that the choice of a confidence coefficient is a matter for the investigator to decide.

The interval estimators described above can be termed "two-sided" since there are two random endpoints, θ' and θ''. A "one-sided" interval estimator for θ can also be set up. For example, a function (say θ''') of a sample can be chosen so that $\Pr(\theta''' < \theta) = \lambda$. For a particular sample the value of θ''' would simply be a *lower* 100λ percent confidence limit for θ, and no upper confidence limit would be specified. Similarly, an *upper* confidence limit can be set up without specifying a lower confidence limit. Illustrative examples of one-sided and two-sided confidence intervals are given in Section 3.2.a.

3.2. Interval Estimation of the Parameter of a Binomial Distribution

The binomial distribution has one parameter (see Table III). This parameter, denoted by p, is the probability that an element drawn at random from a binomial population is a "success" (see Section 2.2). Confidence intervals for p are slightly conservative since the binomial distribution is discrete.

3.2.a. Tables and Charts. Confidence intervals for the binomial parameter can be obtained from various tables in the statistical literature. Three sets of tables will be described briefly. (For complete references to each set see the Bibliography.)

(1) *The Harvard Computation Laboratory's tables of the cumulative binomial distribution.* These tables can be used to obtain one-sided and two-sided confidence intervals for any given confidence coefficient. The sample sizes included in this table range from 1 to 1000.

(2) *Pachares' tables of confidence limits.* These tables give one-sided confidence intervals for the following confidence coefficients: 0.95, 0.975, 0.99, and 0.995. Both types of one-sided intervals are given—namely, the type specified by an upper confidence limit and the type specified by a lower confidence limit. From these confidence limits one immediately obtains two-sided confidence intervals for the following confidence coefficients: 0.90, 0.95, 0.98, and 0.99. The sample sizes included in the tables are 55, 60, 65, ..., 100. Various other tables of confidence limits for the binomial parameter are listed in the paper by Pachares.

(3) *Table XI in Hald's Statistical Tables.* This table gives two-sided 95 and 99 percent confidence intervals for various values of x and of $(n-x)$ from 0 to 500. x represents the number of "successes" and n represents the sample size. From these tables one immediately obtains 97.5 percent and 99.5 percent one-sided confidence intervals.

There are several published charts for conveniently determining confidence intervals for p (for example, see Dixon and Massey). Figure 1 below gives the Clopper-Pearson chart for two-sided 95 percent confidence intervals. The sample sizes associated with the chart are 10, 15, 20, 30, 50, 100, 250, and 1000. Rough interpolation for an intermediate sample size can be carried out easily. The chart can also be used to obtain one-sided 97.5 percent confidence intervals for p. (See the second illustrative example below.)

Example. Using the data in the example in Section 2.2, find a two-sided 95 percent confidence interval for the proportion, p, of local vehicles. (It will be recalled that in the example 50 vehicles were observed, of which 35 were local and 15 were non-local.)

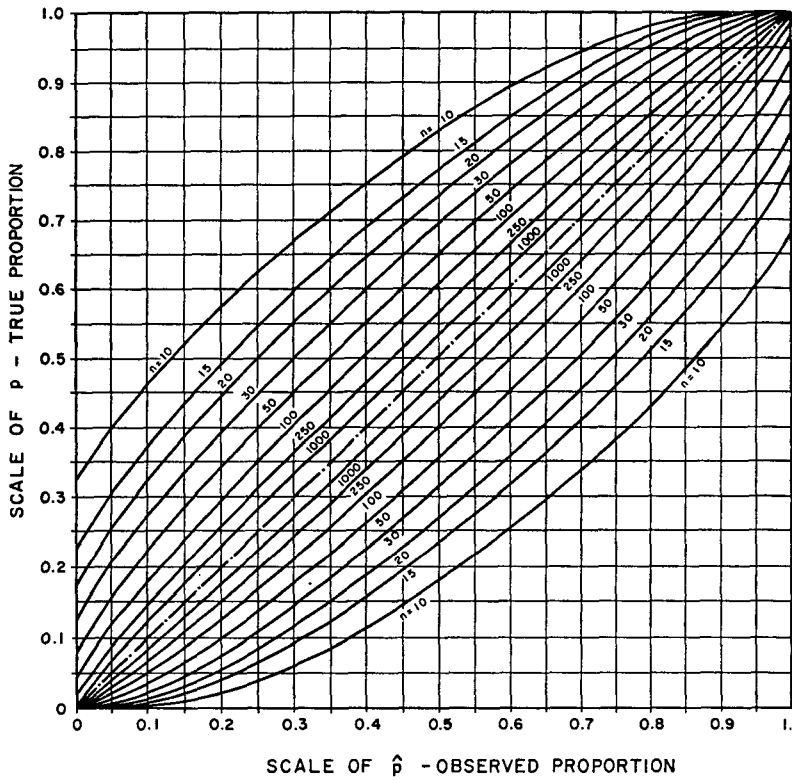


Figure 1. Confidence Belts for Proportions—
Confidence Coefficient 0.95.

Note: This chart is reproduced with the permission of the authors C. J. Clopper and E. S. Pearson and the publishers, The Biometrika Office.

The observed proportion of local vehicles is $35/50 = 0.70$. Using the belt in Figure 1 for a sample of size 50, one finds that the lower and upper 95 percent confidence limits are, approximately, 0.56 and 0.83, respectively. The 95 percent confidence interval for p is therefore

$$0.56 < p < 0.83.$$

In view of these results, the traffic engineer can conclude, with 95 percent confidence, that the proportion, p, of local vehicles is between 0.56 and 0.83.

Second Example. Suppose that in a sample of size 100 from a binomial population there are 40 "successes" and 60 "failures." Find an upper 97.5 percent confidence limit for the probability, p, of a "success." The observed proportion is $40/100 = 0.40$. Using the upper boundary of the belt in Figure 1 for a sample of size 100, one finds that the upper 97.5 percent confidence limit is 0.51. In other words, a (one-sided) 97.5 percent confidence interval for p is as follows:

$$p < 0.51.*$$

(Note that, from Figure 1, the two-sided 95 percent confidence interval is found to be $0.31 < p < 0.51$.)

3.2.b. Formulas. Formulas for calculating two-sided approximate confidence intervals for the binomial parameter are given in Table IV. These formulas are based on the normal approximations to the binomial and hypergeometric distributions.† A discussion of the accuracy of these approximations is given in the book by Hald (pp. 676–691).

The symbols used in Table IV are defined below:

(1) p' and p'' are lower and upper confidence limits, respectively (thus the confidence interval is $p' < p < p''$);

(2) n = sample size;

(3) $\hat{p}$ = (observed number of "successes" in n trials) / n;

(4) N = number of elements in the population when the population is finite;

*Since $p \geq 0$ by definition, the statement that $p < 0.51$ is equivalent to the statement that $0 \leq p < 0.51$.

†The hypergeometric distribution is associated with sampling from a *finite* binomial population (see Section 4 of the Appendix).

Table IV—Formulas for Approximate 100λ Percent Confidence Intervals for the Binomial Parameter*

Infinite Population		*Finite Population*	
$p' = \dfrac{\hat{p}+\dfrac{z_a^2}{2n}-z_a\sqrt{\left(\dfrac{\hat{p}(1-\hat{p})}{n}+\dfrac{z_a^2}{4n^2}\right)}}{1+\dfrac{z_a^2}{n}}$ $p'' = \dfrac{\hat{p}+\dfrac{z_a^2}{2n}+z_a\sqrt{\left(\dfrac{\hat{p}(1-\hat{p})}{n}+\dfrac{z_a^2}{4n^2}\right)}}{1+\dfrac{z_a^2}{n}}$	(3:2)	Same as (3:2) with the exception that z_a is replaced by $z_a\sqrt{\left(\dfrac{N-n}{N-1}\right)}$	(3:2A)
$p' = \hat{p} - z_a\sqrt{\left(\dfrac{\hat{p}(1-\hat{p})}{n}\right)}$ $p'' = \hat{p} + z_a\sqrt{\left(\dfrac{\hat{p}(1-\hat{p})}{n}\right)}$	(3:3)	Same as (3:3) with the exception that z_a is replaced by $z_a\sqrt{\left(\dfrac{N-n}{N-1}\right)}$	(3:3A)

*The notation used in Table IV is defined on pages 25 and 27.

(5) λ = confidence coefficient;

(6) $a = (1+\lambda) / 2$;

(7) z_a = the 100a percent point of the standard normal distribution.*

The relation between z_a and the confidence coefficient λ is indicated in Table V. For example, $z_a = 1.960$ when $\lambda = 0.95$.

Table V: Values of z_a Associated with Certain Confidence Coefficients†

Confidence Coefficient λ	$a=(1+\lambda) / 2$	z_a
0.80	0.90	1.282
0.90	0.95	1.645
0.95	0.975	1.960
0.98	0.99	2.326
0.99	0.995	2.576
0.998	0.999	3.090
0.999	0.9995	3.291
0.9999	0.99995	3.891
0.99999	0.999995	4.417

The confidence limits obtained from (3:2) and (3:3) are called large-sample confidence limits since they closely approximate exact limits when n is large. Formula (3:3) is somewhat less accurate than formula (3:2), but it is computationally simpler. For large values of n the difference between formulas (3:2) and (3:3) is negligible. (See the illustrative example below.)

When the binomial population is finite (and the sampling is without replacement), (3:2) should be replaced by (3:2A) and (3:3) should be replaced by (3:3A). As indicated in Table IV, the formulas for the finite case are obtained from those for the infinite case simply by replacing z_a by $z_a\sqrt{[(N-n) / (N-1)]}$. The quantity $\sqrt{[(N-n) / (N-1)]}$ is known as the "finite population correction" or "finite population factor." This factor lies between 0 and 1. For fixed N the factor decreases as n increases.

*See Section 3 of the Appendix for a definition of a percent point of a distribution.

†The values of z_a and a are given as x and $F(x)$, respectively, beneath Appendix Table 1.

Example. Using the example in Section 2.2, find an approximate 95 percent confidence interval for the proportion, p, of local vehicles in the traffic flow. (50 vehicles were observed of which 35 were local and 15 were non-local.) With regard to this example the quantities n, $\hat{p}$, λ, a, and z_a in Table IV have the following values:

$$n = 50,$$
$$\hat{p} = 35/50 = 0.7,$$
$$\lambda = 0.95,$$
$$a = (1 + 0.95) / 2 = 0.975,$$
$$z_a = 1.96 \text{ (see Table V)}.$$

Substituting in formula (3:2), one finds that

$$p' \doteq \frac{0.70 + 0.04 - (1.96)\sqrt{(0.0046)}}{1.08} \doteq (0.69 - 0.13) = 0.56,$$

$$p'' \doteq 0.69 + 0.13 = 0.82.$$

(The symbol "$\doteq$" means "equals approximately.")

Accordingly, the approximate 95 percent confidence interval for p is

$$0.56 < p < 0.82.$$

Substituting in formula (3:3), one finds that

$$p' = 0.700 - 1.96\sqrt{(0.0042)} \doteq 0.700 - 0.127 \doteq 0.57,$$

$$p'' \doteq 0.700 + 0.127 \doteq 0.83.$$

The approximate 95 percent confidence interval derived from (3:3) is therefore

$$0.57 < p < 0.83.$$

The two confidence intervals above are both based on the data used in the first example in Section 3.2.a. In that example the confidence interval was obtained by means of the chart in Figure 1. Each of the three intervals closely approximates the correct interval. The end-points of the correct interval (obtained from Hald's tables) are as follows (to two decimal places):

$$p' = 0.55 \text{ and } p'' = 0.82.$$

Example For the Case of a Finite Binomial Population. Suppose that in an origin-destination study it is reasonable to assume that a certain zone, say zone 1, generated 10,000 trips. Suppose further that in a 10 percent sample of the trips there were 200 between zones 1 and 2. Find an approximate 95 percent confidence interval for the proportion, p, of the total (10,000) trips that are between zones 1 and 2. The information relevant to the problem can be summarized as follows:

$$n = 1{,}000,$$

$$\hat{p} = \frac{200}{1{,}000} = 0.2,$$

$$N = 10{,}000,$$

$$\lambda = 0.95,$$

$$a = \frac{1 + 0.95}{2} = 0.975,$$

$$z_a = 1.96 \text{ (see Table V)}.$$

Formulas (3:2A) and (3:3A) are nearly equal when $n = 1{,}000$. Since (3:3A) is computationally simpler, it will be used in this example. Substituting in (3:3A), one finds that

$$p' \doteq 0.2 - (1.96)\sqrt{(0.00016)}\sqrt{(0.90009)} \doteq 0.2 - (1.96)\sqrt{0.000144}$$
$$\doteq 0.2 - 0.024 = 0.176,$$

$$p'' \doteq 0.2 + (1.96)\sqrt{(0.00016)}\sqrt{(0.90009)} \doteq 0.2 + 0.024 = 0.224.$$

It follows that an approximate 95 percent confidence interval for p is

$$0.176 < p < 0.224.$$

(Expressed in trips, the 95 percent confidence interval would range from 1760 to 2240 trips.)

Using formula (3:3) (thus assuming an infinite population), one obtains the following approximate 95 percent confidence limits for p:

$$p' = 0.2 - (1.96)\sqrt{(0.00016)} \doteq 0.175$$

$$p'' = 0.2 + (1.96)\sqrt{(0.00016)} \doteq 0.225.$$

The approximate 95 percent confidence interval for p is therefore

$$0.175 < p < 0.225.$$

Clearly the length of the above interval is only slightly greater than that of the interval obtained from (3:3A). This is because n is so small in relation to N that the finite population factor is nearly equal to 1.

3.3. Interval Estimation of the Parameter of a Poisson Distribution

Table VI gives 95 and 99 percent confidence limits for the Poisson parameter for observed values (of the Poisson variable) from 0 to 50. (Table VI can also be used to obtain one-sided 97.5 and 99.5 percent confidence intervals for the Poisson parameter.) For a wider range of observed values such confidence limits can be computed easily by means of Table II in Molina. Approximate confidence limits for the Poisson parameter can be computed by means of formulas (3:4) and (3:5) below.

3.3.a. Use of Table VI. Let $x_1, x_2, \ldots, x_n$ be a random sample from a Poisson distribution, and let m be the parameter of the distribution. The sample sum $S(x) = \sum_1^n x_i$ then has a Poisson distribution with parameter nm. Using the sample sum one obtains confidence limits for nm from Table VI; dividing those limits by n one then has confidence limits for m. (nm is the quantity represented by M in the Note under Table VI.)

Example. Suppose that for five one-minute counts the average number of vehicles (per minute) passing a point on a rural road is 1.4. Assuming that the observations come from a Poisson distribution, find a (two-sided) 95 percent confidence interval for the Poisson parameter, m. (m is the expected (mean) number of vehicles passing the point in any given minute.) In this example $n = 5$ and the sample sum, $S(x)$, equals $5(1.4) = 7$. From Table VI one finds that lower and upper 95 percent confidence limits for $5m$ are 2.8 and 14.4, respectively. Dividing these limits by 5, one obtains 0.56 and 2.88 as the lower and upper confidence limits for m. The 95 percent confidence interval for m is therefore

$$0.56 < m < 2.88.$$

Table VI: Confidence Limits for the Parameter of the Poisson Distribution

Observed Value of Poisson Variable x	*Confidence Coefficient* 0.99 *Lower Limit*	0.99 *Upper Limit*	0.95 *Lower Limit*	0.95 *Upper Limit*	x	*Confidence Coefficient* 0.99 *Lower Limit*	0.99 *Upper Limit*	0.95 *Lower Limit*	0.95 *Upper Limit*
0	0.0	5.3	0.0	3.7					
1	0.0	7.4	0.1	5.6	26	14.7	42.2	17.0	38.0
2	0.1	9.3	0.2	7.2	27	15.4	43.5	17.8	39.2
3	0.3	11.0	0.6	8.8	28	16.2	44.8	18.6	40.4
4	0.6	12.6	1.0	10.2	29	17.0	46.0	19.4	41.6
5	1.0	14.1	1.6	11.7	30	17.7	47.2	20.2	42.8
6	1.5	15.6	2.2	13.1	31	18.5	48.4	21.0	44.0
7	2.0	17.1	2.8	14.4	32	19.3	49.6	21.8	45.1
8	2.5	18.5	3.4	15.8	33	20.0	50.8	22.7	46.3
9	3.1	20.0	4.0	17.1	34	20.8	52.1	23.5	47.5
10	3.7	21.3	4.7	18.4	35	21.6	53.3	24.3	48.7
11	4.3	22.6	5.4	19.7	36	22.4	54.5	25.1	49.8
12	4.9	24.0	6.2	21.0	37	23.2	55.7	26.0	51.0
13	5.5	25.4	6.9	22.3	38	24.0	56.9	26.8	52.2
14	6.2	26.7	7.7	23.5	39	24.8	58.1	27.7	53.3
15	6.8	28.1	8.4	24.8	40	25.6	59.3	28.6	54.5
16	7.5	29.4	9.2	26.0	41	26.4	60.5	29.4	55.6
17	8.2	30.7	9.9	27.2	42	27.2	61.7	30.3	56.8
18	8.9	32.0	10.7	28.4	43	28.0	62.9	31.1	57.9
19	9.6	33.3	11.5	29.6	44	28.8	64.1	32.0	59.0
20	10.3	34.6	12.2	30.8	45	29.6	65.3	32.8	60.2
21	11.0	35.9	13.0	32.0	46	30.4	66.5	33.6	61.3
22	11.8	37.2	13.8	33.2	47	31.2	67.7	34.5	62.5
23	12.5	38.4	14.6	34.4	48	32.0	68.9	35.3	63.6
24	13.2	39.7	15.4	35.6	49	32.8	70.1	36.1	64.8
25	14.0	41.0	16.2	36.8	50	33.6	71.3	37.0	65.9

Note: The Poisson distribution is specified by the frequency function $f(x) = \frac{e^{-M}M^x}{x!}$ $(x = 0, 1, 2, \ldots)$. M is the mean of the distribution. For an illustration of the use of the table, suppose the observed value, x, equals 40; then 95 percent confidence limits for the parameter, M, are 28.6 and 54.5, respectively. (Table VI is reproduced with the permission of the author, W. E. Ricker, and the editor of the *Journal of the American Statistical Association.*)

3.3.b. Formulas. Formula (3:4) below gives approximate lower and upper 100λ percent confidence limits for the Poisson parameter, m.

$$m' = \hat{m} + \frac{z_a^2}{4n} - z_a\sqrt{\left(\frac{\hat{m}}{n} + \frac{3}{8n^2}\right)},$$
$$m'' = \hat{m} + \frac{z_a^2}{4n} + z_a\sqrt{\left(\frac{\hat{m}}{n} + \frac{3}{8n^2}\right)}, \tag{3:4}$$

where

$\hat{m}$ = the sample mean,

n = sample size,

z_a = 100a percent point of the standard normal distribution,

$$a = \frac{1+\lambda}{2}.$$

When the sample *sum*, $n\hat{m}$, exceeds 50, the confidence limits obtained from (3:4) closely approximate the exact limits. When the sample sum is less than or equal to 50, exact confidence limits (to one decimal place) can be obtained from Table VI.

Ignoring the terms $z_a^2 / 4n$ and $3 / 8n^2$ in (3:4) one obtains the following simpler, but less accurate, formulas for m' and m'':

$$m' = \hat{m} - z_a\sqrt{\frac{\hat{m}}{n}},$$
$$m'' = \hat{m} + z_a\sqrt{\frac{\hat{m}}{n}}. \tag{3:5}$$

For large n the difference between (3:4) and (3:5) is negligible.

Example. Suppose that for 100 one-minute counts the average number of vehicles (per minute) passing a point on a rural road is 5.5. Find an approximate (two-sided) 99 percent confidence interval for the expected (mean) number, m, of vehicles in a one-minute period, assuming that the observations form a sample of values of a Poisson variable. In this example, $\hat{m} = 5.5$, $n = 100$, $\lambda = 0.99$, $a = (1+\lambda) / 2 = 0.995$ and $z_a = 2.576$ (see Table V).

Substituting in formula (3:5) one finds that

$$m' = 5.500 - (2.576)\sqrt{(0.055)} \doteq 5.500 - 0.603 \doteq 4.9,$$

$$m'' \doteq 5.500 + 0.603 \doteq 6.1.$$

The following interval is therefore an approximate 99 percent confidence interval for m:

$$4.9 < m < 6.1.$$

3.4. Interval Estimation of the Parameters of a Normal Distribution

The normal distribution is of central importance in the field of statistics.* One of its many important properties is that it approximates various other distributions. This property underlies several formulas given in this section. For example, formula (3:2) is based on the normal approximation to the binomial, and formula (3:4) is based on the normal approximation to the distribution of $\sqrt{(Y+3/8)}$, where Y is a Poisson variable.

Since it approximates other important distributions, the normal distribution is naturally of interest in traffic engineering. A second (and perhaps more important) reason for this interest is that certain observable quantities in traffic engineering have normal or nearly normal distributions. Two examples are "spot" speeds and reaction times. The second reason indicates the utility of interval estimates of parameters of the normal distribution.

There are two parameters in the normal distribution—namely, the mean, u, and the variance, σ^2 (see Table III). The "center" of the distribution is represented by u, and the "spread" of the distribution is represented by the standard deviation, σ (see Section 3 of the Appendix). Confidence limits for u and σ^2 (and σ) are given below.

3.4.a. Confidence Limits for the Mean. Suppose that a random sample has been drawn from a normal distribution, and assume that the values of the mean, u, and variance, σ^2, are unknown to the investigator. *Student's t-distribution* provides a means of obtaining confidence limits for u when the value of σ^2 is unknown. Let u'

*For a comprehensive account of its role in statistics see Mood (pp. 142–143).

and u'' denote, respectively, lower and upper 100λ percent confidence limits for u. The quantities u' and u'' can be expressed as follows:

$$u' = \bar{x} - \frac{t_a S_x}{\sqrt{(n-1)}}, \qquad u'' = \bar{x} + \frac{t_a S_x}{\sqrt{(n-1)}}, \tag{3:6}$$

where

n = the sample size,
$\bar{x}$ = the sample mean,
S_x = the sample standard deviation,
$a = (1+\lambda) / 2$ (λ = the confidence coefficient),
t_a = the 100a percent point of *Student's t-distribution* with $n-1$ *degrees of freedom* (see Appendix Table 3).

The quantity called *degrees of freedom* represents a special constant involved in Student's t-distribution (see Mood [p. 206]).

Example. In the example in Section 2.4 the mean speed of 15 vehicles passing a given location was 47.33 miles per hour and the standard deviation was 5.37 miles per hour. (It is assumed that the observations form a sample from a normal population whose mean and variance are unknown to the investigator.) Find a 99 percent confidence interval for the mean, u, of the population. In this example $n=15$, $\bar{x}=47.33$, $S_x=5.37$, $\lambda=0.99$, $a=0.995$ and $t_a=2.977$.* Substituting these values in (3:6) one finds that $u' = 47.33 - (2.977)\ (5.37) / \sqrt{(14)} \doteq 47.33 - 4.28 = 43.05$ and $u'' = 47.33 + (2.977)\ (5.37) / \sqrt{(14)} \doteq 47.33 + 4.28 = 51.61$. The following interval is therefore a 99 percent confidence interval for u:

$$43.05 < u < 51.61.$$

In other words, one concludes with 99 percent confidence that the mean of the population of speeds lies between 43.05 and 51.61 miles per hour.

It should be emphasized that in the above example the value of σ is assumed to be unknown. When the value of σ can be regarded

*The value 2.977 is the 99.5 percent point of Student's t-distribution for 14 degrees of freedom (see Appendix Table 3).

as known, u' and u'' should be calculated by means of the following formula instead of (3:6):

$$u' = \bar{x} - \frac{z_a \sigma}{\sqrt{n}}, \qquad u'' = \bar{x} + \frac{z_a \sigma}{\sqrt{n}}, \tag{3:7}$$

where z_a is the 100a percent point of the standard normal distribution and σ is, of course, the (known) standard deviation of the population.

3.4.b. Confidence Limits for the Variance and Standard Deviation. Suppose that a sample has been drawn from a normal distribution, and assume that the values of the mean, u, and variance, σ^2, are unknown to the investigator. Let $(\sigma')^2$ and $(\sigma'')^2$ be, respectively, lower and upper 100λ percent confidence limits for σ^2. These confidence limits can be expressed as follows:

$$(\sigma')^2 = \frac{n S_x^2}{\chi^2_{v'}}, \qquad (\sigma'')^2 = \frac{n S_x^2}{\chi^2_{v''}}, \tag{3:8}$$

where

n = the sample size,

S_x^2 = the sample variance,

$v' = \dfrac{1+\lambda}{2}$,

$v'' = \dfrac{1-\lambda}{2}$,

λ = the confidence coefficient,

$\chi^2_{v'}$ = the $100v'$ percent point of the *Chi-square distribution* with $(n-1)$ degrees of freedom,

$\chi^2_{v''}$ = the $100v''$ percent point of the *Chi-square distribution* with $(n-1)$ degrees of freedom.

The *Chi-square distribution* has a very wide range of statistical applications. For a detailed discussion of its nature and utility see, for example, Duncan, Greenshields and Weida, or Mood. Percent points of this distribution are given in Appendix Table 2.

It should be noted that σ' and σ'' are, respectively, lower and upper confidence limits for the *standard deviation*, σ.

Example. Suppose that one wishes to study the variability of speeds at a certain roadway location in relation to speed zoning. Suppose further that 21 speeds have been observed there and that these observations can be regarded as a sample from a normal distribution. The mean and standard deviation of the distribution are unknown. Of these two parameters the only one of interest is the standard deviation since it is associated with variability ("spread") of the distribution (whereas the mean is not). Assuming that the sample variance, S_x^2, equals 4.87*, find a 95 percent confidence interval for the standard deviation, σ. In this example $n=21$, $S_x^2=4.87$, $\lambda=0.95$, $v'=0.975$, $v''=0.025$; thus $\chi^2_{v'}=34.2$ and $\chi^2_{v''}=9.59$. (The values 9.59 and 34.2 are, respectively, the 2.5 and 97.5 percent points of the Chi-square distribution with 20 degrees of freedom (see Appendix Table 2).) Substituting in (3:8), one finds that

$$(\sigma')^2=\frac{(21)\ (4.87)}{34.2}\doteq 2.99,$$

$$(\sigma'')^2=\frac{(21)\ (4.87)}{9.59}\doteq 10.7.$$

It follows that lower and upper confidence limits for σ are, respectively, $\sigma'\doteq\sqrt{(2.99)}\doteq 1.73$ and $\sigma''\doteq\sqrt{(10.7)}\doteq 3.27$. A 95 percent confidence interval for σ is, therefore, in miles per hour

$$1.73<\sigma<3.27.$$

3.5. Interval Estimation of the Mean of a Distribution

In this section formulas are given for confidence limits on the mean of a distribution (i.e., the mean of a population). No special

*The sample variance is a *point estimate* of σ^2; thus a point estimate of σ in this example is $\sqrt{(4.87)}\doteq 2.21$ (see Section 2.4).

assumption is made about the form of the distribution; for example, it is not assumed to be a normal distribution.

Let μ and σ be the mean and standard deviation, respectively, of a distribution, and suppose a sample $x_1, x_2, \ldots, x_n$ is to be drawn from the distribution. The expected value (mean) of the sample mean, $\bar{x}$, is μ. The sample mean is, therefore, an unbiassed point estimate of the population mean. The variance, $\sigma^2_{\bar{x}}$, of $\bar{x}$ is $(\sigma^2/n)\,[N-n)\,/\,(N-1)]$ or (σ^2/n) according as the population from which the sample is drawn is finite* or infinite. (N is the size of the population when the population is finite.)

In a wide variety of practical problems it is reasonable to assume that the distribution of $(\bar{x}-\mu)\,/\,\sigma_{\bar{x}}$ is approximately a standard normal distribution. This assumption is especially suitable if n is large. When this assumption is appropriate and the population is finite, approximate lower and upper 100λ percent confidence limits for μ are as follows:

$$\begin{aligned} \mu' &= \bar{x} - z_a \frac{\sigma}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}, \\ \mu'' &= \bar{x} + z_a \frac{\sigma}{\sqrt{n}} \sqrt{\left(\frac{N-n}{N-1}\right)}, \end{aligned} \tag{3:9}$$

where $a = (1+\lambda)\,/\,2$ and z_a is the $100a$ percent point of the standard normal distribution. Formula (3:9) can be used when the investigator can regard the value of σ as known. If the value of σ cannot be regarded as known, the following formula should be used in place of (3:9):

$$\begin{aligned} \mu' &= \bar{x} - t_a \frac{S_x}{\sqrt{n-1}} \sqrt{\left(\frac{N-n}{N-1}\right)}, \\ \mu'' &= \bar{x} + t_a \frac{S_x}{\sqrt{n-1}} \sqrt{\left(\frac{N-n}{N-1}\right)}, \end{aligned} \tag{3:10}$$

where S_x is the sample standard deviation and t_a is the $100a$ percent

*It is assumed here that the sampling from a finite population is *without replacement* (see Section 1 of the Appendix). If, in fact, the sampling is *with replacement*, $\sigma^2_{\bar{x}} = \sigma^2/n$.

point of the t-distribution with $(n-1)$ degrees of freedom* (see Cochran (p. 20)).

When the population is infinite, the finite population factor, $\sqrt{[(N-n)/(N-1)]}$, in (3:9) and (3:10) should be omitted. It should be emphasized that the validity of (3:9) and (3:10) rests on the assumption of approximate normality of $\bar{x}$. For an interesting discussion of the adequacy of this assumption see Cochran (pp. 22–28).

When there is good reason to doubt that $\bar{x}$ is approximately normally distributed, one can construct *conservative* confidence limits for μ. Formula (3:11)† below gives conservative lower and upper $100\,(1-1/B^2)$ percent confidence limits for μ when the population is finite.

$$\mu' = \bar{x} - B\frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)},$$
$$\mu'' = \bar{x} + B\frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}. \qquad (3:11)$$

When the population is infinite, the finite population factor, $\sqrt{[(N-n)/(N-1)]}$, in (3:11) should, of course, be omitted. The use of (3:11) requires that the value of σ be regarded as known; however, it does not require that $\bar{x}$ be approximately normally distributed. Formula (3:11) is valid irrespective of the distribution of $\bar{x}$.‡ This means, among other things, that (3:11) is valid irrespective of the value of n.

Some of the practical applications of formulas in this section are given in Section 5.7 which deals with short-count estimation of traffic volume.

3.6. Interval Estimation of Percent Points of a Distribution

There are various problems in traffic engineering in which estimates of percent points are of interest. For example, an engineer

*It will be noted from Appendix Table 3 that $t_a \doteq z_a$ for $n \geq 30$, say.

†For example, if $B=2$, then the confidence coefficient is not less than $(1-1/2^2)=0.75$. Formula (3:11) is based on the Bienaymé-Tchebycheff inequality (see Section 3 of the Appendix).

‡Provided that σ is finite.

might wish to estimate the *median* walking distance of parkers, or the 85 *percent point* of the distribution of "spot" speeds at a certain point on a roadway. Accordingly, this section gives formulas for confidence limits on percent points.

Let $x_1, x_2, \ldots, x_n$ be a sample of n values of a random variable X. The only assumption made about the distribution function, $F(x)$, of X is that it is continuous and increasing for each possible value of X. The methods of estimation used here are called *non-parametric* (or *distribution-free*) since the functional form of the distribution is not assumed to be known. (Such methods were also used in Section 2.5.) Non-parametric methods are very simple computationally (although somewhat less accurate than parametric methods). Because of their simplicity, they are sometimes used even when the functional form of the distribution is known. These methods involve the use of order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ of the sample. (For a description of order statistics see Section 2 of the Appendix.)

The interval $(x_{(r)}, x_{(s)})$ $(r<s)$ is a confidence interval for the $100b$ percent point, x_b, of $F(x)$ with confidence coefficient*

$$\sum_{i=r}^{s-1} C_i^n \, b^i \, (1-b)^{n-i} \quad (0<b<1), \tag{3:12}$$

where C_i^n is a binomial coefficient (defined in Section 4 of the Appendix). The expression in (3:12) is a sum of terms of the binomial distribution. This sum can be evaluated by means of tables of the cumulative binomial distribution (e.g., the tables of the Harvard Computation Laboratory). When n is not small, the sum can be approximated by means of the normal approximation to the binomial distribution; formula (3:13) below is based on this approximation. For a given value of b and a given confidence coefficient λ, a good choice of values of r and s is as follows:

$$\begin{aligned} r &= nb - z_a \sqrt{[nb\,(1-b)]}, \\ s &= nb + z_a \sqrt{[nb\,(1-b)]}, \end{aligned} \tag{3:13}$$

where $a=(1+\lambda)/2$ and z_a is the $100a$ percent point of the standard normal distribution.

*See Mood (p. 389).

Example. Suppose that from a sample of 100 distances walked by parkers one wishes to obtain an approximate 95 percent confidence interval for the *median* of the population of walking distances. (It should be recalled that the population median is the 50 percent point of the distribution from which the sample is drawn.) In this example $n=100$, $b=0.50$, $\lambda=0.95$, $a=0.975$, and $z_a=1.96$. Substituting in (3:13) one finds that

$$r = 100(1/2) - (1.96)\sqrt{[100(1/2)\ (1/2)]} = 50 - 9.80 \doteq 40,$$
$$s = 50 + 9.80 \doteq 60.$$

Accordingly, $(x_{(40)}, x_{(60)})$ would be an approximate 95 percent confidence interval for the median, $x_{.50}$. For instance, suppose $x_{(40)}=0.40$ miles and $x_{(60)}=0.65$ miles. An approximate 95 percent confidence interval for $x_{.50}$ would then be

$$0.40 \text{ miles} < x_{.50} < 0.65 \text{ miles}.$$

Second Example. Suppose that from a sample of 200 "spot" speeds one wishes to obtain an approximate 95 percent confidence interval for the 85 percent point, $x_{.85}$, of the distribution of "spot" speeds. In this example $n=200$, $b=0.85$, $\lambda=0.95$, $a=0.975$, and $z_a=1.96$. Substituting in (3:13) one finds that

$$r = 200(0.85) - 1.96\sqrt{[200(0.85)\ (0.15)]} \doteq 160,$$
$$s = 200(0.85) + 1.96\sqrt{[200(0.85)\ (0.15)]} \doteq 180.$$

It follows that $(x_{(160)}, x_{(180)})$ would be an approximate 95 percent confidence interval for $x_{.85}$. For example, if $x_{(160)}=45$ mph and $x_{(180)}=50$ mph, then an approximate 95 percent confidence interval for $x_{.85}$ would be:

$$45 \text{ mph} < x_{.85} < 50 \text{ mph}.$$

Chapter 4: Tests of Hypotheses (Significance Tests)

A statistical hypothesis is a hypothesis (assumption) about the distribution of one or more random variables. A test of a statistical hypothesis is a procedure for deciding, on the basis of a sample, whether to "accept" or "reject" the hypothesis. Such a test is also called a *significance test.* The subject of significance tests is a classical subject in statistics.

In this chapter the essential elements of a significance test will be presented and several significance tests of practical importance in traffic engineering will be given. For a more general treatment of significance tests see, for example, Hoel, Mood, or Wallis and Roberts.

4.1. Introduction

Significance testing is a basic tool of the research worker in traffic engineering. Illustrative examples of traffic problems that can be dealt with by significance testing are:

(1) *After* a change in its design, is the safety of a certain intersection the same as it was *before* the change? (For example, suppose there were 10 accidents at the intersection the year before the change and 9 accidents the following year. While this represents an apparent improvement of 10 percent, it may be inconclusive statistically since the change may be due to chance.)

(2) Are travel modes of CBD shoppers the same in one city as in another?

(3) Is the average speed of vehicles at a given location the same on Tuesday afternoon as on Friday afternoon?

In this section the elements and terminology of significance testing will first be presented. Two examples will then be given to show how the elements are combined to deal with questions like those stated above.

The essential elements of a significance test are:

(1) A *null hypothesis*—e.g., the hypothesis that a population parameter is equal to a specified value (see the examples below).

(2) An alternative hypothesis.*

(3) A significance level.

(4) A rule for "accepting" or "rejecting" a null hypothesis on the basis of a sample once the sample has been drawn. (To reject a null hypothesis is to conclude that the hypothesis is *not consistent* with the sample; to accept a null hypothesis means simply *not* to reject it.)

The conclusion reached in carrying out a significance test is subject to error. Either one of the following two types of error can occur:

(1) Rejecting the null hypothesis when in fact it is true (this is called a "Type I" error);

(2) Accepting a null hypothesis when in fact it is false (this is called a "Type II" error).†

Although it is desirable to minimize the probabilities of these errors, it is impossible to make them both arbitrarily small when the sample size is fixed. Usually the probability of a Type I error is chosen and then the probability of a Type II error is minimized. The *significance level* is equal to the probability of a Type I error (i.e., the probability of rejecting a true hypothesis). The value chosen for the significance level is naturally "small" (e.g., 0.05 [5 percent] or 0.01 [1 percent]). As the discussion proceeds in this chapter, it will become increasingly clear that the traffic engineer must exercise good judgement in specifying and testing statistical hypotheses.

Example A. Suppose that at a given intersection in a community the long-run average of traffic accidents per year has been eight. Suppose further that the intersection is redesigned in an attempt to achieve greater safety, and that in the 12 months following redesign the number of accidents is only five. The question arises as to whether greater safety has been achieved. Clearly there is the possibility that it has been; however, the fact that only five accidents occurred does not automatically prove that it has been.

*Usually in practical problems there are many alternative hypotheses (see the examples given in this chapter).

†The *power* of a significance test relative to a specified alternative hypothesis is the probability of avoiding a Type II error.

Since the long-run average is only eight, there may have been years in which there were five or even fewer accidents. In brief, there is a possibility that the safety of the intersection is the same *after* redesign as *before*. (Of course, one hopes this is not the case.) The null hypothesis to be considered is essentially the assumption that safety is the same after redesign as before. If the observed result is not consistent with this assumption, one feels justified in rejecting the assumption and regarding the safety as greater after redesign. If, however, the observed result is consistent with the assumption, then the assumption is accepted.*

Let X be the number of accidents in a year and suppose X has a Poisson distribution (see Figure 2). The *null hypothesis* states that the true mean number of accidents is 8 for the year in question. The alternatives to the null hypothesis are of the form that for the given year the true mean number of accidents is *less than* 8.† The *significance level* is the probability of rejecting the null hypothesis when it is true (Type I error). The particular *rule* that should be used for acceptance or rejection will depend on the significance level chosen; for example, with a significance level that is to be as close as possible to 0.05 (but not to exceed 0.05) the rule would be to *reject* the null hypothesis when and only when $X \leq 3$. The set of possible values of X for which the null hypothesis is rejected is called the *critical region*. In the case at hand the critical region consists of $X=0$, 1, 2, and 3 (i.e., $X \leq 3$).

The elements of the significance test are indicated in Figure 2 which gives a graph of the Poisson distribution with a true mean of 8. Values of x are shown on the horizontal scale and the relative frequencies, $f(x)$, with which they occur are shown on the vertical scale $[f(x)=\Pr(X=x)]$. The critical region is also shown in Figure 2. Since the Poisson distribution is discrete, the significance

*In a sense there is an analogy between the considerations above and those in a court trial. The intersection prior to redesign is analogous to the defendant. The point of view contained in the null hypothesis is analogous to regarding the defendant as innocent until proved guilty.

†An example of an alternative is that the true mean number is 4, say. Note that the alternatives do not include the possibility that redesign has made the intersection *more unsafe*. In other words, it is presumed that redesign at its worst leaves the intersection at least as safe as before redesign.

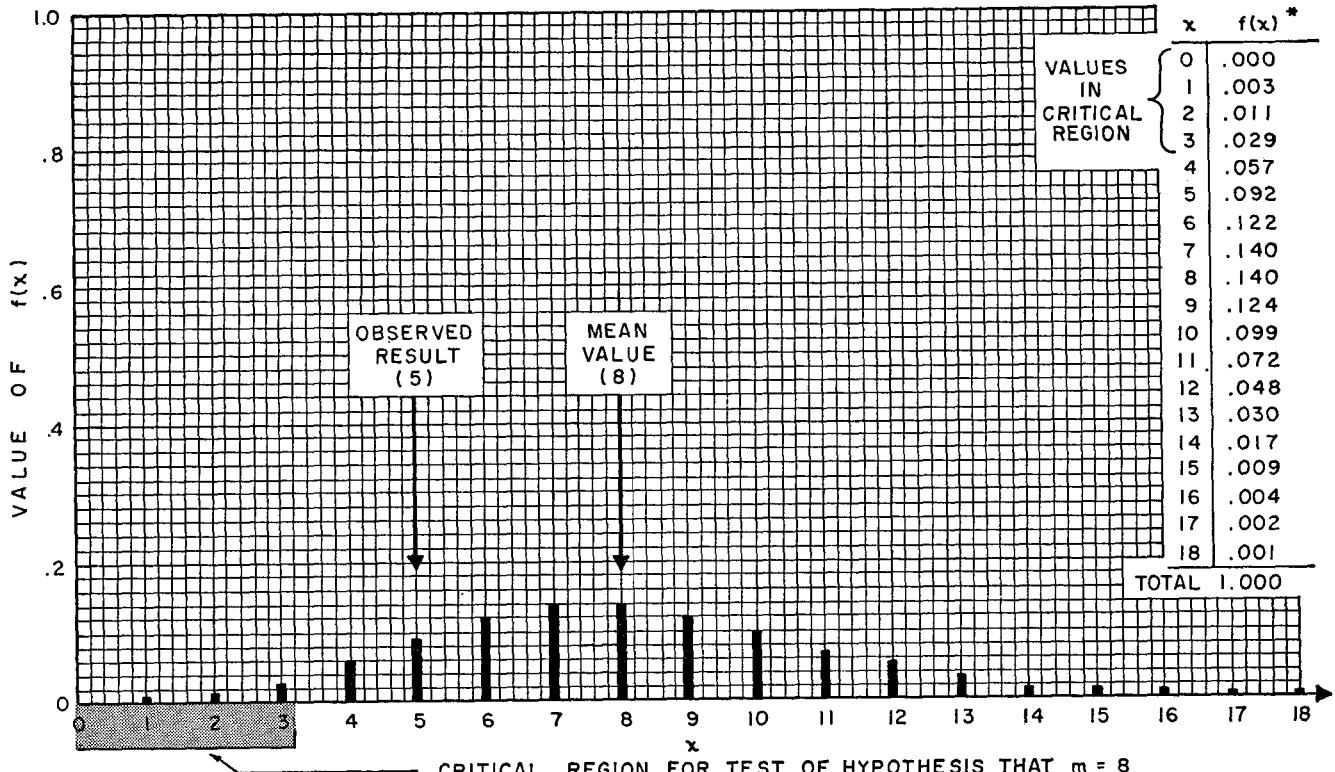


Figure 2. Line Graph of the Poisson Frequency Function.

$f(x) = e^{-m} m^x / x!$ $(m = 8)$

(See Example A—Section 4.1.)

*These values were obtained from Molina's table by rounding to three decimal places.

level is not 0.05 precisely. In fact, it is Pr $(X \leq 3) = 0.043$, which is the sum of $f(0)$, $f(1)$, $f(2)$, and $f(3)$ in the table of values of $f(x)$ given in Figure 2. The critical region cannot be enlarged to include $X = 4$ since Pr $(X \leq 4) = 0.100$, which exceeds 0.05.

Since the observed result, 5, does not lie in the critical region, the null hypothesis is accepted. More specifically, one concludes that the observed result is *consistent* (at a significance level equal to 0.05, approximately) with the hypothesis that the given year is like the past years.

Example B. The traffic signal sequence at a certain intersection was modified to eliminate a left turn phase. Following this modification it was decided to study the safety of the intersection. It was not clear whether the safety of the intersection would be affected, and so judgment on the matter was to be deferred until there had been a full year of experience with the new signal sequence. The number of accidents during the first year after the change was to be used as an indication of safety. In the past, the long-run average number of accidents at the intersection was 8 per year. An important question involved in the study was as follows: When the one-year period has been completed (and thus a one-year total of accidents after the change has become known), how should the judgment regarding safety be made?

This question will be answered by means of a significance test. The test will be carried out on the assumption that the total number of accidents during the first year after the change is 15, and that there are no marked changes in the volumes on the two intersecting streets.

As in the preceding example, it will be assumed that the number, X, of accidents in a year at the intersection follows a Poisson distribution (see Figure 3). The null hypothesis is that the true mean equals 8. The alternative hypotheses are of the form that the true mean is less than 8 and that the true mean is greater than 8.* With regard to a significance level of 0.05 (approximately) a satisfactory

*There is the possibility that the intersection is safer than it was before the change. There is also the possibility that it is less safe. Since *both* possibilities are relevant to the problem, alternatives less than 8 and alternatives greater than 8 are included.

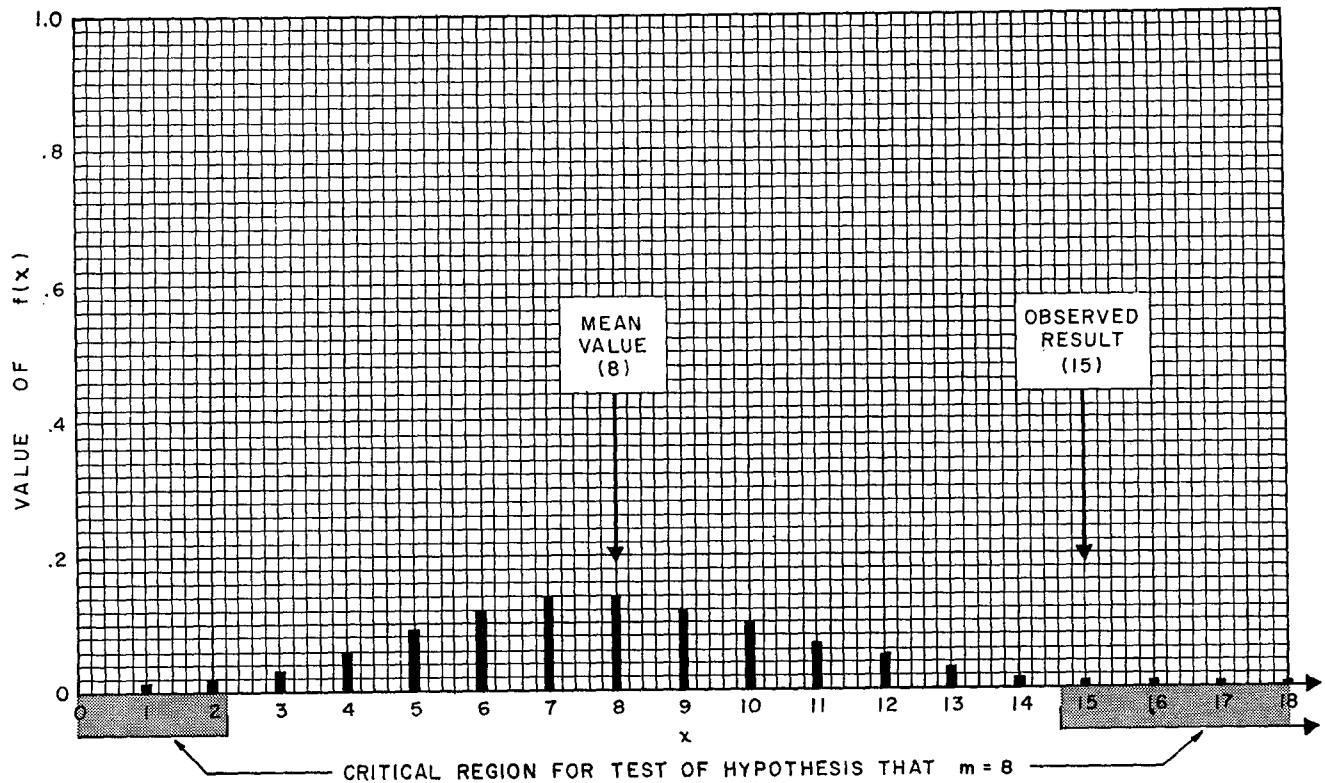


Figure 3. Line Graph of the Poisson Frequency Function.
$f(x) = e^{-m} m^x / x!$ $(m = 8)$*.
(See Example B—Section 4.1.)

*Values of $f(x)$ to three decimal places are given in Figure 2.

rule here is to reject the null hypothesis when and only when $X \geq 15$ or $X \leq 2$. This means of course that the *critical region* consists of the following values of X: 0, 1, 2, and all values that are 15 or greater.

Figure 3 gives a graph of the Poisson distribution* having a mean of 8 and shows the critical region for testing the null hypothesis. The test is an example of a so-called "two-tail" test since the critical region consists of both "tails" of the distribution. This is in contrast with the test carried out in the first example (see Figure 2) which illustrates a "one-tail" test. There the critical region consists of only one "tail" of the distribution (specifically, the lower "tail").†

To construct the critical region one finds the largest lower "tail region" whose probability content is not more than 0.05/2 and the largest "upper tail" region whose probability content is not more than 0.05/2. The complete critical region consists of both these tail regions, and the exact significance level is the sum of their probability contents. The lower tail selected has a probability content of 0.014 (instead of $0.05/2=0.025$) and the upper tail selected has a probability content of 0.016 (see the table of values of $f(x)$ in Figure 2). Accordingly, the significance level is $0.014+0.016=0.03$. If the distribution were continuous, one would have been able to select each "tail" so that its probability content was exactly 0.025. The significance level would then be $0.025+0.025=0.05$.

Since the observed result, 15, lies in the critical region, one *rejects* the null hypothesis of *no change* in safety (at a significance level equal to 0.03). In other words, the evidence suggests that the revised signal sequence may have affected the safety of the intersection.

4.2. Significance Tests Based on Confidence Intervals

In certain situations confidence intervals can be used to carry out

*The same distribution is also graphed in Figure 2.

†A "two-tail" or "one-tail" test is used according as the set of alternatives lies on both sides or only one side of the value (namely 8) in the null hypothesis. The set of alternatives lying on both sides of 8 is called "two-sided." The set lying on only one side is called "one-sided."

a test of significance (see Wilks [p. 217]). This will now be illustrated.

Example. With regard to the second example given in Section 3.3 suppose there is a hypothesis that the true mean $m=5.0$, and assume that this is to be tested at a 1 percent (0.01) significance level. The alternative hypotheses are of the form that $m<5$ and that $m>5$. The number in the sample is 100, and 5.5 is the average of the 100 observations. 99 percent confidence limits for m are 4.9 and 6.1, respectively. Since $m=5$ lies within the 99 percent confidence interval (i.e., *between* the limits) one concludes that the null hypothesis is consistent with the data at the 0.01 significance level. (Had the interval *not* included the value given in the null hypothesis, the null hypothesis would have been rejected at the 0.01 significance level.)

As indicated in the example above, the set of alternative hypotheses is "two-sided." If the set had been "one-sided," a "one-sided" confidence interval would have been used to make the test. For example, if the alternatives were of the form $m<5$, then an upper 99 percent confidence limit would be used. The null hypothesis would be accepted or rejected at the 0.01 level according as the upper limit exceeds 5 or does not exceed 5.

4.3. Contingency Tables

It is sometimes desired to test whether two binomial populations are the same. The following questions are examples of those that can be dealt with by this kind of test: With regard to reaction to lighting, do drivers going in one direction on a freeway differ from those going in the opposite direction? Do shoppers in one community differ from those in another with regard to use of bus transportation and non-bus transportation? In a given community are two outdoor advertising posters better than one for getting a certain advertising message across to the public?

The data obtained for testing whether two binomial populations are the same can be arranged conveniently in the form of a so-called 2×2 *contingency table* (see Table VII). A simple generalization of the 2×2 table can be used when more than two binomial populations are involved and when two or more multinomial populations are involved (see Section 4.3.d).

4.3.a. Two by Two Contingency Table. A 2×2 contingency table is shown in Table VII in abstract form. The table consists of two rows and two columns (apart from totals).

Table VII: 2 x 2 Contingency Table

	Number Having Attribute A	*Number Not Having Attribute A*	*Totals*
Sample 1	a	c	$a+c=m$
Sample 2	b	d	$b+d=g$
Totals	$r=a+b$	$s=c+d$	N

It is assumed that Sample 1 has been drawn from a binomial population and that Sample 2 has been drawn from a binomial population. The quantities m and g are the sizes of Samples 1 and 2, respectively. The quantity a represents the number of members of Sample 1 having Attribute A, and the quantity c represents the number of members of Sample 1 not having Attribute A. The meaning of "Attribute A" depends of course on whatever binomial populations are under study. (For example, in the shopper survey considered in Section 4.3.b. Travel by Bus represents Attribute A.) N is the grand total over both samples. r is the total number having Attribute A and s is the total number *not* having Attribute A—these totals being taken over both samples. The quantities r, s, m and g are called *marginal totals*.*

4.3.b. Testing Equality of Proportions Against "Two-Sided" Alternatives. Let p_1 and p_2 be the respective probabilities of Attribute A in the populations from which Samples 1 and 2 have been drawn (see Table VII). The null hypothesis to be tested is that $p_1 = p_2$. The alternative hypotheses of interest are all those for which $p_1 \neq p_2$. The set of alternatives is "two-sided" since it includes cases where p_1 is *less than* p_2 and cases where p_1 is *greater than* p_2.

*In the 2×2 tables considered in this book the marginal totals m and g are regarded as fixed (in advance of sampling). In the statistical literature on 2×2 tables other situations are also treated. For example, in some problems all four marginal totals are regarded as fixed; in other problems, none are fixed. In these cases the statistical methods used are computationally similar to those given herein.

(A "one-sided" set of alternatives is considered in Section 4.3.c. below.)

The following quantity can be used in testing the null hypothesis:

$$\frac{\left(a-\frac{mr}{N}\right)^2}{\frac{mr}{N}}+\frac{\left(b-\frac{gr}{N}\right)^2}{\frac{gr}{N}}+\frac{\left(c-\frac{ms}{N}\right)^2}{\frac{ms}{N}}+\frac{\left(d-\frac{gs}{N}\right)^2}{\frac{gs}{N}}=w^2, \text{ say,} \quad (4:1)$$

$(r, s, > 0)$*. When the null hypothesis is true, w^2 in (4:1) has a Chi-square distribution (approximately) with 1 degree of freedom.† This distribution is given in Appendix Table 2. At a significance level α, say, the null hypothesis is to be rejected when and only when w^2 exceeds the 100 $(1-\alpha)$ percent point of the Chi-square distribution.

It can be shown that w^2 in (4:1) can be expressed in the following simple form:

$$w^2=\frac{N(ad-bc)^2}{(mgrs)}. \quad (4:2)$$

Example. In a pilot study of drivers' reactions to lighting, 133 persons westbound on a certain segment of highway were interviewed as to their reaction to two types of lights (mercury vapor and fluorescent). 117 stated that they noticed a difference of lights whereas 16 stated that they didn't notice a difference. For 90 eastbound drivers, the corresponding numbers were 87 and 3. Are these results consistent with the assumption that the population of eastbound drivers and the population of westbound drivers are alike with regard to detection of a difference? (Are the two

*When r or s equals 0, w^2 is defined to be 0. m and g are sample sizes and are thus both greater than 0.

†The accuracy of the approximation depends on m, g, and the binomial parameter. The accuracy is good when neither m nor g is small and the binomial parameter is not close to 0; these requirements are often met in traffic engineering studies. For further discussion of the approximation, including the so-called "correction for continuity," see Duncan (pp. 506–507).

samples from the same binomial population?) The set of alternatives to the null hypothesis consists of all cases in which the two population proportions are unequal.

The data can be presented in a 2×2 contingency table as follows:

	Difference Noticed	*No Difference Noticed*	*Totals*
Eastbound	87	3	90
Westbound	117	16	133
Totals	204	19	223

From formula (4:2) one finds that

$$w^2 = \frac{(223)[(87)\,(16) - (117)\,(3)]^2}{(90)\;\;(133)\;\;(204)\;\;(19)} \doteq 5.2.$$

Under the null hypothesis the quantity w^2 has a Chi-square distribution (approximately) with 1 degree of freedom. In the above example the value obtained for w^2 exceeds the critical value, 3.84, of the Chi-square distribution at the 0.05 level of significance (see Appendix Table 2). Hence it is concluded at the 0.05 level that the data are *not* consistent with the assumption that eastbound and westbound populations are alike.

The finding of this pilot study that eastbound and westbound groups are significantly different leads one to ask, "Why should they be different?" It would appear that there are unknown factors associated with drivers' reactions to lighting. A study of such factors would appear warranted. (An interesting by-product of this pilot study is the finding that the majority of motorists in both directions of travel noticed differences between the two kinds of illumination.)

Second Example. Table VIII gives data obtained in a sampling study of travel modes of shoppers entering central business district department stores in Pawtucket and Woonsocket, Rhode Island (during a typical January 1956 day).

Are the two towns alike with regard to proportions of shoppers traveling by bus?

Table VIII: Travel Modes of CBD Store Customers—Pawtucket and Woonsocket, Rhode Island*

(Typical January 1956 Day)

Town	*Travel By Bus*		*Non-bus Travel*		*Totals*	
	No.	%	*No.*	%	*No.*	%
Pawtucket	538	38	877	62	1415	59
Woonsocket	168	17	821	83	989	41
Totals	706	29	1698	71	2404	100

The null hypothesis here is that the two towns have the same population proportions traveling by bus. The alternatives are all cases in which the two population proportions are unequal. The null hypothesis can be tested by means of w^2. Using formula (4:2) one finds that

$$w^2 = \frac{(2404)\,[(538)\,(821) - (877)\,(168)]^2}{(706)\,(1698)\,(1415)\,(989)} \doteq 124.$$

The 99 percent point of the Chi-square distribution with one degree of freedom is 6.63, approximately. Accordingly, the critical region for rejecting the null hypothesis at the 0.01 significance level consists of all values of $w^2 \geq 6.63$. Since the observed value of w^2 exceeds 6.63, the null hypothesis is rejected at the 0.01 significance level. As a matter of fact, the null hypothesis would have been rejected even if the significance level had been considerably smaller than 0.01.

The statistical analysis given above indicates that the large apparent difference between the two towns is *not* a chance difference. One could easily believe that there are important differences between them with regard to one or more of such factors as car ownership, quality of bus service, availability of parking, density of population, etc.

4.3.c. Testing Equality of Proportions Against "One-Sided" Alternatives. In the examples above, the alternative hypotheses were all cases of the form $p_1 \neq p_2$ where p_1 and p_2 are the respective

*Source: Wilbur Smith and Associates.

probabilities of Attribute A in the two populations (see Table VII). Situations arise in which all alternatives of interest are "one-sided" —for example, that p_1, say, is less than p_2. To test the null hypothesis (that $p_1 = p_2$) against "one-sided" alternatives, one can use:

$$\frac{\frac{a}{m} - \frac{b}{g}}{\sqrt{\left[\left(\frac{r}{N}\right)\left(\frac{s}{N}\right)\left(\frac{1}{m}+\frac{1}{g}\right)\right]}} = w', \text{ say, } (r, s, > 0)^* \qquad (4:3)$$

(see Table VII). The quantity, w', above is the difference between the sample proportions divided by the estimate of the standard deviation of their difference. Under the null hypothesis w' is approximately a standard normal variable (for sufficiently large m and g). To test the null hypothesis against alternatives of the form $p_1 < p_2$, one would reject when and only when the observed value of w' has a sufficiently large negative value.

It should be remarked that $(w')^2$ equals the quantity w^2 given in (4:1) and (4:2); thus w' is also suitable for carrying out the test required in the examples in Section 4.3.b. (The null hypothesis would be rejected when the observed value of w' has a sufficiently large positive value and when it has a sufficiently large negative value.)

Example. Suppose that in an outdoor advertising study in a certain town it is desired to compare the effectiveness of a single poster with the effectiveness of a pair of widely separated posters. After a prominent poster has been displayed for a month, a sample of 200 residents is obtained of whom 160 recognize a photograph of the poster. Suppose that subsequently two prominent, *identical* posters are displayed—one in the same place as the original poster and the other in a different part of town. After the two posters have been displayed for a month, a sample of 250 residents is obtained of whom 225 recognize a photograph of the poster. Are these results consistent with the hypothesis that two posters are no better than one? The set of alternatives consists of all cases for which $p_1 < p_2$, where p_1 represents the population proportion of

*When r or s equals 0, w' is defined to be 0. m and g are sample sizes and are thus both greater than 0.

residents who would recognize the poster in a single location and p_2 represents the population proportion of residents who would recognize the poster that is in two locations.

To test the null hypothesis one can use formula (4:3):

$$w' = \frac{\frac{160}{200} - \frac{225}{250}}{\sqrt{\left[\left(\frac{385}{450}\right)\left(\frac{65}{450}\right)\left(\frac{1}{200}+\frac{1}{250}\right)\right]}} \doteq -3.0.$$

Since the observed value (-3.0) is less than the 5 percent point (-1.645) of the standard normal distribution, the null hypothesis is rejected at the 0.05 level of significance.* In other words, it would appear that the pair of identical posters is better than one.

4.3.d. The h × k Contingency Table. The 2×2 contingency table in Section 4.3.a. is a special case of an "$h \times k$" contingency table given in abstract form in Table IX. With regard to the $h \times k$ table the following assumptions are generalizations of the assumptions in Section 4.3.a. For each i ($i=1, \ldots, h$) there is a multinomial population involving k categories. The probability that an individual drawn at random from the ith population belongs to category j will be denoted by p_{ij} ($i=1, \ldots, h; j=1, \ldots, k$). The null hypothesis is that for every j:

$$p_{1j}=p_{2j}=\ldots=p_{hj} \quad (j=1, \ldots, k). \tag{4:4}$$

The set of alternative hypotheses consists of all cases in which two or more of the populations are not the same.

A test of the null hypothesis can be carried out by means of the quantity W^2 defined as follows:

$$W^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i\cdot}\, n_{\cdot j}}{N}\right)^2}{\frac{n_{i\cdot}\, n_{\cdot j}}{N}} = N\left(\sum_{i,j} \frac{n_{ij}^2}{n_{i\cdot}\, n_{\cdot j}} - 1\right). \tag{4:5}$$

When $h=k=2$, W^2 in (4:5) equals w^2 in (4:1). When the null hypothesis is true, W^2 has a Chi-square distribution approximately

*For a "two-tail" test (at the 0.05 level) the null hypothesis would be rejected if $w' < -1.96$ or $w' > 1.96$. Since the observed w' equals -3.0, the null hypothesis would also have been rejected if a "two-tail" test had been carried out.

with $(h-1)$ $(k-1)$ degrees of freedom. A sufficiently large value of W^2 leads to rejection of the null hypothesis. (For further discussion of contingency tables see Duncan, Hoel, or Mood.)

Table IX: h x k Contingency Table

		Categories							
Samples	1	2	. . .	j	. . .	k	Totals
1	n_{11}	n_{12}	. . .	n_{1j}	. . .	n_{1k}	$n_{1\cdot}$
2	n_{21}	n_{22}	. . .	n_{2j}	. . .	n_{2k}	$n_{2\cdot}$
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
i	n_{i1}	n_{i2}	. . .	n_{ij}	. . .	n_{ik}	$n_{i\cdot}$
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
h	n_{h1}	n_{h2}	. . .	n_{hj}	. . .	n_{hk}	$n_{h\cdot}$
Totals	$n_{\cdot 1}$	$n_{\cdot 2}$	. . .	$n_{\cdot j}$	. . .	$n_{\cdot k}$	N

(Note: n_{ij} is the number of members of the ith sample that fall in category j).

4.4. Significance Tests Regarding Population Means

The testing of hypotheses regarding population means is treated at length in the statistical literature. Some special cases of such hypotheses are discussed below in relation to traffic studies.

4.4.a. Testing Whether a Population Mean Has a Given Value. Situations often arise in which there is a null hypothesis that a population mean has a given value. For example, in the case of a binomial population, suppose there is a hypothesis that the probability, p, of Heads in tossing a certain coin equals 1/2. (Note that p is the *true mean* relative frequency of Heads in the population.) After observing a number of tosses of the coin one could test the hypothesis. The test could be carried out against "two-sided" alternatives by setting up a confidence interval for p and noting whether the value of 1/2 lies in the interval (see Sections 3.1 and 4.2). The hypothesis is accepted or rejected according as 1/2 lies inside or outside the confidence interval.

With respect to the Poisson distribution, suppose there is a null hypothesis that the mean, say m, equals a given value m_0. The example in Section 4.2 shows, in detail, how to test the null hypothesis by means of a confidence interval; in this example $m_0 = 5$.

Confidence intervals can also be used to test whether the mean of a normal population equals a specified value. For example, when the value of σ is unknown, formula (3:6) could be used. To test the null hypothesis against, say "two-sided" alternatives, one accepts or rejects the hypothesis according as the specified value does or does not lie within the confidence interval.

When the appropriate conditions hold, one can use a confidence interval of the type described in Section 3.5 to test the hypothesis that a population mean has a specified value.

4.4.b. Testing Whether the Means of Two Normal Populations are Equal. Tests of whether the means of two binomial populations are equal are described in Sections 4.3.b. and 4.3.c. A test of whether the means of two normal populations are equal will now be considered. It will be assumed that the *variances* of the two *populations* are equal.*

Let $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ be two samples from normal populations having means u_1 and u_2, respectively, and a common variance, σ^2 (whose value is unknown). Let the null hypothesis be that $u_1 = u_2$, and suppose that the alternatives of interest are that $u_1 < u_2$ and that $u_1 > u_2$. When the null hypothesis is true, the following quantity has Student's t-distribution† with $n_1 + n_2 - 2$ degrees of freedom:

$$\frac{(\bar{x} - \bar{y}) \sqrt{\left(\frac{n_1 n_2}{n_1 + n_2}\right)}}{\sqrt{\left(\frac{n_1 S_x^2 + n_2 S_y^2}{n_1 + n_2 - 2}\right)}} = t, \text{ say,} \qquad (4:6)$$

*From a theoretical standpoint this assumption is important. When the ratio of the variances is unknown, the problem is referred to as the Behrens-Fisher problem, which has been a controversial topic in the field of statistics. (For further discussion of the problem see Cramér (pp. 520–523) or Mood (pp. 264–265)). For practical purposes the Aspin-Welch test can be used when the ratio of variances is unknown (see Duncan (pp. 476–478)).

†Appendix Table 3 is a table of Student's t-distribution.

where S_x^2 and S_y^2 are the sample variances. The null hypothesis is rejected when t has a sufficiently large positive value and when it has a sufficiently large negative value.

Example. A spot speed study was made of a given location on a Tuesday and Friday of a typical week. The purpose was to obtain information as to whether Tuesday and Friday mean speeds were the same. A sample of 16 speeds obtained on Tuesday had a mean of 42.0 (mph) and a variance of 30.1. A sample of 26 speeds obtained on Friday had a mean of 50.1 (mph) and a variance of 28.7. All observations were made during off-peak periods, and the weather was clear on both days. It is assumed that the samples come from normal populations with the same variance.* Test the hypothesis that the population means are equal at, say, the 0.05 level of significance.

From Appendix Table 3 one finds that for 40† degrees of freedom the 97.5 percent point of the t-distribution is 2.021; thus the null hypothesis will be accepted or rejected depending on whether t in (4:6) does or does not satisfy the following inequality‡:

$$-2.021 < t < 2.021.$$

Substituting the numerical values of the sample means and variances in (4:6) one finds that

$$t = \frac{(-8.1)\sqrt{\left(\frac{(16)\ (26)}{42}\right)}}{\sqrt{\left(\frac{481.6+746.2}{40}\right)}}$$

$$\doteq \frac{-(8.1)\sqrt{(9.90)}}{\sqrt{(30.7)}} \doteq -4.6.$$

Since the value of t does not lie in the interval $-2.021 < t < 2.021$, the null hypothesis is rejected at the 0.05 level of significance.

*A test for equality of the variances is carried out in the first example in Section 4.5. The hypothesis of equality is accepted there at the 0.05 level of significance.

†Note that $n_1 + n_2 - 2 = 16 + 26 - 2 = 40$.

‡The frequency function of t is symmetrical about 0; thus the amount of probability *below* -2.021 and *above* $+2.021$ equals $.025 + .025 = .05$, which is the required significance level.

Since the mean speeds are significantly different, the question arises as to what might account for the difference. Presumably additional field investigation would be required to settle this question.

If it were desired to determine whether curb parking regulations affect mean peak hour speeds, the procedure would be generally the same. Spot speed studies would be conducted *before* and *after* implementation of the regulations, and a significance test for the difference between the two means would be carried out. (See Section 5.5.)

Second Example. The data given below regarding reaction times were obtained for vehicles on Orange Street, New Haven, Connecticut.* It is desired to test whether the population means are equal (at an 0.05 level of significance). The populations are assumed to be normal with the same variance.†

Northbound	*Southbound*
$\bar{x}=1.68$ seconds	$\bar{y}=1.86$ seconds
$S_x=1.45$ seconds	$S_y=1.75$ seconds
$n_1=28$	$n_2=29$

$$\bar{x}-\bar{y}=1.68-1.86=-0.18$$

$$n_1+n_2-2=28+29-2=55.$$

Substituting in formula (4:6) one finds that

$$t=\frac{-(0.18)\sqrt{\left(\frac{(28)\ (29)}{57}\right)}}{\sqrt{\left(\frac{28(1.45)^2+29(1.75)^2}{55}\right)}}\doteq -0.41.$$

The 97.5 percent point of the t-distribution for 55 degrees of freedom is approximately 2.0 (see Appendix Table 3). Since the observed value of t lies between -2.0 and 2.0, the null hypothesis is accepted at the 0.05 level. (Another way of stating this conclu-

*See Greenshields, Ericksen, and Schapiro.

†As indicated in the second example in Section 4.5, a test for equality of the variances resulted in acceptance of the hypothesis of equality (at the .05 level).

sion is to say that the difference between the two means is *not* significant at the 0.05 level.)

4.5. A Test for Equality of Variances (The F-test)

The question of whether the variances of two normal populations are equal is of interest in various traffic engineering problems. Illustrations of such problems are indicated by the two examples given at the end of this section. Speed zoning provides another example; the zoning may have little or no effect on the mean speed, but it may influence the "spread" of the distribution of speeds.* For example, when minimum and maximum speed limits are put into effect on freeways, the engineer may wish to ascertain whether there is a change in the speed "spread."

The hypothesis that the variances of two normal populations are equal can be tested by means of the sample variances. Let n_1 and n_2 be the sizes of the samples and let S_x^2 and S_y^2 be the respective sample variances. Let F denote the ratio

$$\frac{\dfrac{n_1 S_x^2}{n_1 - 1}}{\dfrac{n_2 S_y^2}{n_2 - 1}}. \tag{4:7}$$

When the null hypothesis is true, this ratio has the *F-distribution* with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. A more detailed discussion of this well-known distribution is given in Hoel.

Let σ_x^2 and σ_y^2 be the true variances of which S_x^2 and S_y^2 are point estimates, respectively. To test the null hypothesis against all alternatives of the form $\sigma_x^2 < \sigma_y^2$ and $\sigma_x^2 > \sigma_y^2$, one rejects the null hypothesis when F is sufficiently large and when it is sufficiently small. In making use of tables of the F distribution one forms the ratio in (4:7) so that the observed F is not less than 1. The null hypothesis is then to be rejected at the significance level α when and only when F exceeds the $100(1-\alpha/2)$ percent point of the F-distribution.

*The "spread" is indicated by the standard deviation, which is simply the square root of the variance.

Example. With regard to the first example in Section 4.4 test for equality of variances (at the 0.05 level of significance). It is given that

$$n_1 = 16;\ S_x^2 = 30.1;\ n_2 = 26;\ S_y^2 = 28.7.$$

Forming the F-ratio so that $F \geq 1$, one obtains

$$F = \frac{\frac{16}{15}(30.1)}{\frac{26}{25}(28.7)} \doteq 1.08.$$

For 15 and 25 degrees of freedom one finds from a table of F that 2.41 is the 97.5 percent point of the F-distribution.* Since the observed value, 1.08, is less than 2.41, one accepts the null hypothesis at the 0.05 level of significance.

Second Example. With regard to the second example in Section 4.4 test for equality of variances (at the 0.05 level). ($S_x = 1.45$, $n_1 = 28$; $S_y = 1.75$, $n_2 = 29$.) Forming the F-ratio so that $F \geq 1$, one finds that the ratio equals 1.45 approximately. For 28 and 27 degrees of freedom the 97.5 percent point of the F-distribution is 2.15. Since 1.45 is less than 2.15, one accepts the hypothesis of equal variances (at the 0.05 level).

*The statistical literature contains many tables of the F-distribution. An extensive table of it is given in Hald's *Statistical Tables and Formulas*. He refers to it as the v^2 distribution.

Chapter 5: Case Studies and Applications

Applications of sampling concepts and methods have been illustrated in Chapters 1, 2, 3, and 4 by means of brief examples pertaining to traffic engineering. The purpose of the present chapter is to give additional and more detailed traffic applications. The following topics are considered: (1) Sample Size and Survey Design, (2) Techniques of Sampling, (3) Absolute and Relative Error in Estimating The Binomial Parameter (for example, in estimating the proportion of local cars in traffic), (4) Determining Sample Size for Estimating The Mean of a Population, (5) "Before-and-After" Studies, (6) Randomness of Traffic, and (7) Estimation of Traffic Volume by Means of Short Counts.

5.1. Sample Size and Survey Design

A survey based on sampling (instead of complete enumeration) is called a *sample survey*.* This type of survey is used frequently in traffic engineering studies. Proper design and execution of such a survey call for a thorough knowledge of both sampling and the subject matter involved in the survey.

An important factor in the design of a sample survey is the size of the sample involved. Fortunately, this factor is often under the control of the investigator. In advance of sampling, information about the population is incomplete. (In fact, if it were complete, there would be no point in drawing a sample.) As the sample size increases, information about the population becomes less incomplete. Unfortunately, however, there is usually an increase in sampling cost as the sample size increases. In practice there is a need to achieve a balance between the *cost* and *incompleteness* of the information obtained in sampling. The balance depends on the resources and requirements of the investigator.

Further general remarks are given below regarding the relation†

*The principal steps of a sample survey are described by Cochran (pp. 2–4).

†Specific aspects of this relation are given in other parts of this book (e.g., see Table 1 and Sections 5.3, 5.4, and 5.7).

between sample size and the amount of information in the sample. The relation between cost and size of a sample is beyond the scope of this book since it depends on the particular circumstances in which a sample survey is carried out.

5.1.a. Effect of Sample Size on Estimation. Almost all point estimators have the property that the "spread" (in some sense) of the estimator's distribution tends to 0 as the sample size becomes indefinitely large. An illustration of this property is provided by the distribution of the mean of a sample from an infinite population. The sample mean, say $\bar{x}$, is a point estimator of the population mean. The "spread" of the distribution of $\bar{x}$ can be regarded as the standard deviation, $\sigma_{\bar{x}}$, of $\bar{x}$*; and $\sigma_{\bar{x}}=\sigma/\sqrt{n}$, where n is the sample size and σ is the population standard deviation. It is clear that $\sigma_{\bar{x}}$ tends to 0 as n becomes indefinitely large.

In general the *average* length of a confidence interval (for a parameter of a distribution) depends on the sample size. As the sample size becomes indefinitely large, the average length of the interval tends to 0. Various formulas in Chapter 3 indicate how confidence limits depend on sample size (e.g., see (3:3), (3:5), and (3:6)); similarly, Figure 1 (Chapter 3) provides a clear, comprehensive idea of this dependence with regard to confidence limits for the binomial parameter.

5.1.b. Sample "Percent." In many traffic engineering studies a quantity called "percent" is associated with the sample drawn. A "20 percent" sample, say, is one that contains 20 percent of the population involved. A "5 percent" sample is one that contains 5 percent of the population. This quantity has the appearance of being an index of the adequacy of the sample. Actually, however, the appearance is deceptive since the quantity is not sufficient by itself for that purpose. The variance of a sample mean can be used to show why this is so. (Usually, the smaller this variance is, the more adequate the sample is for estimating the population mean.) The formula for this variance is $(\sigma^2/n)\ (N-n)\ /\ (N-1)$, where σ^2 is the population variance, N is the size of the (finite) population, and n is the sample size. The formula can also be written as $(\sigma^2/n)\ (1-r)\ (N)\ /\ (N-1)$, where $r=n/N$. (The sample "percent" equals

*$\sigma_{\bar{x}}$ is also called the *standard error* of the mean.

$100r$.) For fixed N the variance decreases as r increases; however, for comparing samples from populations of different sizes r is not satisfactory by itself. A 20 percent sample from a population of size 1,000 would have a *larger* variance than a 5 percent sample from a population of size 100,000 (when σ^2 is the same in both cases). It should be noted also that when the sampling is from an infinite population, the sample "percent" is irrelevant to the adequacy of the sample. In this case $N=\infty$; thus the sample "percent" equals 0 irrespective of the sample size.

5.1.c. Finite Population Adjustment of Sample Size. In designing a sample survey one may wish to choose the sample size so that the variance, $\sigma_{\bar{x}}^2$, of the sample mean satisfies some requirement. For example, the requirement might be that $\sigma_{\bar{x}}^2$ be a certain fraction of the population variance, σ^2. As previously indicated, $\sigma_{\bar{x}}^2=\sigma^2/n$ when the population is infinite; from this simple formula one can easily compute the required value, say n', of n. When the sampling is without replacement from a finite population (with the same variance σ^2), the required sample size, say n'', is smaller than n'.* The relation between n'' and n' is as follows:

$$n'' = \frac{n'}{1+\dfrac{n'-1}{N}} \doteq \frac{n'}{1+\dfrac{n'}{N}}, \qquad (5:1)$$

where N is the size of the finite population. The quantity $n'-n''$ can be looked upon as an increase (in the required sample size) that results from regarding a population as infinite when it is actually finite. Formula (5:1) has of course been obtained by solving the equation

$$\frac{\sigma^2}{n'} = \left(\frac{\sigma^2}{n''}\right)\left(\frac{N-n''}{N-1}\right) \qquad (5:2)$$

for n''.

5.2. Techniques of Sampling

A sampling procedure that is based on probability theory is referred to as *probability sampling* (see Cochran, pp. 6–7). In practice the principal types of probability sampling are:

*The reason is that $\sigma_{\bar{x}}^2$ equals $(\sigma^2/n)\,(N-n)\,/\,(N-1)$ instead of (σ^2/n).

1. random sampling,
2. cluster sampling,
3. stratified sampling.

Combinations of these types are used frequently. Covault, for example, has carried out an interesting comparison of these probability sampling procedures with regard to estimation of highway needs.

In this book most of the sampling considered is random sampling. Systematic sampling (a special case of cluster sampling) and stratified sampling are discussed in Section 5.7 which deals with short-count estimation. The books by Cochran, Deming, McCarthy, and Sukhatme will be of interest to those wishing to read further about sampling theory and methods.

The sampling techniques described in this section are suitable for sampling from *synthetic* as well as actual populations. Sampling from a synthetic population is of interest in traffic engineering in connection with *simulation*—e.g., the simulation of traffic flow characteristics. A further matter of interest regarding these sampling techniques is that they can be "computerized" easily. This is important since the use of high-speed, electronic computers will often greatly increase the value of a simulation study.

Detailed procedures of drawing a random sample are given below. The procedures are based on the use of random digits, which will now be described.

5.2.a. Random Digits and Random Selection. A random digit is an observed value of a random variable, X, having the following frequency function:

$$\Pr(X=i) = 1/10 \quad (i=0, 1, \ldots, 9). \tag{5:3}$$

(For a definition of the *frequency function* of a random variable see Section 3 of the Appendix.) A table of random digits is simply a set of random digits selected independently; Appendix Table 4 is an example of such a table. Two random digits can be regarded as forming a number from 00 to 99 selected purely at random. Three random digits can be regarded as a number from 000 to 999 selected purely at random. A similar statement holds for any number of random digits. For example, the number 4926 formed by the first four digits in the first row of Appendix Table 4 can be regarded as

a number selected purely at random from the 10,000 numbers ranging from 0000 to 9999. A number formed in this way is called a *random number.*

5.2.a.1. *Random Selection of an Element from a Finite Population.* By means of random digits one can easily select an element at random from a finite population. First, the elements are numbered from 0 to N-1, where N is the total number of elements in the population. One then picks a random number and selects the element that has this number. If the random number exceeds N-1, it is discarded and another one is picked. The example below illustrates random selection of an element from a population in which the total number of elements is twelve.

Example. Select a five-minute short-count period at random from the twelve five-minute short count periods in an hour. (The need for such a selection could arise in making the "random start" used in the systematic sampling procedure described in 5.7.c.) Let the twelve periods be numbered 00, 01, 02, . . ., 10, 11, where 00 is associated with the first period, 01 with the second period, etc. Since the largest of the twelve numbers consists of two digits, a two-digit random number will be used. To pick a two-digit random number use the first pair of digits in, say, the fifth row of Appendix Table 4. Turning to the table, one finds that the random number is 41. Since this number exceeds the largest of the twelve numbers, it must be discarded. To pick another random number use the second pair of digits (in the fifth row). This number is 01; thus the period selected is the one associated with the number 01—namely the second period. In other words, the five-minute short-count period selected at random is the one extending from five minutes to ten minutes after the beginning of the hour.

5.2.a.2. *Random Selection of a Value of a Random Variable.* The random selection of a value of a random variable can be carried out by means of random digits. A detailed procedure for making such a selection is described below.

Let $X_1, X_2, \ldots, X_h$ be random digits, and let

$$\frac{X_1}{10}+\frac{X_2}{10^2}+\ldots+\frac{X_h}{10^h}= Y, \text{ say.} \qquad (5:4)$$

Since each random digit $X_1, \ldots, X_h$ is one of the numbers 0, 1, ..., 9, the possible values of Y lie between 0 and 1. It can be shown that the distribution of Y is approximately a *uniform distribution* (described at the end of Section 4 of the Appendix).

To illustrate the way in which Y is formed, let h be 5, say, and let X_1, X_2, X_3, X_4, X_5 be the last group of five digits in, say, the first row of Appendix Table 4. Turning to Appendix Table 4 one finds that $X_1=1$, $X_2=2$, $X_3=5$, $X_4=7$, and $X_5=4$. It follows that Y then equals $1/10+2/10^2+5/10^3+7/10^4+4/10^5$ which equals 0.12574. The number 0.12574 can be regarded as a value selected at random from a distribution that is approximately uniform. The quantity $1/10^h$ is a bound on the difference between the cumulative distribution function of Y and the cumulative uniform distribution. For example, when h equals 5 the error of the approximation is not more than 1/100,000. The approximation is improved by replacing Y by Y', where $Y'=Y+5/10^{h+1}$. The bound is then reduced to $5/10^{h+1}$.

Let X be a random variable and represent the cumulative distribution function of X by $F(x)$. Let z' be a value selected at random from a uniform distribution, and let x' be the least value of x such that

$$F(x) \geq z'. \tag{5:5}$$

The number x' is a randomly selected value of X. When $F(x)$ is a continuous and increasing function, x' is simply the value of x satisfying the equation

$$F(x) = z'. \tag{5:6}$$

It will be evident from the discussion below that formulas (5:5) and (5:6) are very useful. An illustration of the use of (5:6) will now be given for the special case in which X has a standard normal distribution.

Example. Select a value at random from the standard normal distribution. This will be done in two steps. The first step is to select a value, z', at random from the uniform distribution. The second step is to solve (5:6) for x', where $F(x)$ is the cumulative standard normal distribution. Choosing $h=3$, say, select a value of z' by means of the first three digits in, say, the second row of Appendix Table 4. These digits are 3, 3, 8, respectively; thus $z'=0.338$. Since $F(x)$ is the cumulative standard normal distribution,

Appendix Table 1 can be used to solve (5:6). Setting $F(x)$ equal to 0.338, one finds from Table 1 that x' equals -0.42 approximately.* The number -0.42 can be regarded as a value selected at random from the standard normal distribution. The final result is approximate (rather than exact) for two reasons. One is that in the first step the distribution used is an approximation to the uniform distribution. The second reason is that the solution of (5:6) is approximate. The error in the first step can be reduced by simply using a larger value of h. The error in the second step can be reduced by using a table that gives x to a larger number of decimal places.

A value of the standard normal distribution can also be selected at random by means of a table of random normal numbers (see 5.2.b.2. below).

5.2.b. Drawing a Random Sample. A random sample of values of a random variable X can be regarded as observed values of X that have been obtained independently. Accordingly, to draw a random sample of values of X one can simply carry out independent repetitions of the random selection of a value of X. In Sections 5.2.b.1. and 5.2.b.2. the drawing of a random sample is described with regard to three different distributions of X—namely the binomial, Poisson, and normal distributions.

When a sample is drawn from a finite population *without replacement*, it is called a random sample if the procedure is such that all possible samples have the same chance of being selected. This type of sampling is described in Section 5.2.b.3.

It should be noted that the sample elements are not obtained independently in random sampling without replacement from a finite population. This type of sampling is therefore not precisely the same as random sampling in which the elements are obtained independently. This latter type of sampling can be regarded as an extension of random sampling without replacement. The extension involves a process in which the size of a finite population increases without limit.

5.2.b.1. *Sampling From Binomial and Poisson Populations.* Let X have a binomial distribution with parameter p. This means that

*Since 0.338 is less than 0.5, one uses the fact that $F(-x) = 1 - F(x)$ in solving the equation. ($1 - 0.338 = 0.662$; thus $-x' \doteq 0.42$ and $x' \doteq -0.42$.)

Pr $(X=1)=p$ and Pr $(X=0)=1-p$, where the values 1 and 0 of X mean "success" and "failure," respectively. To select a value of X at random, choose an h-digit number, say w', from Appendix Table 4 and then assign the value 0 or 1 to X according as $(w'/10^h)$ is less than $1-p$ or not less than $1-p$. The procedure is exact if h is chosen to be equal to the number of decimal places to which p is expressed. The procedure is based almost entirely on the use of (5:5).

Example. Select values at random from a binomial population in which the probability, p, of a "success" is 0.70. Since p is expressed to two decimal places, choose $h=2$. Choosing the first two digits of, say, the third row of Appendix Table 4, one finds that $w'=23$. Since $w'/10^2$ $(=0.23)$ is less than $1-p$ $(=0.30)$, one assigns the value 0 to X as the outcome of this random selection. A second random selection of a value of X can be made by means of the next two digits in the third row of Appendix Table 4. The number formed by these two digits is 31. Since $31/10^2$ $(=0.31)$ is not less than 0.30, one assigns the value 1 to X as the outcome of the second random selection. For a third random selection one can use the third set of two digits in the same row. The number formed by those two digits is 87, thus one assigns the value 1 to X as the outcome of the third random selection. It should be noted that a random sample of three values of X has been obtained. The elements of the sample are 0, 1, and 1; thus the sample consists of one "failure" and two "successes."

Second Example. Suppose that it is desired to simulate the occurrence (and non-occurrence) of left-turning vehicles on one approach to a given intersection. Suppose further that from a general knowledge of traffic on this approach, one can assume that the probability an arriving vehicle turns left is 0.12. To determine whether a given (simulated) vehicle turns left, one can use a two-digit random number. The vehicle is regarded as one that turns left or does not do so according as the random number is greater than (or equal to) 0.88 or is less than 0.88. (Note that $1-0.12=0.88$.) Starting with, say, the 15th row of Appendix Table 4 one obtains the results tabulated below for the first 20 (simulated) vehicles. (In the table, L means "turns left" and T means "does not turn left.")

Vehicle	1	2	3	4	5	6	7	8	9	10
Random Number	92	47	01	88	40	76	01	19	31	09
Action of Vehicle	L	T	T	L	T	T	T	T	T	T
Vehicle	11	12	13	14	15	16	17	18	19	20
Random Number	14	34	45	56	14	50	28	41	58	65
Action of Vehicle	T	T	T	T	T	T	T	T	T	T

Random selection of a value from a Poisson population will now be considered. The procedure, which is based mainly on formula (5:5), will be described by means of an example. It could be used, for instance, in simulating the arrival of cars at a drive-in bank or at a garage.

Example. Select values at random from a Poisson distribution having a mean of 8. Let X be the random variable having this distribution, and let $f(x)$ be the frequency function of X. This frequency function and the cumulative distribution function, say $F(x)$, are tabulated below. ($f(x)$ is also given in Figure 2 (to three decimal places).)

x	$f(x)$	$F(x)$
0	0.000	0.000
1	0.003	0.003
2	0.011	0.014
3	0.029	0.043
4	0.057	0.100
5	0.092	0.192
6	0.122	0.314
7	0.140	0.454
8	0.140	0.594
9	0.124	0.718
10	0.099	0.817
11	0.072	0.889
12	0.048	0.937
13	0.030	0.967
14	0.017	0.984
15	0.009	0.993
16	0.004	0.997
17	0.002	0.999
18	0.001	1.000

Since $F(x)$ is expressed to three decimal places, a three-digit random number from Appendix Table 4 can be used to randomly select a value of X. Let s represent a three-digit number selected from Appendix Table 4, and let s' be $s/10^3$ $(=s/1000)$. The possible values of s' are 0.000, 0.001, . . ., 0.999. If s' turns out to be less than 0.003, the value 1 is assigned to X; if s' is less than 0.014 but not less than 0.003, the value 2 is assigned to X; etc. The complete correspondence between values of X and ranges of values of s' is tabulated below.

Range of values of s'	*Value of* X
$0 \leq s' < 0.003$	1
$0.003 \leq s' < 0.014$	2
$0.014 \leq s' < 0.043$	3
$0.043 \leq s' < 0.100$	4
$0.100 \leq s' < 0.192$	5
$0.192 \leq s' < 0.314$	6
$0.314 \leq s' < 0.454$	7
$0.454 \leq s' < 0.594$	8
$0.594 \leq s' < 0.718$	9
$0.718 \leq s' < 0.817$	10
$0.817 \leq s' < 0.889$	11
$0.889 \leq s' < 0.937$	12
$0.937 \leq s' < 0.967$	13
$0.967 \leq s' < 0.984$	14
$0.984 \leq s' < 0.993$	15
$0.993 \leq s' < 0.997$	16
$0.997 \leq s' < 0.999$	17
$s' = 0.999$	18

It should be noted that the endpoints of the ranges of values of s' are values of $F(x)$. The results of the procedure involved are exact for the distribution, $F(x)$, tabled above.

The actual selection of values of X will now be illustrated. Let s be the first three digits in, say, the fourth row of Appendix Table 4. Turning to this table, one finds that these digits are 5, 7, 5; thus $s = 575$ and $s' = 0.575$. Since 0.575 lies in the range $0.454 \leq s' < 0.594$,

one assigns the value 8 to X. Using the next three digits in the fourth row, one can select a second value of X at random. These three digits are 1, 7, 5; thus $s=175$ and $s'=0.175$. Since 0.175 lies in the range $0.100 \leq s' < 0.192$, one assigns the value 5 to X. Continuing in this way, one can select additional values of X at random. For example, the values of s' associated with the third, fourth, and fifth sets of three digits in that row are 0.525, 0.650, and 0.281, respectively. The corresponding values of X are 8, 9, and 6, respectively. A random sample of five values of X has now been obtained. The elements of the sample (in the order in which they were drawn) are: 8, 5, 8, 9, and 6.

The binomial and Poisson distributions are examples of distributions associated with discrete random variables. The procedure for randomly selecting a value of any given discrete random variable is similar to the procedures described above.

5.2.b.2. *Sampling From a Normal Population.* To select a value at random from a standard normal distribution, one can use a table of random normal numbers. An example of such a table is given in *A Million Random Digits* (by The Rand Corporation). To obtain a random sample of n values from the distribution, one simply reads out n entries from the table. The values that make up such a table are obtained in much the same manner as that described in the example in 5.2.a.2.

The mean and variance of the standard normal distribution are 0 and 1, respectively. A randomly selected value from this distribution can be transformed easily to a randomly selected value from a normal distribution with mean u and variance σ^2. Let x' be a randomly selected value from the standard normal. The quantity $\sigma x' + u = y'$, say, is then a randomly selected value from the normal distribution whose mean and variance are u and σ^2. For example, suppose one wishes to obtain a value at random from a normal distribution whose mean and variance are 7 and 25, respectively. Using -0.42 as the value of x' (see the example in 5.2.a.2.), one can regard 4.90 as a randomly selected value from the distribution. (Note that $y' = \sigma x' + u = 5\ (-0.42) + 7 = -2.10 + 7 = 4.90$.)

5.2.b.3. *Sampling Without Replacement From a Finite Population.* A table of random numbers can be used to draw without replacement

a random sample of size n ($n > 1$)* from a finite population of size N. The procedure is to number the N objects in the population from 0 to $N-1$ and then pick random numbers until n are obtained. The actual sample consists of the objects whose numbers were drawn. In selecting random numbers discard any that have already been selected and discard any that exceed $N-1$.

Example. From the 95 residences in a certain survey area draw a random sample of size 20 without replacement. (The need for such a sample might arise in making an origin-destination study.) Set up a convenient numbering of the residences from 00 to 94. Turning to Appendix Table 4, proceed along, say the sixth row, using pairs of digits to select members of the sample. One finds that the first 20 pairs of digits are as follows:

73, 88, 98, 02, 36, 99, 53, 12, 30, 53,

71, 23, 74, 88, 61, 59, 04, 67, 62, 83.

This set of 20 numbers is not entirely satisfactory since there are two 88's, two 53's and two numbers (namely 98 and 99) that exceed the maximum number (94) among the population elements. The second 88, the second 53, and the 98 and 99 are discarded. Continuing along the sixth row, one finds that the next number (60) is acceptable. The next number (53) is not. The next number (81) is acceptable. The next number (97) is not. The next number (32) is acceptable. Having used all digits in the sixth row, one starts in another row—say the seventh. The first two-digit number there is 93, which is acceptable. This completes the random sampling procedure. The sample consists of the population elements whose numbers are as follows:

73, 88, 02, 36, 53, 12, 30, 71, 23, 74,

61, 59, 04, 67, 62, 83, 60, 81, 32, 93.

5.3. Absolute and Relative Error in Estimating the Binomial Parameter

5.3.a. Introduction. Certain populations commonly encountered in traffic engineering may be regarded as "binomial" in that each

*The special case in which $n = 1$ is treated in 5.2.a.1.

element either has or does not have a given attribute. Illustrative examples are: *Through* versus *local* traffic; *car* versus *bus* travel; traffic having origins in a *given zone* versus traffic from *all other zones*; parking spaces *occupied* versus parking spaces *not occupied* in a given facility.

This section discusses the relation between sample size and accuracy of estimation of the parameter of a binomial population. The results given are applied to a wide variety of traffic problems. Specifically, the three main objectives of this section are to show how to determine:

(1) the size of a sample so that the absolute or relative error of the point estimate (of the binomial parameter) is not more than a given amount with preassigned probability;

(2) an upper bound (in advance of sampling) on the relative error associated with a given sample size for a preassigned probability;

(3) "relative error" confidence limits for the binomial parameter.

5.3.b. Formulas for Sample Size. Let p be the probability that an element drawn at random from the population has a given attribute. The probability that an element drawn at random does not have the given attribute is $1-p=q$, say. For a sample of size n from the population, the point estimate, say $\hat{p}$, of p is

$$\hat{p} = \frac{\text{(number having attribute)}}{n}$$

(see Section 2.2). The point estimate, say $\hat{q}$, of q is $\hat{q}=1-\hat{p}$.

5.3.b.1. *Absolute Error.* It is often desirable to determine the sample size so that there is a high probability that the *absolute error* of $\hat{p}$ is not more than a preassigned amount, say D. (More precisely, one wishes to determine the sample size so that $\Pr\{-D<\hat{p}-p<D\} = \lambda$, where λ is a preassigned (high) probability (e.g., 0.95)). Let n_0 be the required sample size. It can be shown that

$$n_0 \doteq \left(\frac{z_a^2}{D^2}\right) p\,(1-p), \tag{5:7}$$

where $a=(1+\lambda)/2$ and z_a is the $100a$ percent point of the standard normal distribution. This result is based on the normal approximation to the distribution of $\hat{p}$.* When $p=0.5$, the quantity $p(1-p)$ in (5:7) assumes its maximum value, which is 0.25. This means that

$$n_0 \leqq \left(\frac{z_a^2}{D^2}\right)(0.25). \tag{5:8}$$

When there is no information about p in advance of sampling, the sample size should be chosen equal to the quantity on the right in (5:8). It is very helpful, of course, to have information regarding p (in advance of sampling) that makes it reasonable to assume $p(1-p)$ does not exceed a given number less than 0.25.

Example (Travel Modes). It is desired to estimate the proportion of shoppers entering a downtown department store that travel by bus to the central business district. The sample size should be sufficiently large so that the absolute error of the observed proportion $\hat{p}$ will not exceed 0.05 with a probability of 0.95. In this problem $\lambda=0.95$, $a=0.975$, $z_a=1.96$, and $D=0.05$. Substituting in (5:8), one finds that $n_0 \leq [(1.96)^2/(0.05)^2]\ (0.25)=384$. The sample size should be 384 if there is no information available (in advance of sampling) regarding the true proportion, p.

If previous studies had shown that p would not exceed 0.2, a smaller sample size would be acceptable. In this case $p(1-p)$ would not exceed $(0.2)\ (0.8)=0.16$ and n_0 would not exceed $[(1.96)^2/(0.05)^2]\ (0.16) \doteq 246$. A sample of size 246 would at least meet the requirement stated above.

5.3.b.2. *Relative Error.* In many engineering studies consideration is given to *relative* rather than *absolute* error as a basis for selecting appropriate sample sizes. It is often desirable to determine the sample size so that there is a high probability that the *relative error* of $\hat{p}$ is not more than a preassigned amount. (More precisely, one wishes to determine the sample size so that $\Pr\{-d<(\hat{p}-p)/p<d\}=\gamma$, where γ is a preassigned high probability (e.g., 0.95) and d is a preassigned bound on relative error (e.g., 0.10).) The type of problem under consideration is illustrated by the following

*When the population is a finite binomial population, n_0 can be reduced in accordance with formula (5:1).

question: How many vehicles should be counted on a given street to estimate with a high probability the proportion of in-state vehicles within a relative error of, say, 10 percent?

Let n_1 be the required sample size. It can be shown that

$$n_1 \doteq \left(\frac{\chi^2_{1,\gamma}}{d^2}\right)\frac{1-p}{p}, \tag{5:9}$$

where p is the true proportion and $\chi^2_{1,\gamma}$ is the 100γ percent point of the Chi-square distribution with 1 degree of freedom (tabled in Appendix Table 2).* Formula (5:9) is based on the normal approximation to the distribution of $\hat{p}$. Since $(1-p)/p$ becomes indefinitely large as p tends to 0, it is apparent that formula (5:9) cannot be used when there is no information regarding p in advance of sampling. Accordingly, it will be assumed that there is a *known* lower bound, say p', on p in advance of sampling. Since $p \geq p'$, it is apparent that $n_1 \leq (\chi^2_{1,\gamma}/d^2)\,(1-p')/p'$. This bound on n_1 will be regarded as the appropriate sample size. In summary, the appropriate sample size, say n, is

$$n = \left(\frac{\chi^2_{1,\gamma}}{d^2}\right)\frac{1-p'}{p'}, \tag{5:10}$$

where p' is a known *lower* bound on p, d is a specified bound on the relative error of $\hat{p}$, and $\chi^2_{1,\gamma}$ is the 100γ percent point of the Chi-square distribution with 1 degree of freedom.† If $p=p'$, the probability is γ (approximately) that in a sample of size n the relative error of $\hat{p}$ is not more than d. If p exceeds p', the required sample size would be less than n; however, the investigator would not know how much less. Accordingly, p' should be chosen as large as possible in keeping with the investigator's knowledge of the subject matter.‡

*$\chi^2_{1,\gamma} = z_a^2$, where z_a is the 100a percent point of the standard normal distribution and $a=(1+\gamma)/2$.

†When the population is a finite binomial population, n can be reduced in accordance with formula (5:1).

‡Since γ is chosen to be less than 1, there is in general a non-zero probability ($\leq 1-\gamma$) that $\hat{p}<p'\,(1-d)$.

When p' in (5:10) is a lower bound on *both* p and q, n is an appropriate sample size for which *both* the relative errors of $\hat{p}$ and $\hat{q}$ are not more than d (with probability γ).

With certain modifications formula (5:10) also gives a conservative value of the sample size when the sampling is from a multinomial distribution.* In this case p' denotes a lower bound on p_1, p_2, ..., p_K, where K is the number of categories and p_1, p_2, ..., p_K are the respective probabilities associated with the categories. $\chi^2_{1,\gamma}$ is then replaced by $\chi_{K-1,\gamma}$, the 100γ percent point of the Chi-square distribution with $(K-1)$ degrees of freedom. For illustration, consider a trinomial population consisting of trucks, buses, and passenger cars in a given traffic flow. The modified form of formula (5:10) would give a sample size more than adequate for *simultaneously* estimating all three proportions within a specified relative error, d.†

Example (Traffic Composition). How large a sample should be selected from a binomial population, consisting of passenger cars and commercial vehicles on a given highway, so that there is a probability of 0.99 that the estimates $\hat{p}$ and $\hat{q}$ are both within 10 percent of p and q, respectively? (On basis of previous experience it is reasonable to assume that both p and q are at least 0.15.) In this problem $p'=0.15$, $d=0.1$, $\gamma=0.99$, and $\chi^2_{1,\gamma}=6.63$ (see Appendix Table 2). Substituting in (5:10), one finds that the appropriate sample size is as follows:

$$n = \left(\frac{6.63}{(0.1)^2}\right)\frac{0.85}{0.15} = 663\left(\frac{17}{3}\right) = 3757.$$

If the probability required in the above example was 0.95 instead of 0.99, the quantity $\chi^2_{1,0.95}=3.84$ would replace $\chi^2_{1,0.99}=6.63$. The appropriate sample size would then be

$$n = \left(\frac{3.84}{(0.1)^2}\right)\frac{0.85}{0.15} = 384\left(\frac{17}{3}\right) = 2176.$$

*For a description of the multinomial distribution see Section 4 of the Appendix.

†The requirement of simultaneous accuracy is more stringent than that of accuracy in only one category. For this reason a larger sample is needed for simultaneous accuracy than for accuracy in only one category.

5.3.c. Graphic Solution for Sample Size, Relative Error, and Confidence Limits. For any given level of probability a family of curves may be drawn on the basis of (5:10). Such a family of curves is given in Figure 4 for a 0.95 level of probability. Each curve is a graph of the appropriate sample size, n, as a function of the relative-error bound, d, for a given value of p'.

Three important ways of using Figure 4 are described below and are summarized in Table X.

(1) *Selecting Sample Size.* Figure 4 can be used to determine the sample size, n, so that the relative error of $\hat{p}$ will be within a specified amount, d, with a probability of 0.95. It is assumed that there is a known lower bound, p', on the true proportion, p. When p' is a lower bound on p and q, the relative errors of both $\hat{p}$ and $\hat{q}$ are not greater than d (with 0.95 probability).

(2) *Determining an Upper Bound (or Limit) on Relative Error.* For a given sample size and a given value of p' Figure 4 can be used to determine a bound, d, on the relative error of $\hat{p}$ with a probability of 0.95. (This is simply a reverse of the first use.)

(3) *Determining "Relative Error" Confidence Limits.* When a sample of size n has been drawn (and the value of $\hat{p}$ has been determined), Figure 4 can be used to obtain "relative error" 95 percent confidence limits (approximately) for the true proportion, p. (In this use p' is replaced by $\hat{p}$ and *probability* is replaced by *confidence.*) The limits are expressed in terms of relative error of $\hat{p}$. (It is assumed that in advance of sampling there is no information regarding p.)

The third use of Figure 4 gives results that are approximately the same as those obtained from confidence interval charts (see Figure 1). The "relative error" confidence limits are of the form $\hat{p} \pm d\hat{p}$; thus the "relative error" confidence interval is of the form

$$\hat{p} - d\hat{p} < p < \hat{p} + d\hat{p}.$$

The approximation is generally adequate for most traffic engineering situations; however, if n is less than 100 or d exceeds 0.30 (30 percent), the approximation is not necessarily close. When accurate confidence limits are required, they can be obtained from Figure 1 or from formulas in Section 3.2.b.

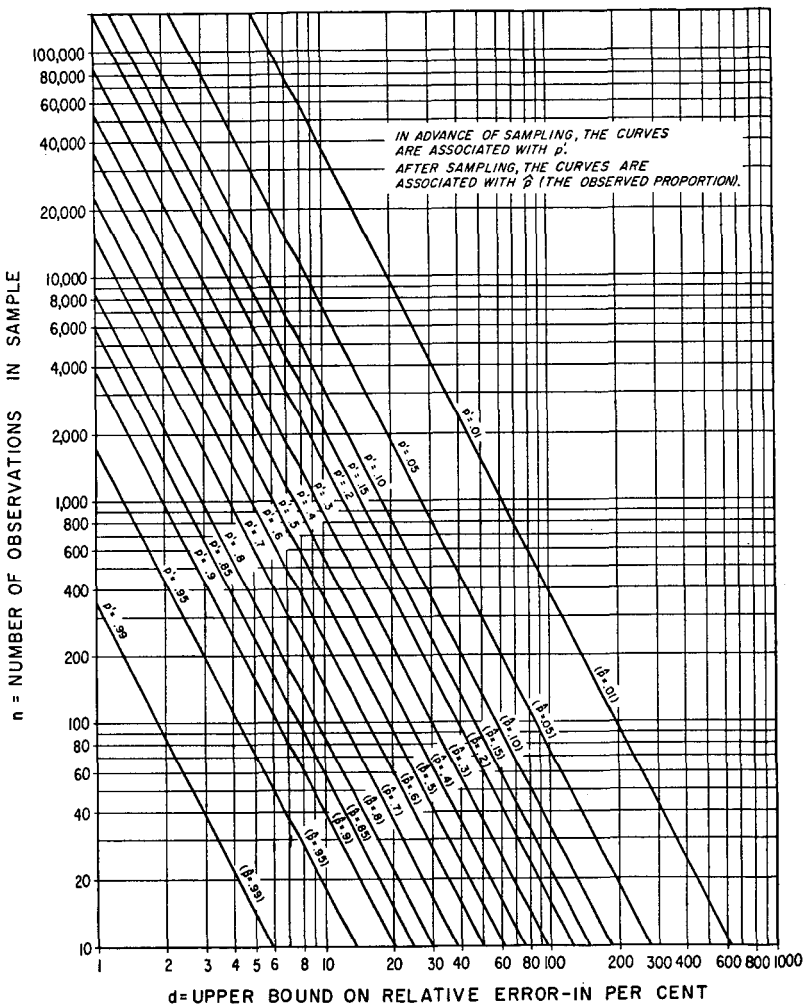


Figure 4. Relative Error in Estimating
The Binomial Parameter—0.95 Probability or Confidence.

Example of Selecting Sample Size (Traffic Composition). With regard to the example given in Section 5.3.b.2, find the appropriate sample size by means of Figure 4. As previously indicated, $p' = 0.15$, $d = 0.1$, and $\gamma = 0.95$. The solution is obtained as follows:

(1) The line labelled $p' = 0.15$ is read to where it intersects the (vertical) line corresponding to $d = 0.10$ (10 percent);

(2) The appropriate sample size is then read (horizontally) from the vertical scale. In the case at hand, $n \doteq 2200$.

Table X: Uses of Figure 4

Type of Use	*Technical Description*
Selecting Sample Size	Determining n when d and p' are given in *advance* of sampling ($\gamma = 0.95$)
Determining an Upper Bound on Relative Error	Determining d when p' and n are given in *advance* of sampling ($\gamma = 0.95$)
Determining "Relative Error" Confidence Limits	Determining d *after* the sample has been obtained (and n and $\hat{p}$ are known) ($\gamma = 0.95$)

Example of Determining An Upper Bound on Relative Error (CBD versus Through Traffic). It is desired to determine an upper bound (with 0.95 probability) on the relative error of the estimate of *through* (non-CBD) *traffic* on a given street. The estimate is to be obtained by interviewing a sample of the drivers. Field conditions limit the number of interviews to 900. Past studies indicate that the true proportion of *through traffic* is at least 0.3. The solution is obtained as follows:

(1) The line labelled $p' = 0.30$ is read to where it intersects the (horizontal) line corresponding to $n = 900$;

(2) The upper bound (limit), d, on relative error is then read (vertically) from the horizontal scale. In the case at hand $d \doteq 0.10$.

Example of Determining "Relative Error" Confidence Limits (Traffic Origins). Suppose that of 2,000 cars observed on the Merritt Parkway in Milford, Connecticut, 400 cars had New York State license plates. Regarding the 2,000 cars as a sample from a binomial population, find 95 percent confidence limits for the proportion, p, of New York State cars in the population. The confidence limits can be found as follows by means of Figure 4:

(1) Since $\hat{p}=400/2000=0.2$, read the line $\hat{p}=0.2$ to where it intersects the horizontal line corresponding to $n=2000$;

(2) The bound on the relative error is then read (vertically) from the horizontal scale; in this case $d \doteq 0.09$ ($=9$ percent), thus the relative error of $\hat{p}$ is at most 9 percent with a confidence coefficient of 0.95;

(3) The 95 percent "relative error" confidence limits are $\hat{p} \pm d\hat{p}$, which are (approximately) $0.20-0.02=0.18$ and $0.20+0.02=0.22$; thus the 95 percent confidence interval for p is

$$0.18<p<0.22.$$

(The confidence limits obtained above are in close agreement with those obtained from (3:2) or (3:3).)

5.4. Determining Sample Size for Estimating the Mean of a Population

Frequently it is desired to determine the sample size such that with a given probability the sample mean* differs from the population mean by not more than a certain amount—expressed as a given fraction of the population standard deviation. In other words, it is desired to determine the sample size so that

$$\Pr\left[-r\sigma < \bar{x} - \mu < r\sigma\right] = \lambda, \tag{5:11}$$

where λ is the given probability (e.g., 0.95), μ is the population mean, $\bar{x}$ is the sample mean, σ is the population standard deviation, and r is a given fraction (e.g., 0.1, 0.25, etc.). Let n_0 be the required value of the sample size. It can be shown that

$$n_0 \doteq \frac{z_a^2}{r^2}, \tag{5:12}$$

where $a=(1+\lambda)/2$ and z_a is the $100a$ percent point of the standard normal distribution. This solution of the problem is based on the approximate normality of the distribution of the sample mean;

*The sample mean is an unbiassed estimator of the population mean (see Section 3.5).

the result is exact (to the nearest integer) if the population is normal. If the population is finite (of size N, say), the quantity n_0 should be replaced by the quantity n'_0, say, where $n'_0 = n_0/(1+n_0/N)$ (see formula (5:1)).

Example. It is desired to estimate the average speed along a given section of highway within one mile per hour with a probability of 0.95. (On the basis of previous studies of speeds on the highway it can be assumed that the standard deviation of speeds is 10 miles per hour.) In this problem $\lambda = 0.95$, $a = 0.975$, $z_a = 1.96$, $\sigma = 10$ and $r = 1/10$. Substituting in (5:12), one finds that

$$n_0 \doteq \frac{(1.96)^2}{(0.1)^2} \doteq 384;$$

thus the required sample size is 384.* In practice about 400 speed measurements would be obtained along the given section of highway.

It is apparent from formula (5:11) that the requirement about the sample size pertains to the *absolute* error of $\bar{x}$. With probability λ the maximum absolute error is $r\sigma$, where σ is the population standard deviation and r is a fraction chosen by the investigator.†

Formula (5:12) can be used to determine the required sample size when the available information regarding σ makes it reasonable to assume that σ is known (as in the preceding example). Formula (5:12) is also useful even when there is limited information about σ (e.g., knowledge of only an upper bound).

When the numerical value of $r\sigma$ is important but there is no useful information available regarding σ, the required sample size cannot be determined from (5:12). One way of dealing with such a situation is to draw a preliminary sample to obtain useful information regarding σ.

*If the population is regarded as finite, the required sample size can be reduced. For example, if the population is regarded as the traffic on a given day and this traffic amounts to 10,000 vehicles, the required sample size is $n'_0 = n_0 / (1+n_0/N) = 384/1.0384 \doteq 370$.

†If the maximum absolute error, $r\sigma$, is expressed as G, say, then formula (5:12) can be written as follows: $n_0 \doteq z_a^2 \sigma^2/G^2$.

5.5. "Before-and-After" Studies

"Before-and-after" studies in traffic engineering are commonly undertaken to obtain information about the effect of a certain change in roadway or traffic conditions (e.g., to estimate the benefit of a specific improvement). The effect is usually expressed in terms of some measurable quantity such as traffic volume, travel time, or operating speed. For example, one might wish to obtain information on the effect of one-way routings on average speed or volume.

A statistical model of "before-and-after" studies is given below. This model is followed by an example illustrating the use of statistical methods in analyzing data obtained from a "before-and-after" study.

5.5.a. Statistical Model of a "Before-and-After" Study. A "before-and-after" study can be regarded as a comparison of one group of observations with another, where one group is obtained *before* and the other group *after* a specified change.* Ideally one wants all factors relevant to the study to be the same after the change as before—with the exception, of course, of the factor(s) in which the change is made. When the study is designed in this way, the effect of the specified change is not obscured by other effects.

From a statistical point of view it is natural to regard the group of "before" observations as a random sample from a "before" population, and the group of "after" observations as a random sample from an "after" population. By means of these samples one can *estimate* a difference between the populations or carry out a *significance test* regarding the populations. For example, one might form a point or interval estimate of the difference between the population means—such estimates are given in Section 5.5.b. below. Alternatively, one might test whether the population means are equal—several tests of this type are given in Chapter 4. Both Examples A and B in Section 4.1 deal with the question of whether the mean of a Poisson population is the same *after* a change as *before*; somewhat similar questions, associated with binomial populations, are treated in Section 4.3; a test of whether

*See Wardrop (pp. 348–351).

the means of two normal populations are equal is given in Section 4.4.b.

The use of estimation and significance testing in a "before-and-after" study will now be illustrated in a numerical example.

5.5.b. An Illustrative Example. Suppose the data tabulated below were obtained in a study of the effect, on average speed, of eliminating peak-hour curb parking.

*Item**	*Conditions*	
	Before Elimination of Peak-hour Curb Parking	*After Elimination of Peak-hour Curb Parking*
(1) Date	June 6, 1961	June 20, 1961
(2) Period of Day	4:30—5:30 p.m.	4:30—5:30 p.m.
(3) Sample Mean (Average Speed) (mph)	22.0	25.0
(4) Sample Standard Deviation (mph)	6.5	4.0
(5) Sample Size (Number of speeds observed)	50	40

*Items (1), (2), and (3) are commonly recorded but items (4) and (5) are often not recorded. It will be shown that (4) and (5) are also important for analysis of the data.

On basis of these data what conclusions can be drawn regarding the effect (if any) of eliminating peak-hour curb parking? (Can the 3 mph difference between the average speeds be regarded as merely a chance difference? What limits can be placed reasonably on the amount of error associated with the 3 mph difference?)

It is assumed that the "before" observations form a sample from a normal population with mean u_B and variance σ_B^2, say. Similarly, it is assumed that the "after" observations form a sample from a normal population with mean u_A and variance σ_A^2, say. The quantity that is primarily of interest is the difference between the true mean speeds—namely $u_A - u_B$.

Let $H = u_A - u_B$. A point estimate, $\hat{H}$, of H is $\hat{H} = \bar{x}_A - \bar{x}_B$, where $\bar{x}_A$ and $\bar{x}_B$ are, respectively, the means of the "after" and "before" samples. $\hat{H}$ is normally distributed with mean H and variance $\sigma_A^2/n_A + \sigma_B^2/n_B$, where n_A and n_B are, respectively, the sizes of the "after" and "before" samples. Let H' and H'' be lower

and upper 100λ percent confidence limits for H, respectively. H' and H'' can be expressed as follows:

$$H' = \hat{H} - z_a \sqrt{\left(\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right)},$$
$$H'' = \hat{H} + z_a \sqrt{\left(\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right)}, \tag{5:13}$$

where z_a is the $100a$ percent point of the standard normal distribution and $a = (1+\lambda)/2$.

Since $\bar{x}_A = 25.0$ mph and $\bar{x}_B = 22.0$ mph., the point estimate of the difference between mean speeds $(u_A - u_B)$ is

$$\hat{H} = 25.0 \text{ mph} - 22.0 \text{ mph} = 3.0 \text{ mph}.$$

This estimate is of course subject to error. To indicate the amount by which it may be in error, confidence limits for $u_A - u_B$ will be obtained.

If the values of σ_A and σ_B could be regarded as known, confidence limits for $u_A - u_B$ could be obtained directly from (5:13). In practice the values are often unknown, and so in this discussion they are assumed to be unknown. Under these circumstances one can still use (5:13) by replacing the population variances, σ_A^2 and σ_B^2, by the respective sample variances, say S_A^2 and S_B^2, which are estimates of σ_A^2 and σ_B^2. The resulting confidence limits are inexact; however, they are suitable for practical purposes when both sample sizes are large (≥ 30, say). (In the example above both sample sizes can be regarded as large.)

To obtain (inexact) 95 percent confidence limits for H $(= u_A - u_B)$ in the example above, one sets $z_a = 1.96$ and replaces σ_A^2 and σ_B^2 in (5:13) by $S_B^2 = (6.5)^2$ and $S_A^2 = (4.0)^2$, respectively. The resulting confidence limits are as follows:

$$H' = 3.0 - (1.96)\sqrt{\left(\frac{(6.5)^2}{50} + \frac{(4.0)^2}{40}\right)}$$
$$= 3.0 - (1.96)\sqrt{(1.245)} \doteq 3.0 - 2.19 = 0.81,$$
$$H'' \doteq 3.0 + 2.19 = 5.19.$$

Accordingly, the (inexact) 95 percent confidence interval for $u_A - u_B$ is

$$0.81 \text{ mph} < u_A - u_B < 5.19 \text{ mph}.$$

In other words, one concludes (with roughly 0.95 confidence) that the increase in average speed after elimination of peak-hour curb parking is more than 0.81 mph but less than 5.19 mph.

The 95 percent confidence interval obtained above immediately provides a test of the hypothesis that $u_A = u_B$ at an 0.05 significance level. When $u_A = u_B$, the difference $u_A - u_B = 0$. Since 0 is *not* included in the 95 percent confidence interval for $u_A - u_B$, the hypothesis that $u_A = u_B$ (i.e., that $u_A - u_B = 0$) is rejected at the 0.05 significance level. (The use of a confidence interval to carry out a significance test is discussed in Section 4.2.)

It is not necessary to compute a confidence interval for $u_A - u_B$ to carry out a significance test of the hypothesis that $u_A = u_B$. The test carried out above by means of a confidence interval can also be carried out directly by means of the following ratio:

$$\frac{\bar{x}_A - \bar{x}_B}{\sqrt{\left(\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}\right)}}.$$

The hypothesis is accepted or rejected at significance level α according as the absolute value of the ratio is or is not less than the $100(1-\alpha/2)$ percent point of the standard normal distribution. (This direct test and the test based on a confidence interval are equivalent when $\alpha = 1 - \lambda$.) Substituting the data of the example in the ratio, one finds that the ratio equals $3/\sqrt{1.245} \doteq 2.69$. Since this value exceeds 1.96 (the 97.5 percent point of the standard normal distribution), the hypothesis of equality of means is rejected at the 0.05 significance level. (This conclusion is of course the same as that reached in the test based on a 95 percent confidence interval.)

For a detailed discussion of confidence limits and significance tests pertaining to the difference of means of two normal populations see Duncan (pp. 469–481). It is apparent from that discussion that the confidence limits used in the example above can be im-

proved by certain modifications. Use of the modified limits does not require that the sample sizes be large. It should also be remarked that when σ_A^2 and σ_B^2 can be assumed equal (even though unknown), *exact* confidence limits can be obtained for $u_A - u_B$ through the use of Student's t-distribution.* A further matter of interest is that with certain refinements the significance test used in the example above becomes the Aspin-Welch test (referred to in Section 4.4.b).

5.5.c. Concluding Remarks. The illustrative example given above shows clearly that analysis of data from "before-and-after" studies requires information about not only sample means but also sample sizes and sample variances. If there were no information about the sample sizes and variances, it would not have been possible to draw statistical conclusions about the difference between mean speeds.

It should be remarked that a "before-and-after" study does not necessarily involve two samples. For instance, each of Examples A and B in Section 4.1 involves only one sample. In each case the characteristics of the "before" population were assumed to be completely known. It is also of interest that more than just two populations can be compared statistically (e.g., see the $h \times k$ contingency table in Section 4.3.d).

Finally, it should be noted that "before" and "after" represent only one kind of difference between two populations (namely a difference with regard to time). There are countless other ways in which two populations can be distinct (e.g., geographical location, type of vehicle, direction of travel, etc.).

5.6. Randomness of Traffic

Occasionally the traffic engineer wishes to determine whether traffic counts at a particular location are what would be expected from a Poisson distribution. This problem can be dealt with by means of the *Poisson index of dispersion,* which will be described below. For other ways of handling problems of this type see Greenshields and Weida (pp. 163 ff.).

Let $x_1, \ldots, x_n$ be the numbers of vehicles passing a given road-

*Section 4.4.b gives an exact significance test of the hypothesis that $u_A = u_B$ under the assumption that σ_A^2 and σ_B^2 are equal.

way point in each of n equal time intervals, respectively. It is assumed that each x_i $(i=1, \ldots, n)$ is drawn from a Poisson distribution. The null hypothesis* is that all n of the x's come from the same Poisson distribution. To test the null hypothesis one can use the *Poisson index of dispersion*, which can be written as follows:

$$\sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{\bar{x}} = \frac{\sum_{i=1}^{n} x_i^2}{\bar{x}} - n\bar{x}, \qquad (5:14)$$

(see Hoel, p. 178). When the null hypothesis is true, this quantity has a Chi-square distribution (approximately) with $n-1$ degrees of freedom. The null hypothesis is rejected at significance level α when the Poisson index exceeds the $100(1-\alpha)$ percent point of the Chi-square distribution with $n-1$ degrees of freedom.

Example. Suppose that the following counts were made in successive 5-minute periods: 20, 18, 25, 22, 16, 24, 17, 23, 21, 15, 18, 23. It is desired to test, at an 0.05 level of significance, the null hypothesis that all observations have come from the same Poisson distribution. In this example $n=12$, $\bar{x}=(242/12)$, $n\bar{x}=242$, and $\sum_{i=1}^{n} x_i^2=5002$. Substituting these values in (5:14), one finds that

$$\sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{\bar{x}} = \frac{5002}{242/12} - 242 \doteq 248.03 - 242 = 6.03.$$

Since $n=12$, the Chi-square distribution associated with (5:14) has $12-1=11$ degrees of freedom. The 95 percent point of the Chi-square distribution with 11 degrees of freedom is 19.7 (see Appendix Table 2). Since the observed value of the Poisson index is smaller than 19.7, the null hypothesis is accepted at the 0.05 level of significance. (For a general description of significance testing see Chapter 4.)

5.7. Estimation of Traffic Volume by Means of Short Counts

Traffic volume is of interest in almost every traffic study. In some cases the volume must be determined precisely, and this means in general that all the traffic involved must be counted. Often,

*See Chapter 4 for descriptions and illustrations of null hypotheses.

however, a statistical estimate of volume will be satisfactory since pin-point accuracy is not necessary. When extreme accuracy is not required, estimation is much more practical than precise determination since the costs (in money and effort) are much less.

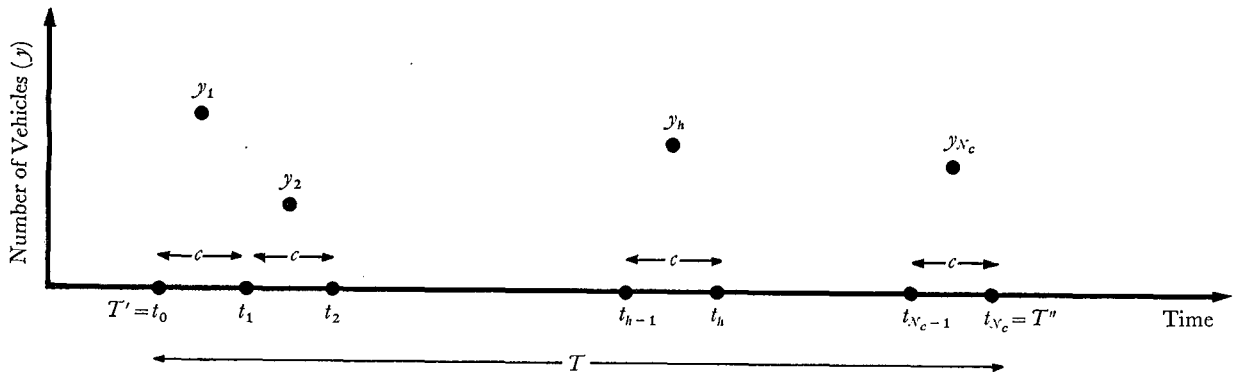
This section deals with the subject of traffic-volume estimation by means of a sample of short counts. Short-count estimation, which began with the work of McClintock, has been the subject of many interesting studies. Vickery, for example, has given some interesting mathematical arguments in support of the use of short counts. For more recent studies see Adams, Burch, White and Pelz, and the unpublished report of the Bureau of Highways of the Department of Public Works of Puerto Rico.*

The *population of short counts*, shown in Figure 5, is a fundamental concept associated with the estimation of traffic volume. In this section two different types of sampling from this population are considered—namely, *random sampling* and *systematic sampling*. Point and interval estimates of volume are given for each type.

5.7.a. The Population of Short Counts. Consider a particular roadway location and a time period (T', T''), which will be called the *base period*. The length of this period is $T'' - T' = T$, say. Divide the base period into N_c equal subperiods (t_0, t_1), (t_1, t_2), ..., (t_{N_c-1}, t_{N_c}). The length of each subperiod is represented by c. The time t_0 equals T' and the time t_{N_c} equals T''. (The quantities described above are shown in Figure 5.)

Let $y_1, y_2, \ldots, y_{N_c}$ be the respective numbers of vehicles passing the location in the N_c subperiods. For example: y_1 is the number of vehicles passing the location in the first period, from t_0 up to the time t_1; y_2 is the number passing in the second period, from t_1 up to the time t_2; ...; and y_{N_c} is the number passing in the last period, from t_{N_c-1} up to (and including) the time t_{N_c} (see Figure 5).

*These and other short count studies have indicated that: 1. Different purposes require different degrees of accuracy; 2. Traffic follows daily and hourly patterns that are generally consistent and often predictable; e.g., the total daily volume does not vary materially among different weekdays; 3. The heavier the traffic volume at a particular location, the greater the accuracy of short count methods; 4. The more counts (even though of short duration), the greater the accuracy; and 5. When traffic is not light or unduly erratic, counts of five or six-minute duration are entirely satisfactory.



REMARKS

Length of Base Period $= T'' - T' = T$,

N_c = Number of Short-Count Periods in the Base Period,

c = Length of Each Short-Count Period,

$T = cN_c$,

Population Elements are $y_1, y_2, \ldots, y_h, \ldots, y_{N_c}$,

Population Mean Equals $\left(\sum_{h=1}^{N_c} y_h\right) / N_c = \mu_c$, say,

μ_c is the Average Volume Per Short-Count Period,

Volume Equals $\sum_{h=1}^{N_c} y_h = V$, say,

$V = N_c \mu_c$,

Population Variance Equals

$$\sum_{h=1}^{N_c} (y_h - \mu_c)^2 / N_c = \sigma_c^2, \text{ say.}$$

Figure 5. Schematic Representation of the Base Period (T', T''), The Short-Count Periods. and the Population of Short Counts.

The total number of vehicles passing the location in the entire base period is $\sum_{h=1}^{N_c} y_h = V$, say. V is called the *volume* of traffic past the location in the base period. Each of the quantities $y_1, y_2, \ldots, y_{N_c}$ is called a *short count.* This terminology is used since the length, c, of the subperiod associated with any of the y's is usually "short" (e.g., about five or ten minutes). The subperiod is called a *short-count* period.

It should be noted that V is related to the mean of the population of N_c short counts $y_1, y_2, \ldots, y_{N_c}$. More specifically, V equals $N_c \mu_c$, where μ_c is the population mean. Since N_c is a known constant (chosen by the investigator), the problem of estimating V is equivalent to that of estimating the population mean, μ_c. For example, an interval estimate of V can be obtained from a random sample of short counts; such estimates are given in Section 5.7.b below. They are very simple modifications of formulas given in Section 3.5 for interval estimates of population means.

The variance of the population $y_1, y_2, \ldots, y_{N_c}$ will be represented by σ_c^2. The quantity σ_c^2 is a critical parameter with regard to both point and interval estimation of μ_c (and thus V). For example, if σ_c^2 is known to be 0*, then V can be estimated with perfect accuracy by observing only one of the $N_c y$'s. The reason is not hard to find. When $\sigma_c^2 = 0$, all population elements, $y_1, y_2, \ldots, y_{N_c}$ are equal; in fact, each equals V/N_c. After observing the value of any one of the y's, the investigator would multiply that value (namely V/N_c) by N_c to obtain the value of V. When σ_c^2 is relatively large, the error of estimate of V is not likely to be small unless extensive counting is done.

The general problem under consideration in Section 5.7 is the estimation of the volume, V, by means of a sample from the population of short counts. The subject of 5.7.b is estimation by means of a *random sample* of short counts. The subject of 5.7.c is estimation by means of a *systematic sample* of short counts. Estimation of V by means of *stratified sampling* is discussed briefly in 5.7.d.

In practice, short-count periods are usually selected according

*Of course in practice σ_c^2 would seldom, if ever, be equal to 0.

to some predetermined schedule so as to be equally spaced apart. A systematic sample of short counts has this property of equal spacing, but a random sample does not; thus there is a serious question as to the practicality of drawing a random sample of short counts. Although the results set forth in 5.7.b are based on random sampling, they are nevertheless useful since they indicate the relation between accuracy of volume estimates and the number of short counts obtained.

5.7.b. Estimation of Volume By Means of a Random Sample of Short Counts. Let $Y_1, Y_2, \ldots, Y_n$ be a random sample (drawn without replacement) from the population $y_1, y_2, \ldots, y_{Nc}$ described in 5.7.a above. Let $\bar{Y}_c$ be the mean of the sample.

5.7.b.1. *Point Estimation of Volume.* A point estimate, $\hat{V}$, of V is

$$\hat{V} = N_c \bar{Y}_c. \tag{5:15}$$

The expected value of $\hat{V}$ is $N_c \mu_c = V$; thus $\hat{V}$ is an unbiassed estimate of V. The variance, $\sigma^2_{\hat{V}}$, of $\hat{V}$ is

$$\sigma^2_{\hat{V}} = N_c^2 \left(\frac{\sigma_c^2}{n}\right)\left(\frac{N_c - n}{N_c - 1}\right) \tag{5:16}$$

since $\sigma^2_{\hat{V}} = N_c^2 \sigma^2_{\bar{Y}_c}$ (see Section 3.5).

5.7.b.2. *Interval Estimation of Volume.* When σ_c can be regarded as known and the distribution of $\bar{Y}_c$ is approximately normal, formula (5:17) below gives approximate lower and upper 100λ percent confidence limits for V:

$$\begin{aligned} V' &= N_c \bar{Y}_c - z_a N_c \frac{\sigma_c}{\sqrt{n}} \sqrt{\left(\frac{N_c - n}{N_c - 1}\right)}, \\ V'' &= N_c \bar{Y}_c + z_a N_c \frac{\sigma_c}{\sqrt{n}} \sqrt{\left(\frac{N_c - n}{N_c - 1}\right)}, \end{aligned} \tag{5:17}$$

where $a = (1 + \lambda)/2$ and z_a is the $100a$ percent point of the standard normal distribution. Formula (5:17) is obtained from formula (3:9) by simply multiplying μ' and μ'' in (3:9) by N_c. The remarks in Section 3.5 regarding the validity of (3:9) also apply to the validity of (5:17).

Example. Suppose that from a four-hour base period a random sample of 20 short counts is drawn—the length of each short-count period being five minutes. Suppose also that the sample mean equals 90. Finally, suppose that past experience suggests that the population standard deviation, σ_c, equals 10. Find approximate lower and upper 95 percent confidence limits for the volume, V. In this example $N_c=(4\times 60)/5=48$, $n=20$, $\bar{Y}_c=90$, and $\sigma_c=10$. Substituting in (5:17) one obtains the following results:

$$V' = (48)(90) - (1.96)(48)\left(\frac{10}{\sqrt{20}}\right)\sqrt{\frac{28}{47}}$$
$$\doteq 4320 - 162.4 \doteq 4158,$$
$$V'' \doteq 4320 + 162.4 \doteq 4482.$$

An approximate 95 percent confidence interval for V is therefore

$$4158 < V < 4482.$$

When the value of σ_c is unknown and the distribution of $\bar{Y}_c$ can be regarded as approximately normal, formula (5:18) below gives approximate lower and upper 100λ percent confidence limits for V:

$$V' = N_c\,\bar{Y}_c - t_a\frac{N_c\,S_y}{\sqrt{(n-1)}}\sqrt{\left(\frac{N_c-n}{N_c-1}\right)},$$
$$V'' = N_c\,\bar{Y}_c + t_a\frac{N_c\,S_y}{\sqrt{(n-1)}}\sqrt{\left(\frac{N_c-n}{N_c-1}\right)}, \qquad (5:18)$$

where $a=(1+\lambda)/2$, t_a is the $100a$ percent point of the t-distribution with $(n-1)$ degrees of freedom, and S_y is the sample standard deviation. Formula (5:18) is obtained from formula (3:10) by multiplying μ' and μ'' in (3:10) by N_c. Formula (5:18) is valid whenever formula (3:10) is.

5.7.b.3. *Sample Size For Relative Error of Point Estimate To Be Within a Preassigned Amount (With High Probability).* In setting up a schedule of short counts one may wish to know how many periods to select so that the relative error of estimate is likely to be small. In other

words, one may wish to determine the sample size, n, so that there is a "high" probability that the following inequality is true:

$$-d \leq \frac{\hat{V}-V}{V} \leq +d,$$

where d is a "small" fraction (e.g., 0.10). Let n' be the least value of n such that the probability is at least λ, say, that the above inequality holds. (In practice one might choose λ to be, say, 0.75, 0.90, or 0.95.) It can be shown that n' is (approximately) the smallest value of n satisfying the following inequality:

$$n \geq \frac{N_c^3 z_a^2 \sigma_c^2}{N_c^2 \sigma_c^2 z_a^2 + d^2 V^2 (N_c - 1)}, \tag{5:19}$$

where V is the volume, $a=(1+\lambda)/2$, z_a is the $100a$ percent point of the standard normal distribution, and σ_c^2 is the variance of the short-count population. This result is based on the assumption that the distribution of $\bar{Y}_c$ (and thus that of $\hat{V}$) is approximately normal. If the functional form of the distribution of $\hat{V}$ cannot be regarded as known (at least approximately), the quantity z_a in (5:19) should be replaced by $1/\sqrt{(1-\lambda)}$; in general this leads to a conservative value of n'.

It is evident that formula (5:19) cannot be used in practice unless there is some information available (in advance of sampling) regarding V* and σ_c^2. Two assumptions will be made regarding the available information about V and σ_c^2. These assumptions, which are discussed below, lead to a modified form of (5:19) that is generally useful. The first assumption is that there is a known lower bound, V_L, on V, where $V_L > 0$. The second assumption is that

$$\sigma_c^2 \doteq \frac{V}{N_c}\left(1-\frac{1}{N_c}\right). \tag{5:20}$$

*Of course, this does not mean that V would be known precisely. If V were known precisely (in advance of sampling), there would be no need to estimate V and thus no need to obtain short counts.

With these assumptions it can be shown that n' is (approximately) the smallest value of n satisfying the following inequality:*

$$n \geq \frac{N_c z_a^2}{z_a^2 + d^2 V_L}. \tag{5:21}$$

Engineering considerations indicate that the assumption stated in (5:20) is satisfactory if there are no marked changes in the intensity of traffic flow during the base period. It is also of interest to assume that $y_1, y_2, \ldots, y_{N_c}$ form a sample of size V from a multinomial population having a probability of $1/N_c$ for each of its N_c categories. This alternative assumption leads to simplifications in the sampling theory associated with traffic-volume estimation. For example, formula (5:21) can then be derived in a very simple way. Another interesting consequence of this assumption is that the expected value of the variance of the y's is equal to the quantity on the right-hand of (5:20).

It is clear from (5:21) that the assumption of a known positive lower bound V_L (on V) is important. If V_L were 0 (which, of course, is unlikely), then the right-hand side of (5:21) would equal N_c and so the smallest sample size satisfying (5:21) would be N_c. This would mean that counting should be done throughout the base period to achieve the required accuracy.

Example. Suppose that one wishes to estimate the volume over a four-hour base period to within 10 percent relative error with a probability of at least 0.95. Suppose also that the volume can be assumed to be at least 2,800. Finally, suppose that five-minute short counts will be used. Find the minimum sample size necessary to meet the requirements stated above. In this example $c=5$ minutes, $N_c=(4\times 60)/5=48$, $V_L=2800$, $\lambda=0.95$, $a=(1+0.95)/2$, and $z_a=1.96$. Substituting in (5:21) one obtains the following inequality:

$$n \geq \frac{(48)\ (1.96)^2}{(1.96)^2+(0.01)\ (2800)} \doteq 5.79.$$

*In advance of sampling this inequality could be used to obtain a bound on the relative error of $\hat{V}$ (with probability λ).

It follows that the minimum sample size, n', that meets the requirements is 6. Since the length of each short-count period is 5 minutes, the total counting time would be $6 \times 5 = 30$ minutes. The ratio of the length of the base period to the total counting time would therefore be $240/30 = 8$; thus the percent of time counted would be $(30/240)\ 100 = 12.5$.

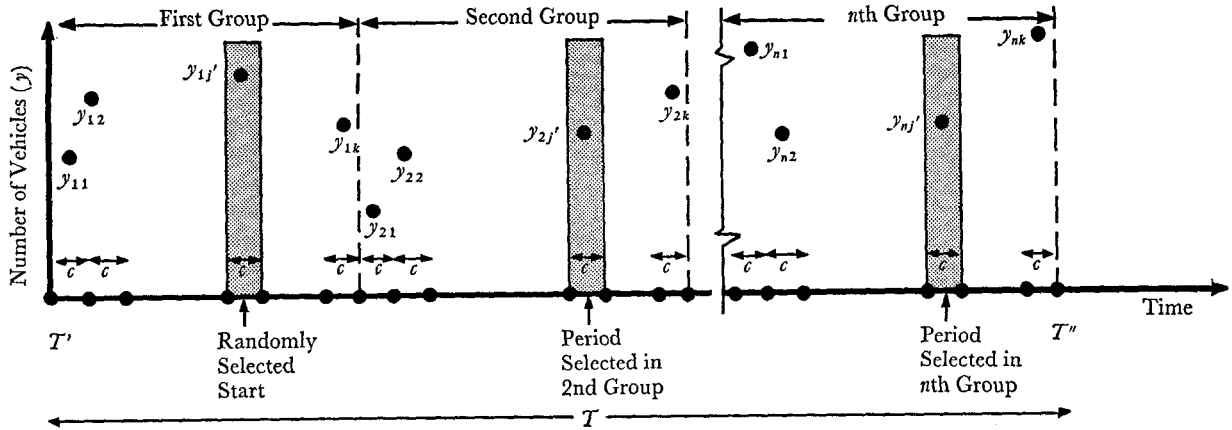
5.7.c. Estimation of Volume by Means of a Systematic Sample of Short Counts. In Section 5.7.b above the type of sampling under consideration was random sampling. In the present section a different type of sampling—namely, *systematic sampling*—is considered. A new description of the population of short counts is given to simplify the description of this type of sampling. After these descriptions are given, point and interval estimation of volume will be considered. The final topic of Section 5.7.c is the design of a sample so that with high probability the relative error of estimate does not exceed a preassigned amount.

5.7.c.1. *The Population and the Sampling Procedure.* Divide the total base period (T', T'') into n equal periods, and divide each of these periods into k equal subperiods (see Figure 6). Let the length of each subperiod be represented by c. It is clear that $nkc = T$, where T is the length of the base period. Each of the subperiods is a short-count period. It will be convenient to represent these nk short-count periods as $(1, 1), (1, 2), \ldots, (1, k), (2, 1), (2, 2), \ldots, (2, k), \ldots, (n, 1), (n, 2), \ldots, (n, k)$; for example, (i, j) represents the jth short-count period in the ith group of short-count periods $(i = 1, \ldots, n; j = 1, \ldots, k)$. Let y_{ij} be the number of vehicles passing a roadway point in the short-count period (i, j). The set of nk numbers $y_{11}, y_{12}, \ldots, y_{1k}, \ldots, y_{n1}, y_{n2}, \ldots, y_{nk}$ is the population of short counts (see Figure 6). The volume, V, over the base period is the sum of the y_{ij}'s—i.e.,

$$V = \sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij}.$$

The sampling procedure is as follows:

(1) Select one of the first k short-count periods $(1, 1), \ldots, (1, k)$ purely at random. To accomplish this one selects an integer, say j', purely at random from $1, 2, \ldots, k$; the short-count period



REMARKS

$T'' - T' = T$,

n = Number of Groups,

k = Number of Short-Count Periods in Each Group,

c = Length of Each Short-Count Period,

$nkc = T$,

Population Elements:

$y_{11}, y_{12}, \ldots, y_{1k}, \ldots, y_{n1}, y_{n2}, \ldots, y_{nk}$,

Volume = Sum of All Population Elements,

Elements of Sample Described in Diagram above:

$y_{1j'}, y_{2j'}, \ldots, y_{nj'}$,

Sample Sum:

$y_{1j'} + y_{2j'} + \ldots + y_{nj'} = W_{j'}$ say $(j' = 1, \ldots, k)$,

Set of Possible Sample Sums: $W_1, W_2, \ldots, W_k$,

Volume $= W_1 + W_2 + \ldots + W_k = V$,

Point Estimate of Volume: $\hat{V} = kW_{j'}$,

Variance of $\hat{V}$: $\sigma_{\hat{V}}^2 = k^2\sigma_k^2$, where σ_k^2 is the Variance of $W_1, W_2, \ldots, W_k$.

Figure 6. Schematic Representation of the Base Period, the Short-Count Periods, the Population of Short Counts, and the Systematic Sample.

$(1, j')$ is then a short-count period selected at random from $(1, 1), \ldots, (1, k)$. (See the illustrative example in 5.2.a.1.)

(2) Then select the short-count periods $(2,j')$, $(3,j')$, $\ldots$, (n,j').

(3) Count the number of vehicles in each short-count period selected—i.e., in $(1,j')$, $(2,j')$, $\ldots$, (n,j'). These counts are $y_{1j'}$, $y_{2j'}$, $\ldots$, $y_{nj'}$. Thus in each of the n groups one out of the k subperiods is counted.

Since the first short-count period selected—namely $(1, j')$—is selected at random, the sampling procedure involves a "random start." All other selections of short-count periods are determined by the outcome of the random start. It should also be noted that the short-count periods selected are equally spaced apart—the spacing being k short-count periods (see Figure 6). In a sense the sample size is n; however, the sample is *systematic* rather than *random*. Because of its systematic feature such a sample is more suitable for practical use than a random sample.

5.7.c.2. *Point Estimation of Volume.* When a sample has been drawn in accordance with the procedure described above, a point estimate, $\hat{V}$, of the volume, V, can be expressed as follows:

$$\hat{V} = k\, W_{j'}, \tag{5:22}$$

where $W_{j'} = \sum_{i=1}^{n} y_{ij'}$. $W_{j'}$ is simply the sum of the counts that make up the sample. It should be noted that $W_1, W_2, \ldots, W_k$ are the possible values of the sample sum in advance of sampling. Selecting a "start" at random implies selecting one of the W's at random. It should also be noted that $W_1 + W_2 + \ldots + W_k = V$; hence

$$\frac{V}{k} = \mu_k, \tag{5:23}$$

where μ_k is the mean of the W's.

The expected value of $\hat{V}$ is $k\mu_k = V$ (see (5:23)). This means that $\hat{V}$ is an unbiassed estimate of V. The variance, $\sigma^2_{\hat{V}}$, of $\hat{V}$ is

$$\sigma^2_{\hat{V}} = k^2\, \sigma_k{}^2, \tag{5:24}$$

where $\sigma_k{}^2$ is the variance of $W_1, W_2, \ldots, W_k$.

From an engineering point of view it seems satisfactory to assume that

$$\sigma_k^2 \doteq V\left(\frac{1}{k}\right)\left(1-\frac{1}{k}\right), \tag{5:25}$$

provided there are no marked changes of traffic-flow intensity in any one of the n periods of the base period. From a theoretical point of view it is of interest to make the following, alternative assumption: in the ith period ($i=1, \ldots, n$) the y_{ij}'s form a sample of size V_i* from a multinomial population having a probability of $1/k$ for each of its k categories. An interesting consequence of this assumption is that $\hat{V}$ (defined in (5:22)) is approximately normal with mean V and variance $V(k-1)$. (This property of $\hat{V}$ is used in 5.7.c.4 below.) Another interesting consequence is that the expected value of the variance of the W's is equal to the quantity on the right-hand side of (5.25).

5.7.c.3. *Interval Estimation of Volume.* Conservative lower and upper 100 $(1-1/B^2)$ percent confidence limits for the volume, V, are as follows:†

$$\begin{aligned} V' &= \hat{V} - B\,\sigma_{\hat{V}}, \\ & \qquad (B > 1) \\ V'' &= \hat{V} + B\,\sigma_{\hat{V}}, \end{aligned} \tag{5:26}$$

where $\sigma_{\hat{V}}$ is defined in (5:24). Such an interval can be used when the value of $\sigma_{\hat{V}}$ is known or when a satisfactory engineering estimate of it is available.

It is also of interest to have confidence limits for V when there is good reason to assume that $\sigma_k^2 \doteq V\,(1/k)\,(1-1/k)$. This assumption is stated in (5:25) and discussed immediately below (5:25). When the assumption holds, conservative lower and upper 100 $(1-1/B^2)$ percent confidence limits for V are as follows:†

*V_i is the volume associated with the ith period (i.e., $V_i = \sum_{j=1}^{k} y_{ij}$). The quantities $V_1, \ldots, V_n$ need not be equal; in fact, they may vary considerably from one part of the base period to another.

†Formulas (5:26) and (5:27) are based on the Bienaymé-Tchebycheff inequality. (See Section 3 of the Appendix.) When $\hat{V}$ is approximately normal with mean V and variance $V(k-1)$, a slight modification of formula (5:27) provides approximate 100λ percent confidence limits for V. The modification consists of replacing B by z_a $(a=(1+\lambda)/2)$.

$$V' \doteq \hat{V} + \frac{B^2(k-1)}{2} - B\sqrt{\left(\hat{V}(k-1) + \frac{B^2(k-1)^2}{4}\right)},$$

$$V'' \doteq \hat{V} + \frac{B^2(k-1)}{2} + B\sqrt{\left(\hat{V}(k-1) + \frac{B^2(k-1)^2}{4}\right)}. \tag{5:27}$$

Example. Suppose that an engineer has obtained a systematic sample of five-minute short counts from a two-hour base period, as follows:

(1) The base period was divided into four 30-minute periods, and each of these four periods was divided into six 5-minute short-count periods.

(2) A short-count period was selected at random from the first six, and the corresponding short-count periods in the three remaining 30-minute periods were selected. The sum of the four counts was 405.

On the assumption stated in (5:25), find a conservative 75 percent confidence interval for V. In this example $k=6$, $B=2$ (since $1-1/2^2=0.75$), and $\hat{V}=6\times(405)=2430$. Substituting in (5:27) one obtains the following results:

$$V' \doteq 2430 + 4\times\frac{5}{2} - 2\sqrt{\left(2430\times 5 + \frac{4\times 25}{4}\right)}$$
$$= 2440 - 2\sqrt{(12175)} \doteq 2440 - 221 = 2219,$$
$$V'' \doteq 2440 + 221 = 2661.$$

The required confidence interval is therefore

$$2219 < V < 2661.$$

5.7.c.4. *Choice of k For Relative Error of Estimate to be Within a Preassigned Amount (With High Probability).* The problem considered here is that of choosing k so that there is high probability that the relative error of $\hat{V}$ is not greater than a preassigned amount, d. This problem is similar in some respects to the one treated in Section 5.7.b.3. The solution, given in (5:29) below, is based on two assumptions. One is the special assumption introduced in the discussion below (5:25); the other assumption is that there is a known lower bound, V_L, on the volume, V.

It is required that a value, say k', of k be chosen so that the probability is at least λ that

$$-d \leq \frac{\hat{V}-V}{V} \leq +d. \tag{5:28}$$

The value of λ is specified in advance of choosing k. Under the assumptions stated above the required value, say k', of k is the largest value of k satisfying the following inequality:

$$k \leq 1 + \frac{V_L d^2}{(z_a)^2}, \tag{5:29}$$

where V_L is the known lower bound on V, z_a is the $100a$ percent point of the standard normal distribution, and $a=(1+\lambda)/2$.*

Example. Suppose an engineer wishes to estimate the volume over a four-hour base period to within 10 percent with a probability of at least 0.95. Suppose further that it is reasonable to assume that the volume V is at least 2800. What value of k is required? In this example $V_L=2800$, $d=0.1$, and $z_a=1.96$. Substituting in (5:29) one finds that

$$k \leqq 1 + 2800\,\frac{0.01}{1.96^2} \doteq 1 + 7.29 = 8.29.$$

The required value k' is therefore equal to 8. (The reader may find it interesting to compare the results of this example with those of the example given in 5.7.b.3.)

Since k represents the ratio of the total time in the base period to the total time actually counted, the selection of 8 as the value of k would mean that the total counting time equals one-eighth of the base period. Since the length of the base period is 240 minutes, the total counting time is $(1/8)\ (240)=30$ minutes. The engineer would still have some choice as to the length, c, of the short-count period and the number, n, of short counts obtained. For example, he could choose c to be 5 minutes and n to be 6 or choose c to be 10 minutes and n to be 3. The first choice would lead to a systematic

*If the assumption in (5:25) is preferred to the one introduced below (5:25), a *conservative* value of k can be obtained by a slight modification of (5:29). The modification consists of replacing z_a by B $(B>1)$; and the value of the probability involved is then $1-1/B^2$ instead of λ.

sample of six 5-minute short counts, and the second would lead to a systematic sample of three 10-minute short counts. It should be remarked, however, that there are theoretical reasons for preferring the choice of six 5-minute short counts (see the discussion below [5:25]).

5.7.d. Estimation of Daily Volume Using Stratified Sampling. Stratified sampling is often used in dealing with very heterogeneous populations. The aim in stratification is to obtain strata such that each stratum is less heterogeneous than the original population.

The procedures associated with stratified sampling can be described as follows (see McCarthy [p. 282]):

1. Divide the population into mutually exclusive and exhaustive subgroups (or *strata*);
2. Draw a sample from each stratum;
3. Make an estimate for each stratum, and combine these estimates to obtain an estimate for the entire population.

It is natural to consider the use of stratified sampling in estimating the daily volume of traffic since the population of counts associated with traffic throughout a given day is quite irregular. The traffic density shows conspicuous "rises" and "falls" which must be taken into account. Practical experience suggests that it would be suitable to divide the population into, say, four strata. These strata would correspond to the following periods of the day: morning rush period, midday, afternoon rush period, and mid-evening.* Each of these periods can be regarded as a *base period* for which an estimate of volume can be made along the lines indicated in Sections 5.7.c (or 5.7.b in some cases). An estimate of daily volume could then be made by combining the four separate estimates.

5.8. Concluding Remarks

The ideas developed in the first four chapters have been used extensively in Chapter 5. Each of the preceding sections in this chapter gives traffic engineering applications of sampling techniques or concepts, and each (except 5.2) gives one or more applications of estimation or significance testing. (The reader may

*In general any part of the day in which there is non-negligible traffic would be included in one of these periods.

be interested to note that *point estimation* is involved in Sections 5.1, 5.3, 5.4, 5.5 and 5.7, that *interval estimation* is involved in 5.1, 5.3, 5.5, and 5.7, and that *significance testing* is involved in 5.5 and 5.6.)

In exploring the relation between sampling and traffic engineering it has been necessary to go into great detail on many statistical and engineering matters. As a result, the pages of this book contain many formulas and a large amount of technical discussion. It remains to provide the reader with an over-all view of the basic role of sampling in traffic engineering. The remarks below are designed to meet this need.

Throughout the course of his work in planning, designing, and operating transportation facilities, the traffic engineer needs to obtain information about the characteristics of populations with which he is concerned. Such information is used in drawing conclusions or making decisions. By means of sampling, the engineer can obtain useful information regarding characteristics of these populations; however, this information is almost always incomplete and therefore involves some uncertainty. (In order that there be no uncertainty it is necessary that the sample consist of the entire population; however, owing to limitations of time, money, manpower, or other resources, it is usually not feasible to draw such a sample.) Procedures of estimation and significance testing make it possible for the engineer to take into account the uncertainty associated with the information obtained from sampling. It is important that the uncertainty be taken into account correctly, since failure to do so may lead to wrong conclusions or unsound decisions. By proper use of sampling the engineer can control the uncertainty associated with the information obtained.

The methods and concepts presented in this book are quite general; thus they are powerful tools for the traffic engineer who understands how to apply them. Many applications of these tools are shown in the illustrative examples given throughout the book; however, their range of applicability includes far more than the specific subject matter of the examples. Equipped with these tools, the traffic engineer can grasp the statistical aspects of his problems more readily and thereby deal with his problems more effectively.

Appendix

1. Populations and Samples

Population. A population is a set (collection) of objects. The set may be finite or infinite.

Sampling. Sampling a population is the selection (drawing) of one or more elements of the population. To select or draw an element does not necessarily mean that the element is removed from the population. It may mean only that the element is observed. Almost all the sampling considered in this book is random sampling.* It should also be remarked that in this book two slightly different meanings are associated with the phrase *random sampling.* The usual meaning is that one or more *random variables* are associated with the sampling procedure—in a manner indicated in the paragraph immediately below. (For a definition of a random variable see Section 3 of the Appendix.) The other meaning associated with random sampling is used only in the special case where the sampling is *without replacement* from a finite population.† Such sampling is said to be random if all possible samples have the same chance of being selected.

Sample. A sample can be regarded as a set of elements drawn from a population. Unless otherwise indicated, the elements of the population are assumed to be numbers. When the sampling is random (in the first sense described in the paragraph above), the sample can be regarded as a set of observed values of one or more random variables. For the most part in this book a random sample is considered to be simply a set of observed values, $x_1, \ldots, x_n$, of a random variable X that have been obtained independently.

The remarks above indicate two different ways of describing the source of a sample. One can say that the *sample comes from a popula-*

*Techniques of random sampling are presented in detail in Section 5.2.

†When sampling is *without replacement,* no element drawn is returned to the population. When sampling is *with replacement,* each element drawn is returned to the population before another drawing is made.

tion or that it *consists of observed values of a random variable.* A third way is to say that the *sample comes from a distribution.* The third mode of expression is connected with the fact that a random variable is characterized by a distribution. (See Section 3 of the Appendix for a description of a distribution.)

2. Functions of Samples

With regard to the definitions given below it is assumed that the sample consists of n numbers, say $x_1, \ldots, x_n$.

Sample Sum. The sample sum is

$$x_1 + \ldots + x_n = S(x), \text{ say.} \tag{1}$$

Sample Mean. The sample mean, say $\bar{x}$, is

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} \tag{2}$$

Note that $\bar{x} = S(x)/n$, thus $S(x) = n\,\bar{x}$.

Sample Variance. The sample variance, say S_x^2, is defined as follows:

$$S_x^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}, \tag{3}$$

where $\bar{x}$ is the sample mean. It can be shown easily that

$$S_x^2 = \frac{x_1^2 + \ldots + x_n^2}{n} - \bar{x}^2. \tag{4}$$

Note: Some writers define the sample variance as

$$\frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{(n-1)}, \quad (n > 1).$$

Sample Standard Deviation. The sample standard deviation, S_x, is defined as follows:

$$S_x = + \sqrt{\left(\frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}\right)}. \tag{5}$$

It should be remarked that the square of the sample standard deviation is the sample variance. (See also the note above regarding the sample variance.)

Order Statistics. Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be an arrangement of $x_1, \ldots, x_n$ in increasing order of magnitude; thus $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. The quantities $x_{(1)}, \ldots, x_{(n)}$ are called *order statistics.* $x_{(r)}$ is called the *rth order statistic.* For example, the order statistics of the sample 7, 5, 12, 2, 9 are $x_{(1)} = 2, x_{(2)} = 5, x_{(3)} = 7, x_{(4)} = 9, x_{(5)} = 12$.

Sample Median. Let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ be the order statistics of a sample of size n. The sample median, say $\tilde{x}$, is defined as follows:

$$\tilde{x} = \begin{cases} x_{(k+1)} \text{ when } n = 2k+1 \ (k = 0, 1, \ldots) \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} \text{ when } n = 2k \ (k = 1, 2, \ldots). \end{cases} \tag{6}$$

In words: when n is odd (say $n = 2k+1$), $\tilde{x}$ is the $(k+1)$th order statistic; when n is even (say $n = 2k$), $\tilde{x}$ is the average of the kth and $(k+1)$th order statistics. In the illustrative example above, the median of the sample of size 5 is $x_{(3)} = 7$.

Sample Range. The sample range is defined as the difference between the largest and smallest elements of the sample. In terms of order statistics of a sample of size n, the sample range is $x_{(n)} - x_{(1)}$. In the illustrative example above, the sample range is $x_{(5)} - x_{(1)} = 12 - 2 = 10$.

For every sample the standard deviation, S_x, and the range satisfy the following inequality:

$$S_x \leq \frac{x_{(n)} - x_{(1)}}{2}. \tag{7}$$

Sample Cumulative Distribution Function. The sample cumulative distribution function, say $F_n(x)$, is defined as follows:

$$F_n(x) = \frac{\text{number of values among } (x_1, \ldots, x_n) \leq x}{n}, \ (-\infty < x < +\infty). \tag{8}$$

It is apparent from this definition that $0 \leq F_n(x) \leq 1 \ (-\infty < x < +\infty)$.

3. Random Variables and Probability Distributions

Random Variable. A random variable is a function defined on a sample space; more specifically, it is an assignment of a number to each point of the sample space (see Feller, p. 199). The set of numbers assigned is the set of *possible values* of the random variable. A unit amount of probability is spread out (distributed) over the set of possible values. This distribution is called the *probability distribution* of the random variable. The probability distribution can be specified by the *cumulative distribution function* or by the *frequency function.*

Cumulative Distribution Function of a Random Variable. The cumulative distribution function, say $F(x)$, of a (one-dimensional) random variable X is such that

$$F(x) = \Pr\ (X \leq x)\ \ (-\infty < x < +\infty). \tag{9}$$

(The notation "Pr $(X \leq x)$" means "the probability that X is less than or equal to x".) $F(x)$ is also termed simply the *distribution function* of X.

The Frequency Function of a Random Variable. When the possible values of a random variable X form a discrete set, say x_1, x_2, etc., the function

$$\Pr\ (X = x_j) = f(x_j), \text{ say, } (j = 1, 2, \ldots) \tag{10}$$

is called the frequency function of X. X is called a *discrete* random variable. When the cumulative distribution function $F(x)$ of X is differentiable, X is called a continuous random variable. The function

$$\frac{dF(x)}{dx} = f(x), \text{ say,} \tag{11}$$

is called the *probability density function* of X. $f(x)$ is also called the *frequency function* of X. Hoel (p. 24) indicates that it is becoming more and more common to use the term *frequency function* for both continuous and discrete variables. The *expected value* (also called the *mean*) of a discrete variable X is defined as

$$x_1 f(x_1) + x_2 f(x_2) + \ldots = m_x, \text{ say,} \tag{12}$$

where $f(x_j)$ is the frequency function of X ($j=1, 2, \ldots$). m_x is also called the mean of the distribution of X.

The *variance* of X is defined as

$$(x_1-m_x)^2 f(x_1) + (x_2-m_x)^2 f(x_2) + \ldots = \sigma_x^2, \text{ say.} \qquad (13)$$

The *standard deviation* of X is defined as the (non-negative) square root of the variance. If X is a continuous random variable, the definitions of its expected value (mean), variance, and standard deviation are similar, respectively, to the definitions above. (It should be remarked, however, that some random variables do not have a mean or variance.)

Let (a, b) be the interval of smallest length containing all possible values of X (and assume that (a, b) is a finite interval). The length, $b-a$, is called the *range* of X. (Sometimes the interval (a, b) is called the range of X.) It is of interest that

$$\sigma_x \leq \frac{(b-a)}{2} \qquad (14)$$

Percent Points of a Random Variable. Let $F(x)$ be the cumulative distribution function of a random variable X, and suppose that x_a is a possible value of X such that

$$F(x_a) = a \ (0 \leq a \leq 1). \qquad (15)$$

The quantity x_a is termed the $100a$ *percent point* of X.* For example, the *median* of X is the number x_a when $a=0.50$; thus the median of X is a number such that exactly half the population is less than or equal to the number and, of course, exactly half exceeds the number. Some other well-known examples of percent points are $x_{.25}$ and $x_{.75}$ which are called the *lower* and *upper* quartiles, respectively. When $F(x)$ is continuous and increasing at every possible value of X, there is a unique x_a for any given a.

The "Center" and "Spread" of a Distribution. In a sense the median of a distribution can be regarded as the "center" of the distribution and the range can be regarded as the "spread" of the distribution. In another sense the mean and standard deviation can be regarded

*x_a is also termed the $100a$ percent point of the distribution of X.

as the "center" and "spread," respectively. The Bienaymé-Tchebycheff inequality indicates how the mean and standard deviation represent the "center" and "spread." This inequality is as follows (see Cramér p. 183):

$$\Pr\left(-B\sigma_x < X - m_x < B\sigma_x\right) \geq 1 - \frac{1}{B^2}, \quad (B \geq 1), \tag{16}$$

where m_x and σ_x are, respectively, the mean and standard deviation of X.* For example, it follows from formula (16) that the probability is at least 0.75 that X will be within *two* standard deviations of its mean. (Note that $1-1/B^2=0.75$ when $B=2$.) The following table gives values of the probability bound, $1-1/B^2$, for several values of B:

B	1.0	1.5	2.0	2.5	3.0
$1-1/B^2$	0.000	0.556	0.750	0.840	0.889

It is interesting to compare the values of the bound in formula (16) with the exact values of $\Pr\left(-B\sigma_x<X-m_x<B\sigma_x\right)$ when X has a normal distribution. The exact values, obtained by means of Appendix Table 1, are as follows (to three decimal places):

B	1.0	1.5	2.0	2.5	3.0
$\Pr\left(-B\sigma_x<X-m_x<B\sigma_x\right)$	0.683	0.866	0.954	0.988	0.997

4. Some Important Probability Distributions

The Binomial Distribution. Let p be the probability that an element selected at random from a binomial population will be a "success" (see Section 2.2). Let X be the number of "successes" in n independent random selections of an element from the population.

The frequency function of X is

$$f(x) = C_x^n p^x (1-p)^{n-x} \quad (x=0, 1, \ldots, n). \tag{17}$$

(Note: $C_x^n=n!/[x!\,(n-x)!]$, $x=0, 1, \ldots, n$.) The mean and variance of X are np and $np(1-p)$, respectively.

*It should be remarked that formula (16) is valid for any random variable having a finite mean and variance.

The Poisson Distribution. Let X have the Poisson distribution. The frequency function of X is

$$f(x) = \frac{e^{-m} m^x}{x!}, \quad (x=0, 1, 2, \ldots) \tag{18}$$

where m is the expected value (mean) of X. (Note: $e=2.71828\ldots.$)

The Normal Density Function. Let X have the normal distribution. The probability density function associated with this distribution is

$$f(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \quad e^{-(x-u)^2/2\sigma^2} \quad (-\infty < x < +\infty), \tag{19}$$

where u is the expected value of X and σ^2 is the variance of X.

The standard normal density function, say $g(x)$, is

$$g(x) = \frac{1}{\sqrt{(2\pi)}} \, e^{-x^2/2} \tag{20}$$

which is a special case of $f(x)$, arising when $u=0$ and $\sigma=1$.

The Hypergeometric Distribution. Consider a *finite* binomial population having $\mathcal{N}$ elements of which $\mathcal{N}_1$ are "successes" and $\mathcal{N}-\mathcal{N}_1$ are "failures." The population proportion of "successes" is $\mathcal{N}_1/\mathcal{N}$. This proportion can be regarded as the parameter of the population since $\mathcal{N}$ is assumed to be known. Suppose a sample of size n is drawn *without replacement,* and let X be the number of successes in the sample. The frequency function of X is

$$f(x) = \frac{C_x^{\mathcal{N}_1} \, C_{n-x}^{\mathcal{N}-\mathcal{N}_1}}{C_n^{\mathcal{N}}}, \tag{21}$$

where $n \leqq \mathcal{N}$ and $x=h, \ldots, H$ ($h=$maximum of 0 and $\mathcal{N}_1+n-\mathcal{N}$; $H=$minimum of $\mathcal{N}_1$ and n). This is called the hypergeometric distribution. The expected value of X is $n\mathcal{N}_1/\mathcal{N}$ and the variance of X is $n(\mathcal{N}_1/\mathcal{N})\,[(\mathcal{N}-\mathcal{N}_1)/\mathcal{N}]\,[(\mathcal{N}-n)/(\mathcal{N}-1)]$. The quantity X/n is an unbiassed estimator of the parameter $\mathcal{N}_1/\mathcal{N}$.

It should be remarked that if the sample is drawn *with replacement,* the probability distribution of the number, X, of successes is

$$f(x) = C_x^n \, p^x \, (1-p)^{n-x}, \tag{22}$$

where $p=N_1/N$ and $x=0, \ldots, n$. This is a binomial distribution.

The Multinomial Distribution. Consider a population having exactly k kinds of elements. Let p_i be the probability that a randomly drawn element is of the ith kind. (Of course, $p_1+p_2+\ldots+p_k=1$.) Suppose a sample of size n is drawn from the population, and let $N_1, N_2, \ldots, N_k$ be the numbers of elements of the first, second, $\ldots$, kth kinds, respectively, in the sample ($N_1+N_2+\ldots+N_k=n$). It can be shown* that

$$\Pr(N_1=n_1, N_2=n_2, \ldots, N_k=n_k) = \left(\frac{n!}{n_1!\ldots n_k!}\right) p_1^{n_1} \ldots p_k^{n_k}, \quad (n_1+n_2+\ldots+n_k=n). \tag{23}$$

This is the so-called *multinomial distribution.* When $k=2$, the multinomial distribution reduces to the binomial distribution.

The Uniform Distribution. Let X have the uniform distribution. The probability density function associated with this distribution is

$$f(x)=1, \quad (0\leq x\leq 1). \tag{24}$$

The cumulative distribution function of X is

$$F(x)=x, \quad (0\leq x\leq 1). \tag{25}$$

The mean and variance of X are 1/2 and 1/12, respectively.

*If the population is finite, the sampling is assumed to be *with replacement.*

Appendix Tables

Appendix Tables

COMMENTS REGARDING TABLES

Appendix Table 1. The Cumulative Standard Normal Distribution

This is a table of the function

$$F(x) = \int_{-\infty}^{x} g(t)dt,$$

where $g(t)$ is the standard normal density function given in formula (20) in the Appendix. It should be noted that $F(x) = \Pr(X \leq x)$, where X has the standard normal distribution. The values of x in the table are 0, 0.01, ..., 3.49. To obtain values of $F(-x)$ use the equation $F(-x) = 1 - F(x)$. Beneath the main part of Appendix Table 1 is a supplementary table giving values of x for $F(x) = 0.90$, 0.95, 0.975, etc. This supplementary table is similar to Table V in Chapter 3. (Table V is used in many examples in Chapter 3 (for instance, see the examples in 3.2.b)).

Appendix Table 2. The Cumulative Chi-square Distribution

This is a table of values of χ^2 (Chi-square) for various values of the cumulative χ^2–distribution, say $F(\chi^2, n)$. n is called the number of "degrees of freedom" of the distribution. The values of $F(\chi^2, n)$ given in the table are 0.005, 0.010, 0.025, 0.050, 0.100, 0.250, 0.500, 0.750, 0.900, 0.950, 0.975, 0.990, and 0.995. When n is larger than 30,

$$\chi_a^2 \doteq \frac{1}{2}[z_a + \sqrt{(2n-1)}]^2,$$

where χ_a^2 is the $100a$ percent point of the Chi-square distribution and z_a is the $100a$ percent point of the standard normal distribution. For illustrations of the use of the Chi-square distribution see examples in 4.3.b.

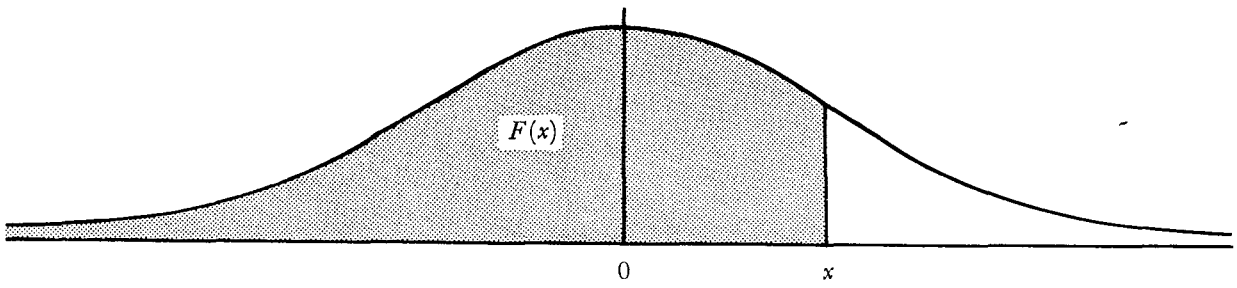
Appendix Table 3. The Cumulative Student's *t*-Distribution

This is a table of values of t for various values of the cumulative Student's t-distribution, say $F(t, n)$. n is called the number of "degrees of freedom" of the distribution. The values of $F(t, n)$ given in the table are 0.75, 0.90, 0.95, 0.975, 0.990, 0.995, and 0.9995. For $n=\infty$, the Student's t-distribution is the standard normal distribution. This accounts for the fact that most of the entries in the last row of Appendix Table 3 also appear in the supplementary table beneath the main part of Appendix Table 1. For an illustration of the use of Student's t-distribution see the first example in 3.4.a.

Appendix Table 4. 2000 Random Digits

This is a table of independently observed values of a random digit. Section 5.2 gives a definition of a random digit and illustrations of how to use Appendix Table 4.

Appendix Table 1—The Cumulative Standard Normal Distribution*



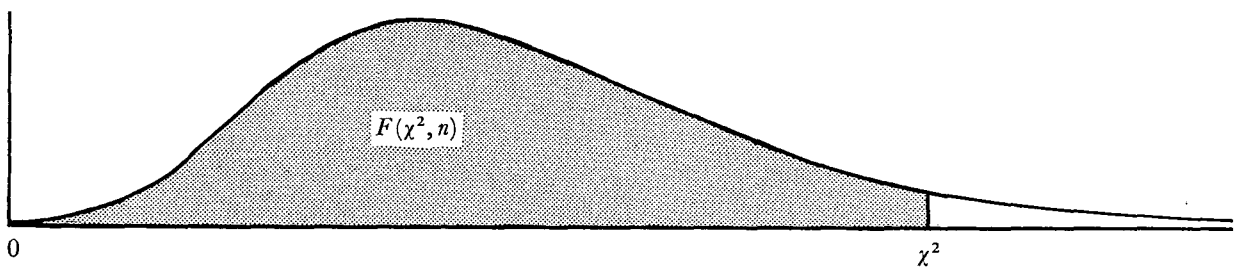
x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319

1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

x	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417
$F(x)$	0.90	0.95	0.975	0.99	0.995	0.999	0.9995	0.99995	0.999995
$2[1-F(x)]$	0.20	0.10	0.05	0.02	0.01	0.002	0.001	0.0001	0.00001

*This table is reproduced from *Introduction to the Theory of Statistics* by A. M. Mood, McGraw-Hill, New York, 1950. It is published here with the kind permission of the author and publishers.

Appendix Table 2—The Cumulative Chi-Square Distribution*

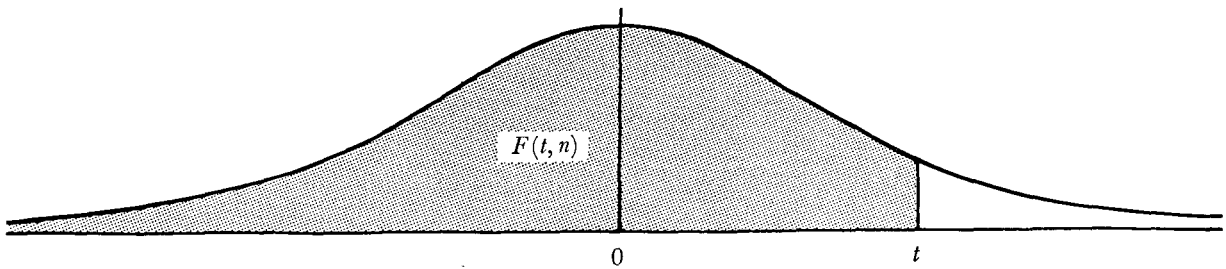


n/F	.005	.010	.025	.050	.100	.250	.500	.750	.900	.950	.975	.990	.995
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	0.0158	0.102	0.455	1.32	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	0.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	0.676	.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2

11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7

*This table is abridged from "Table of Percentage Points of the χ^2 Distribution," *Biometrika*, Vol. 32, Part II (1941), pp. 187–191. It is published here with the kind permission of the author, Catherine M. Thompson, and the editor of *Biometrika*.

Appendix Table 3—The Cumulative Student's t-Distribution*



n/F	.75	.90	.95	.975	.99	.995	.9995
1	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	1.638	2.353	3.182	4.541	5.841	12.941
4	0.741	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	6.859
6	0.718	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	5.405
8	0.706	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	4.073

16	0.690	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.767
24	0.685	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	1.303	1.684	2.021	2.423	2.704	3.551
60	0.679	1.296	1.671	2.000	2.390	2.660	3.460
120	0.677	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.291

*This table is abridged from *Statistical Tables for Biological, Agricultural, and Medical Research*, by R. A. Fisher and Frank Yates, Oliver and Boyd, Edinburgh, 1957 (5th edition). It is published here with the kind permission of the authors and publishers.

Appendix Table 4—2000 Random Digits*

Rows	Digits									
1	49269	27212	46095	37106	64254	27460	49572	51700	27679	12574
2	33891	03867	09925	06476	82018	45094	59014	67113	44192	00075
3	23318	79895	70550	81717	28833	30271	15821	14999	88174	62617
4	57517	55256	50281	51583	96879	05225	42272	05339	20483	57596
5	41011	75937	22767	50120	95938	49753	63882	99616	69083	38721
6	73889	80236	99531	23053	71237	48861	59046	76283	60538	19732
7	93877	30345	64882	66660	17026	70364	45676	08039	96228	89936
8	59141	95585	89552	97247	59325	27848	80058	15950	61481	90906
9	40998	44137	16144	66300	44091	50018	81364	18211	60294	76559
10	20279	27414	10589	39860	23000	31767	95618	56738	50332	16936
11	70342	92481	30702	76264	62619	68678	62284	83112	93032	55203
12	52614	36950	41796	45403	79262	02887	53596	61308	20738	34811
13	27099	90956	65448	03080	75795	29753	97699	80872	23830	85882
14	74427	99523	74904	28017	45898	57232	48525	07086	26805	74533
15	92470	18840	76011	93109	14344	55614	50284	15865	19458	35856
16	13464	53679	64603	51571	56124	79107	29596	89572	78198	57121
17	73649	08804	87977	87959	70859	40909	77295	87877	75158	62810
18	92074	23244	59516	50552	31602	41899	06347	27821	68370	48596
19	88577	30231	25267	84622	31449	12086	56461	22962	78213	62483
20	93966	60437	52239	58113	32526	38708	81607	57016	01695	90110

21	04649	59990	23979	03855	10297	46516	96092	82305	30760	78756
22	04967	82876	04773	86651	16648	53133	82439	78851	49766	24553
23	15273	36417	01901	33386	76979	25920	33372	02695	11982	40911
24	06230	91696	43907	17827	30332	89203	32215	91806	23080	49102
25	09174	11548	54590	75803	66108	73882	62324	26017	72716	33887
26	01285	31604	71039	24337	53514	58964	89901	22040	92751	12617
27	37007	05523	61672	62557	98540	26094	60284	19621	96230	38044
28	06545	09458	42988	02913	86345	67936	90174	40840	44991	24256
29	34989	74086	13652	68706	01363	04294	88008	78693	83068	94746
30	00221	89299	53186	05930	61889	51341	45412	58860	72568	11381
31	59785	36887	10690	31347	93326	96267	86987	57565	86836	49071
32	90331	41248	34629	30240	27270	03864	84308	03035	61369	36902
33	51017	44409	17120	23823	36460	63359	08333	63173	19134	06493
34	00303	18550	26191	19051	81502	66343	06737	90430	65478	58982
35	82484	16483	47704	44640	68322	44548	72787	02335	28749	39320
36	05436	98146	56596	00812	51445	35533	35478	47573	38414	25542
37	38032	13442	42983	97207	77854	57806	81616	52828	79429	47389
38	96795	57764	19605	24767	63253	18809	65093	44449	22952	76872
39	30983	38948	09310	48336	87651	27110	84427	76209	56412	12760
40	16747	14551	82626	31224	98636	75100	84882	79479	83420	05347

*This table is a part of page 13 of *A Million Random Digits with 100,000 Normal Deviates* by The Rand Corporation, The Free Press, Glencoe, Illinois, 1955. It is reproduced here with the kind permission of The Rand Corporation and the publishers.

Bibliography

Adams, Warren T. "Five-Minute-Cluster Sampling for Determining Urban Traffic Volumes," *Proceedings of the Highway Research Board*, Washington, D.C., 1955.

Burch, James. "Total Travel in North Carolina Municipalities," *Proceedings of the Highway Research Board*, Washington, D.C. 1951.

Bureau of Highways, Department of Public Roads, Commonwealth of Puerto Rico, "Application of the Vickery Short Count Method to Urban Traffic Volume Counts in Puerto Rico," 1951 (unpublished).

Clopper, C. J. and Pearson, E. S. "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, Vol. 26 (1934), pp. 404–413.

Cochran, W. G. *Sampling Techniques*. John Wiley and Sons, New York, 1953.

Covault, D. O. "The Use of Sample Survey Methods to Estimate Highway Needs," *Traffic Quarterly*, April 1960, pp. 248–273.

Cramér, H. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, New Jersey, 1946.

Deming, W. E. *Some Theory of Sampling*. John Wiley and Sons, New York, 1950.

Dixon, W. J. and Massey, F. J. *Introduction to Statistical Analysis*. McGraw-Hill, New York, 1951.

Duncan, A. J. *Quality Control and Industrial Statistics*. Richard D. Irwin, Inc., Homewood, Illinois, 1959 (revised edition).

Feller, W. *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, New York, Vol. 1, 1957 (2nd edition).

Fisher, R. A. and Yates, Frank. *Statistical Tables for Biological, Agricultural, and Medical Research*. Oliver and Boyd, Edinburgh, 1957 (5th edition).

Gerlough, Daniel L. "Use of Poisson Distribution in Highway Traffic," in *Poisson and Traffic*, The Eno Foundation for Highway Traffic Control, Saugatuck, Connecticut, 1955.

Greenshields, Bruce D., Ericksen, Elroy, and Schapiro, Donald. "Traffic Performance at Urban Intersections," Technical Report No. 1, Yale Bureau of Highway Traffic, Yale University, New Haven, Connecticut, 1947.

Greenshields, Bruce D. and Weida, Frank M. *Statistics with Applications to Highway Traffic Analysis*. The Eno Foundation for Highway Traffic Control, Saugatuck, Connecticut, 1952.

Hald, A. *Statistical Tables and Formulas*. John Wiley and Sons, New York, 1952.

Hald, A. *Statistical Theory with Engineering Applications.* John Wiley and Sons, New York, 1952.

Harvard University Computation Laboratory. *Tables of the Cumulative Binomial Probability Distribution.* Harvard University Press, Cambridge, 1955.

Hoel, P. G. *Introduction to Mathematical Statistics.* John Wiley and Sons, New York, 1954 (2nd edition).

McCarthy, P. J. *Introduction to Statistical Reasoning.* McGraw-Hill, New York, 1957.

McClintock, M. *Short Count Traffic Surveys and Their Application to Highway Design.* Portland Cement Association, Chicago, November 1935 (2nd and revised edition).

Molina, E. C. *Poisson's Exponential Binomial Limit.* D. Van Nostrand, New York, 1942.

Mood, A. M. *Introduction to the Theory of Statistics.* McGraw-Hill, New York, 1950.

Pachares, J. "Tables of Confidence Limits for the Binomial Distribution," *Journal of the American Statistical Association,* Vol. 55 (1960), pp. 521–533.

Rand Corporation, The. *A Million Random Digits With 100,000 Normal Deviates.* The Free Press, Glencoe, Illinois, 1955.

Ricker, W. E. "The Concept of Confidence or Fiducial Limits Applied to the Poisson Frequency Distribution," *Journal of the American Statistical Association,* Vol. 32 (1937), pp. 349–356.

Smith, Wilbur and Associates. *Trade, Transit, and Traffic—Providence, Rhode Island,* New Haven, Conn., October 1957.

Sukhatme, P. V. *Sampling Theory of Surveys with Applications.* The Iowa State College Press, Ames, Iowa, 1954.

Thompson, Catherine M. "Table of Percentage Points of the χ^2 Distribution," *Biometrika,* Vol. 32, Part II (1941), pp. 187–191.

Vickery, C. W. "On Estimating Street Traffic by Means of Extremely Short Counts," State Highway Department, Austin, Texas, July 12, 1939 (unpublished).

Wallis, W. Allen and Roberts, Harry V. *Statistics, A New Approach.* The Free Press, Glencoe, Illinois, 1956.

Wardrop, J. G. "Some Theoretical Aspects of Road Traffic Research," Road Paper No. 36, Institution of Civil Engineers, London, 1952.

White, Francis and Pelz, V. H. "Short Count Results in a Traffic Survey Made in Fort Wayne, Indiana," *Traffic Engineering,* August 1948, pp. 499–511.

Wilks, S. S. *Elementary Statistical Analysis.* Princeton University Press, Princeton, New Jersey, 1949.

Author Index

Subject Index