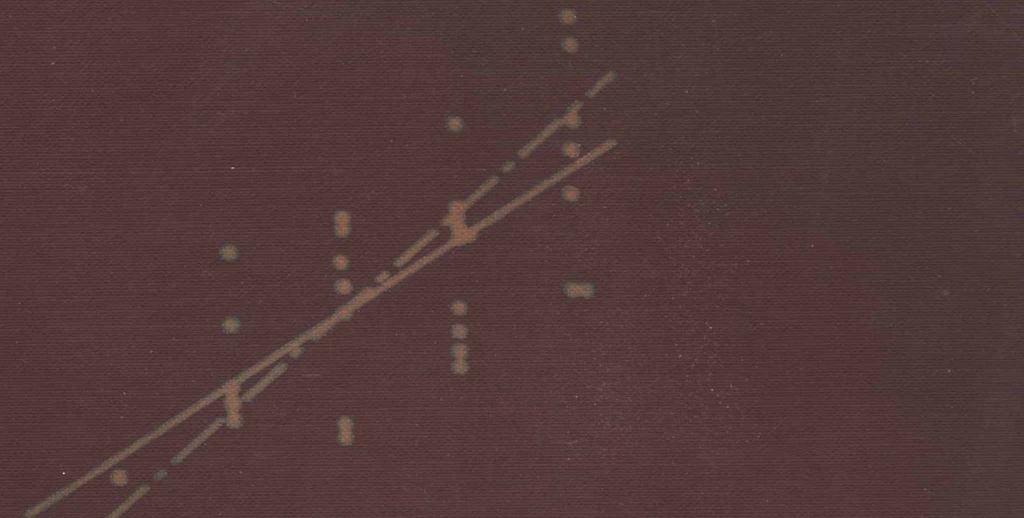
Concepts and Methods of EXPERIMENTAL STATISTICS



H. C. FRYER

Concepts and Methods of

Experimental Statistics

H. C. Fryer

Kansas State University

Allyn and Bacon, Inc.
Boston, 1966

Library of Congress Catalog Card Number: 66-15488

Printed in the United States of America

Preface

Increasingly large numbers of people are learning more about the area of applied statistics. One of the most popular of these areas can be described as experimental statistics, because it deals with the concepts and methods of statistical analyses needed when experimental research culminates in the taking of numerical measurements.

Many persons doing experimental research or preparing to do such research must attain a working understanding of the basic concepts in experimental statistics with a minimum of prerequisite work in mathematics and statistics, for they often have not taken two or three one-year sequences in undergraduate mathematics and statistics while seeking an advanced degree in another field.

Over the past twenty-five years a number of books have been written in the general area of statistical methods. Some of these books do not cover enough of the topics which one wishes to learn. Others cover most of the topics one wishes to learn but are written chiefly as reference books or handbooks seeking to give directions for many statistical operations. It seems to me that at the present time there is need for a book on statistical methods which is adequately comprehensive and also primarily a teaching instrument designed to lead the nonmathematical persons into as thorough an understanding of the basic concepts and reasoning of experimental statistics as is possible in what is essentially an introduction for them. After teaching this subject over a period of roughly twenty-five years I have attempted to develop this textbook in experimental statistics for advanced undergraduates and beginning graduate students in the several areas of experimental research. It is not slanted intentionally toward any particular field of applied statistics.

This attempt to present a teaching book in experimental statistics to nonmathematical but mature students necessarily calls upon the intuition, common sense, and ingenuity of the student to bridge some gaps best filled by mathematical reasoning. However, the student is expected to have or to attain some facility with algebra, with summation symbols, and even with a few elementary matrix operations.

Chapter 1 is intended to be more than a cursory introduction to the subject of this book. It is hoped that here the student will gain an insight into the problems, concepts, and procedures of experimental statistics and so acquire a good background for the other chapters, which present in more detail what already has been touched upon in Chapter 1. Certainly the student will not fully understand what is covered in Chapter 1 when he has finished this chapter, but it is hoped, and believed from past experience, that he will be better able to learn the subject of this book than if he were plunged immediately into detailed discussions of sampling, statistical methods, and statistical interpretations. It will be necessary for an instructor using this book as a text to guard against spending too much time on Chapter 1. It is the aim of that chapter to bring the student to as full an understanding as possible of the points made in Section 1.8, the summary, before that student plunges into a deeper and more detailed study of experimental statistics. It is suggested that Sections 1.1, 1.2, and 1.3 be assigned as a group, Sections 1.4 and 1.5 as another, Section 1.6 by itself, and Section 1.7 by itself, Section 1.8 being made the basis of a final summary of Chapter 1. As I see it, the instructor's discussions should be general and unifying, and he should spend a good deal of the classroom time on the problems at the ends of the sections in Chapter 1, with as much class participation as possible. No more than six or seven class periods, including an "idea" examination, should be spent on this chapter.

This book's precursors were used in classes in statistical methods for several years by members of the Department of Statistics at Kansas State University. This use was the basis of several revisions. The classes in which this material was used comprised mostly graduate students from at least twenty-five different departments in agriculture, arts and sciences, commerce, engineering, home economics, and veterinary medicine. This broad background seems to justify the inclusion of problems and illustrations from a wide variety of experimental situations. It also seems to me to justify a sincere attempt to lead students to a basic understanding of the subject rather than to a mere facility with statistical methods, for these students will be called upon later in their professional lives to apply experimental statistics to a very wide range of problems, some not even specifically conceivable at this time.

Anyone writing a textbook on this subject inevitably will be influenced by Snedecor's classic books and also by several other textbooks, reference books, and handbooks in the general area of statistical methods. There is, further, abundant literature on statistical methods in the many fields of experimental research. When these influences are overlain by more than twenty-five years of teaching of experimental statistics to nonmathematical students, it becomes impossible to acknowledge specific debts to specific writers and teachers with whom I have had contact. I do feel a special debt to Professor George W. Snedecor, who was one of my teachers at Iowa State University and whose books we have used at Kansas State University for many years. I am glad, however, to acquit him of all blame for the inadequacies of the present textbook. I am indebted also to the students in my classes and to my colleagues in the Department of Statistics at Kansas State University for calling my attention to errors, to misprints, and to unclear presentations in earlier versions of this book.

I am indebted to the late Sir Ronald A. Fisher, F.R.S., Cambridge, and Dr. Frank Yates, F.R.S., Rothamsted, and to Messrs. Oliver and Boyd Ltd., Edinburgh, for permission to reprint parts of Table XII from their book, *Statistical Tables for Biological, Agricultural and Medical Research.*

Finally, I should like to thank all other persons and organizations who have granted permission to use parts of their publications. These permissions are acknowledged in the appropriate places.

H. C. FRYER
Manhattan, Kansas

Contents

Chapter 1

Introduction

Chapter 2

Sampling Binomial and Multinomial Populations

Chapter 3

Sampling One Normal Population, $N(\mu, \sigma^2)$

Chapter 4

Sampling Two Normal Populations, $N(\mu_1, \sigma_1^2)$ *and* $N(\mu_2, \sigma_2^2)$

Chapter 5

Sampling Populations with Unspecified Distributions

Chapter 6

Linear Regression and Correlation

Chapter 7

Sampling More Than Two Normal Populations Simultaneously: Introductory Analysis of Variance

Chapter 8

Models and Expected Mean Squares for the Analysis of Variance for Completely Randomized and Randomized Complete Block Designs

Chapter 9

More Complex Analyses of Variance

Chapter 10
Simple Analysis of Covariance

Chapter 11
Multiple Linear Regression and Correlation Analysis

Chapter 12
Curvilinear Regression Analysis

Chapter 13
The Poisson and Negative Binomial Distributions

Chapter 14
Introductory Discriminatory Analysis

Tables

Chapter 1

Introduction

1.1 STATISTICAL POPULATIONS OF NUMERICAL MEASUREMENTS

For centuries human beings have described themselves, their environments, their possessions, their thoughts, and their actions by means of numerical measurements. This sort of description developed from the necessity to arrange efficiently our activities and to make more rapid the communication of ideas. For example, the density of habitation by man, plants, or animals, in a describable geographical or sociological area usually is measured by counting. Weights describe crops, people, soils, fish, or manufactured products. The economic rewards of labor, investment, professional services, or the production of consumable goods are measured numerically, such that highly complex transactions are carried out and described accurately and efficiently. The increasing availability of high-speed computers and data processors is indicative of the benefits to be derived from numerical description by calculation and the storage of information.

When a person makes numerical measurements of some objects or actions, he does not just measure a conglomeration of them; rather, he has at least previously formulated in his mind a reasonably homogeneous group that he wishes to measure in some specific respect. An economist does not measure the net incomes of a random group of people whom he meets during a tour or on his way from one part of a country to another; instead, he measures net incomes of lawyers in a certain area, of cattlemen in western Kansas, of grocers in Hawaii, etc. There is a definable homogeneity about any group which is measured for any specific characteristic. This homogeneity is the starting point for defining what will be called a *statistical population* or, more briefly herein, a *population.*

There is, however, associated with the concept of a homogeneous group that may be measured in a specific way the full awareness that *no* group is perfectly homogeneous: always there is some variability, even in the most closely knit group. If this were not so, there would be no reason at all for statistical concepts, and all experimental research would be exceedingly simple.

The group of actions, objects, or situations to which a set of numerical measurements pertains can be the basis of defining a statistical population. However, the same group of actions, objects, or situations also can be the basis of defining several different statistical populations. The inhabitants of Australia, for example, could be studied with respect to sex ratio, age, weight, blood sugar, political opinions, amount of education, quality of eyesight, and so forth, ad infinitum. Hence, any particular statistical population which is to be studied must be defined carefully, or considerable confusion may result and thus obscure the purposes of the study. The definition of the population must specify the group of objects or actions to be studied, the particular feature which is to be studied numerically, and the unit of measure to be employed; and these must be specified unambiguously. For example, all the legal residents of the State of Oregon on July 1, 1965, could be the basis of the following rather specific statistical populations, and many more:

(*a*) Their ages to the nearest tenth of a year.

(*b*) Their blood pH's in the usual unit of measure for this chemical characteristic.

(*c*) Their weights to the nearest tenth of a pound.

(*d*) Their pulse rates at 9:00 A.M. (PST).

(*e*) The gross annual incomes of all who were teachers in four-year colleges or universities and had been rehired in writing for the next school year.

(*f*) The political faiths of those at least 21 years of age, with Democrat = 1, Republican = 2, and Other = 3.

Note that each of these statistical populations is a collection of numerical measurements, not of people.

Once a statistical population has been defined unambiguously, it usually is of interest to highlight a few features of it by means of some summary numbers. If the objects, actions, or characteristics are merely to be classified in two or more categories and then the numbers in each class recorded, the data will be called *enumerative data.* For example, if insects are to be sprayed with some insecticide and the numbers Dead and Alive counted after a suitable waiting period, then probably the only summary number of interest would be the percent Dead, for this figure succinctly describes the effectiveness of an insecticide. If the percent Dead for the whole population involved is called p, then we say p is the *population parameter* which describes the potential effectiveness of that spray against the particular species of insect studied. This parameter, p, is sufficient to convey all the information wanted in this situation. For example, if one knows that a certain insecticide kills 95% of the houseflies that it hits, and if the lethal quality of the spray is the only matter of interest, then $p = 0.95$ gives this information fully. Put another way, one can expect that the chances are 95 out of 100 that a randomly chosen housefly sprayed with that particular spray will be killed. If some flies are killed, some are merely made moribund, and others are unaffected, one needs two parameters, namely p_1 = percent killed and p_2 = percent made moribund. Obviously, $100 - p_1 - p_2$ = percent unaffected by the spray. Now, to convey all the information needed about the population one parameter (p) is not sufficient, and two parameters (p_1 and p_2) are required. Such an enumerative population, or one with even more parameters, is called a *multinomial population*; a one-parameter population with p = percent killed is called a *binomial population.* More specific definitions of these two types of enumerative population will be given later.

If the objects or actions upon which a statistical population is based are described numerically by measuring something along a continuous scale of measurement, such as weights in ounces, prices in cents, yields of grain in bushels per acre, or breaking loads of concrete beams in pounds, each statistical population probably will be described by means of averages and measures of variability. Perhaps frequency distributions and graphs also will be employed. That is, one usually is concerned with the general level of the sizes of the numbers in the population and with their variability (or dispersion) around that general level of size. For example, intelligence quotients (I.Q.'s) of people are measured along an essentially continuous scale from 25 to 175, approximately. A psychologist might wish to know whether one group of people has, in general, a higher level of intelligence than another group. One way to express the level of intelligence is to compute an average I.Q. of some sort. The arithmetic mean and the median are two popular averages, but there are others, such as the midrange, the geometric mean, and the harmonic mean.

Even if the general level of intelligence in one group is higher than in another, as measured by I.Q.'s, there will be considerable variation within each group. Some individuals in the group with generally lower intelligence will have I.Q.'s higher than those of some members of the higher-intelligence group. Thus, any acceptable study of the levels of intelligence of two or more groups of people must consider not only apparent differences in general level of intelligence but also the variability within each group. Doing this correctly and efficiently is one of the purposes of statistical analysis.

One of the most common and useful measures of variability in a group of measurements taken on what is regarded as a homogeneous group of objects or actions is the *variance*, σ^2, or its square root, σ, which is called the *standard deviation*. These and other measures of variability will be considered in detail later.

The preceding statements may be illustrated by means of large sets of numerical measurements, which simulate actual statistical populations of numerical measurements. The population of Table 1.11 simulates some actual agronomic data taken during a study at Kansas State University. Several hundred differences in wheat yield between the Pawnee and Tenmarq varieties of wheat grown side by side on pairs of plots in a number of regions in Kansas were used as the starting point for the population. Thereafter, similar data were added in such quantity and in such a way that a population

TABLE 1.11. FREQUENCY DISTRIBUTION TABLE FOR A NEAR-N(5, 4) POPULATION

X interval	*Frequency of occurrence (f)*	*X interval*	*Frequency of occurrence (f)*
10.8 to 11.2	2	4.3 to 4.7	268
10.2 10.6	10	3.8 4.2	244
9.6 10.0	18	3.3 3.7	210
9.0 9.4	34	2.8 3.2	166
8.4 8.8	61	2.3 2.7	127
7.8 8.2	104	1.8 2.2	104
7.2 7.7	127	1.2 1.6	61
6.8 7.1	166	0.6 1.0	34
6.3 6.7	210	0.0 0.4	18
5.8 6.2	244	−0.6 −0.2	10
5.3 5.7	268	−1.2 −0.8	2
4.8 5.2	287		
		Total number in population:	2775

of 2775 measurements was produced, which closely conforms to what is called a *normal population* with a mean of $\mu = 5$ and a variance $\sigma^2 = 4$. Such a population is designated by the symbol N(5, 4). Obviously, this mass of 2775 numbers would be quite incomprehensible until it were summarized in some effective way. One helpful summary is obtained by grouping those 2775 numbers into relatively few classes of numbers and essentially considering all the numbers within any class as being at the midpoint of that class.

With a relatively large set of data one usually tries to form about twenty classes whose widths are about one fourth of the size of the standard deviation. When the numbers in a population are symmetrically arranged with respect to the mean, μ, as is true of any normal population, that μ should be at the midpoint of one of the classes, or there will be some distortion of the distribution curve. Minor adjustments in the lengths of the class intervals used may be advisable, if indicated by some trial-and-error preliminary groupings of the data. In the case of the 2775 numbers described above, the standard deviation is $\sqrt{\sigma^2} = \sqrt{4} = 2$, and one fourth of 2 is 0.5; hence this length of interval was tried with $X = 5$ at the midpoint of one interval. It appeared upon actual trial that an excessively large number of classes would result and that some of the end classes would have very few members of the population in them. For this reason a few of the classes at the extremes of the distribution were increased to a length of 0.6 instead of 0.5. Table 1.11 then was constructed by tallying the individual 2775 numbers into their proper classes and counting them. The frequencies, f's, are these counts for each of the classes.

The following features of the population can be noted in the table rather easily:

(*a*) The highest frequency of occurrence of measurements (X's) comes for the X interval which includes the true mean, $\mu = 5$.

(*b*) The class frequencies are smaller and smaller for measurements farther and farther from the population mean, and this decrease is symmetrical with respect to that mean.

(*c*) A high percent (near 95) of the whole population consists of numbers within two standard deviations ($=4$) of the mean ($=5$), and about two thirds of the members of the population are within one standard deviation ($=2$) of the mean, either way.

Thus, one can say that if a pair of adjoining plots of wheat, one Pawnee and the other Tenmarq, is chosen at random, the most probable difference in yield is $\mu = 5$ bushels per acre favoring Pawnee. Furthermore, large differences (and small differences) equally far from the most probable value, μ, are equally likely to occur. The likelihood of occurrence of such differences is closely related to the standard deviation, σ, in the normal population. This information is deducible either from the frequency distribution in Table 1.11 or from its graph in Figure 1.11. The graph was obtained by plotting the class frequencies over the midpoints of the class intervals, because all X's in an interval are considered, for computational purposes, to be at the midpoints of these intervals. Thereafter a smooth curve was drawn to fit these points as well as possible. This is approximately a normal curve with $\mu = 5$ and $\sigma^2 = 4$, so it is designated a near-N(5, 4) population.

The table and figure display, in general, three interesting features of this population of numbers, namely, its general form, its general level of magnitude, and the dispersion (or variation) of the individual measurements from the general level for the whole group. In other words, the Pawnee variety generally outyielded the Tenmarq variety by about 5 bushels per acre, but the differences in yield varied symmetrically by more than 6 bushels per acre in each direction from the mean, $\mu = 5$. Thus, the *form*, *level*, and *dispersion* of a statistical population are generally of interest, and are displayed to some degree by the frequency distribution table and its graph.

To emphasize the above statements additionally, consider the sketches in Figure 1.12 of other frequency distributions of populations.

The curve A indicates a preponderance of relatively small measurements but also the existence of a few relatively very large measurements. Some salary distributions are of this sort, as are insect counts under certain circumstances.

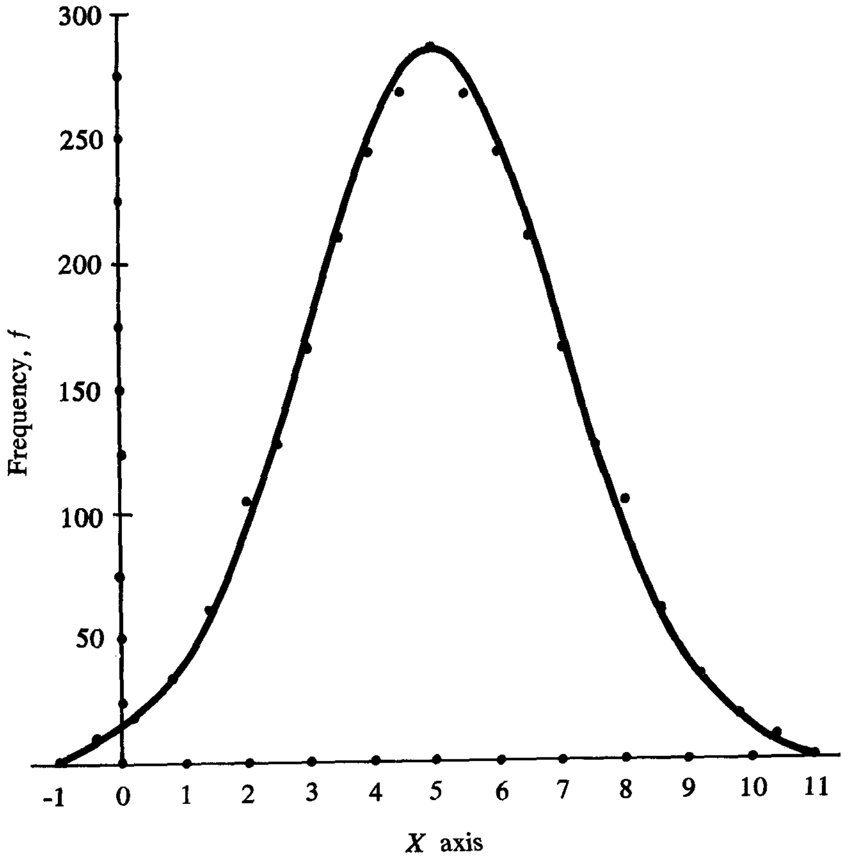


Figure 1.11. Frequency distribution for the population of Table 1.11.

The curve B is basically similar to A but has a preponderance of large numbers. Both A and B are decidedly nonsymmetrical in general form, but the *level* of measurement is much higher in B than in A. As far as these sketches show, there is approximately the same *dispersion* about the respective means.