# Principles and Procedures of STATISTICS

ROBERT G. D. STEEL
JAMES H. TORRIE

# Principles and Procedures of STATISTICS

WITH SPECIAL REFERENCE TO THE BIOLOGICAL SCIENCES

Robert G. D. Steel

*Professor of Statistics*
*North Carolina State College*

James H. Torrie

*Professor of Agronomy*
*University of Wisconsin*

McGRAW-HILL BOOK COMPANY, INC.

New York Toronto London

1960

# PREFACE

A tentative outline for this text was prepared by the authors while both were on staff at the University of Wisconsin. Discussions with faculty and students had led to the conclusion that three major considerations are involved in teaching statistics at the graduate level in colleges of agriculture.

There must be comprehensive coverage of statistical principles and procedures in a limited time. Most students can schedule only one or two terms of statistics since statistics is a supporting course except for the few students majoring or minoring in the subject.

The analysis of variance should be presented as quickly as feasible. The beginning graduate student is generally without statistical training yet assistantship duties often involve field, greenhouse, barn, or laboratory experiments which result in measurement data demanding analysis of variance computations. Also, graduate problems may require knowledge of experimental design and the analysis of variance before the student has acquired such knowledge.

Finally, a nonmathematical approach is desirable. The teacher of applied statistics often finds an inadequacy in and fear of mathematics on the part of the student. Hence we decided to avoid algebraic manipulations. Whether or not we have been completely successful will depend largely on how the student feels about our handling of the binomial distribution in Chap. 20.

Statisticians, like mathematicians, use subscripts as an alternative to lengthy descriptions for recurring and important situations. There seems to be little point in avoiding this convenient notation and subscripts are used with two summation notations, namely, $\Sigma$ and a "dot notation." Greek and English alphabets serve to distinguish between parameters and statistics. The Greek alphabet is given in the Appendix.

In teaching our classes, we have three main aims:

First, to present the student with a considerable number of statistical techniques, applicable and useful in terms of research. A relatively small survey of graduates of a biometry course suggests that the extent to which any statistical method is used depends on how intensively it was treated during the course. Few methods are learned later.

Second, to promote disciplined thinking on the part of the student with respect to the conduct of experiments. For example, the student should be trained to the point where he is able to appreciate that parts of the Salk vaccine trial were not relevant to the intent of the trial; that an experiment is not necessary to determine the relative precision with which a mean is measured for different sample sizes; that charts, averages, ratios, percentages, and nonrandom sampling can be and often are grossly misused.

Third, to promote disciplined thinking on the part of the student with respect to the analysis of experimental data. For example, a student should be able to appraise figures critically, appreciating their fallibility and limitations in terms of natural variation and its effects.

We began teaching from our tentative outline, eliminating inadequacies and inconsistencies as best we could and rearranging and adding material as the result of our teaching experiences. We paid considerable attention to models. We added techniques as they appeared in original sources, for example, those for testing among treatment means. We included some material of a reference nature when we felt it desirable for the student to know of its existence. Finally, realizing that no text can be completely adequate, up to date, and without error, we made a last rewriting and sent the manuscript to the publishers.

Our text generally follows the order of presentation in our courses. Considerable leeway in teaching methods and course content is possible.

In the first five chapters, statistical concepts are presented to provide a basis for understanding the many statistical methods given in the later chapters. These concepts include parent and derived distributions, probability, confidence intervals, tests of hypotheses, type I and II errors, power, sample size, and the analysis of variance. Since these concepts are usually new and difficult to the student, we have stressed them by repetition and elaboration.

In the chapters on statistical methods, we try to show the underlying logic as well. This is done, in part, by discussion of linear models for experimental designs and linear regression.

The placement of Chap. 6 was settled somewhat arbitrarily. The student is certainly not prepared to study it intensively at this point. Some instructors will wish to teach it later in the course while others will touch on it briefly before beginning the analysis of variance and then return to it after Chap. 12.

Discrete variables are discussed after the analysis of variance and covariance have been covered. The application of chi-square to problems involving discrete variables is treated with stress on the fact that chi-square is a continuous distribution and that, consequently, the test criteria here called chi-square are distributed only approximately as chi-square. The discussion of the binomial distribution was intentionally left until after two chapters which deal with discrete data, including binomially distributed data. This is done for two reasons. First, it is possible to overstress the binomial aspects of enumeration data at the expense of the nonparametric potential of the chi-square test criterion. Second, it seems possible to motivate a few more students to an interest in Chap. 20 after they have been exposed to enumeration data and the use of chi-square as a test criterion.

This text is, then, meant for the scientist or scientist-to-be with no special training in mathematics. It attempts to present basic concepts and methods of statistics and experimentation so as to show their general applicability. Most examples are chosen from biological and allied fields since that is where we are most familiar with applications. The order of presentation is such that the analysis of variance appears as early as is deemed practical.

The authors are indebted to Professor Sir Ronald A. Fisher, Cambridge, to Dr. Frank Yates, Rothamsted, and Messrs. Oliver and Boyd, Ltd., Edinburgh, for permission to reprint Table III from their book "Statistical Tables for Biological, Agricultural, and Medical Research."

The authors are also indebted to Fred Gruenberger and the Numerical Analysis Laboratory of the University of Wisconsin for preparing Table A.1; to E. S. Pearson and H. O. Hartley, editors of *Biometrika Tables for Statisticians*, vol. I, and to the *Biometrika* office for permission to reprint Tables A.2, A.6, A.8, and A.15; to C. M. Thompson and the *Biometrika* office for permission to reprint Table A.5; to D. B. Duncan and the editor of *Biometrics* for permission to reprint Table A.7; to C. W. Dunnett and the editor of the *Journal of the Americal Statistical Association* for permission to reprint Table A.9; to C. I. Bliss for permission to reprint Table A.10; to F. N. David and the *Biometrika* office for permission to reprint Table A.11; to L. M. Milne–Thomson and L. J. Comrie, authors of *Standard Four-figure Mathematical Tables*, and MacMillan and Co. Ltd., London, for permission to reprint Table A.12; to G. W. Snedecor, author of *Statistical Methods*, 4th edition, and the Iowa State College Press for permission to reprint Table A.13; to D. Mainland, L. Herrera, and M. I. Sutcliffe for permission to reprint Table A.14; to F. Mosteller and J. W. Tukey, the editor of the *Journal of the Americal Statistical Association*, and the Codex Book Company, Inc., for permission to reprint Table A.16; to Prasert Na Nagara for permission to reprint Table A.17; to Frank Wilcoxon and the American Cyanamid Company for permission to reprint Table A.18; to Colin White and the editor of *Biometrics* for permission to reprint Table A.19; to P. S. Olmstead, J. W. Tukey, the Bell Telephone Laboratories, and the editor of the *Annals of Mathematical Statistics* for permission to reprint Table A.20.

We are grateful to the many individuals, editors, and publishers who granted permission to use experimental data and other material.

For helpful comments on mimeographed notes, we wish to thank W. J. Drapala, W. T. Federer, and students in our classes.

Special thanks are due to Mrs. Jane Bamberg and Miss Helen Fuller for excellent work in typing the final draft of the manuscript.

*Robert G. D. Steel*
*James H. Torrie*

# CONTENTS

# Chapter 1

# INTRODUCTION

**1.1 Statistics defined.** Modern statistics provides research workers with knowledge. It is a young and exciting subject, a product of the twentieth century. For the scientist, particularly the biological scientist, statistics began about 1925 when Fisher's *Statistical Methods for Research Workers* appeared.

Statistics is a rapidly growing subject with much original material still not available in texts. It grows as statisticians find answers to more and more of the problems posed by research workers. Men who were among the earliest contributors to statistics are still productive and newcomers find diverse opportunities for their research talents. In the application of statistics, principles are general though techniques may differ, and the need for training in statistics grows as increased application is made in the biological and social sciences, engineering, and industry.

This young, vigorous subject affects every aspect of modern living. For example, statistical planning and evaluation of research have promoted technological advances in growing and processing food; statistical quality control of manufactured products makes automotive and electric equipment reliable; pollsters collect data to determine statistically the entertainment preferences of the public. More and more, the research team has, or has access to, a statistician and approximately 5 billion dollars was spent on research in the United States in 1957.

The extent of statistics makes it difficult to define. It was developed to deal with those problems where, for the individual observations, laws of cause and effect are not apparent to the observer and where an objective approach is needed. In such problems, there must always be some uncertainty about any inference based on a limited number of observations. Hence, for our purposes, a reasonably satisfactory definition would be: *Statistics is the science, pure and applied, of creating, developing, and applying techniques such that the uncertainty of inductive inferences may be evaluated.*

To most scientists, statistics is logic or common sense with a strong admixture of arithmetic procedures. The logic supplies the method by which data are to be collected and determines how extensive they are to be; the arithmetic, together with certain numerical tables, yields the material on which to base the inference and measure its uncertainty. The arithmetic is often routine, carried out on a desk calculator, and need involve no special mathematical training for the user. We shall not be directly concerned with mathematics although there is hardly an area of this subject which has not provided some usable theory to the statistician.

**1.2 Some history of statistics.** A history of statistics throws considerable light on the nature of twentieth century statistics. The historic perspective is also important in pointing to the needs and pressures which created it.

The term statistics is an old one and statistics must have started as state arithmetic since, to levy a tax or wage a war, a ruler needed to know the wealth and number of his subjects. Presumably all cultures that intentionally recorded history also recorded statistics. We know that Caesar Augustus sent out a decree that all the world should be taxed. Consequently he required that all persons report to the nearest statistician, in that day the tax collector. One result of this was that Jesus was born in Bethlehem rather than Nazareth. William the Conqueror ordered that a survey of the lands of England be made for purposes of taxation and military service. This was called the Domesday Book. Such statistics are history.

Several centuries after the Domesday Book, we find an application of empirical probability in ship insurance, which seems to have been available to Flemish shipping in the fourteenth century. This can have been little more than speculation or gambling but it developed into the very respectable form of statistics called insurance.

Gambling, in the form of games of chance, led to the theory of probability originated by Pascal and Fermat, about the middle of the seventeenth century, because of their interest in the gambling experiences of the Chevalier de Méré. To the statistician and the experimental scientist, the theory contains much of practical use for the processing of data.

The normal curve or normal curve of error has been very important in the development of statistics. The equation of this curve was first published in 1733 by de Moivre. De Moivre had no idea of applying his results to experimental observations and his paper remained unknown till Karl Pearson found it in a library in 1924. However, the same result was later developed by two mathematical astronomers, Laplace, 1749–1827, and Gauss, 1777–1855, independently of each other.

An essentially statistical argument was applied in the nineteenth century by Charles Lyell to a geological problem. In the period 1830–1833, there appeared three volumes of *Principles of Geology* by Lyell, who established the order among the Tertiary rocks and assigned names to them. With M. Deshayes, a French conchologist, he identified and listed fossil species occurring in one or more strata and also ascertained the proportions still living in certain parts of the seas. On the basis of these proportions he assigned the names Pleistocene (most recent), Pliocene (majority recent), Miocene (minority recent), and Eocene (dawn of the recent). Lyell's argument was essentially a statistical one. With the establishment and acceptance of the names, the method was almost immediately forgotten. There were no geological evolutionists to wonder if discrete steps, as implied by the names, were involved or if a continuous process was present and could be used to make predictions.

Other scientific discoveries of the nineteenth century were also made on a statistical basis with little appreciation of the statistical nature of the technique and with the method unfortunately soon forgotten. This is true in both the biological and physical sciences.

Charles Darwin, 1809–1882, biologist, received the second volume of Lyell's book while on the *Beagle*. Darwin formed his theories later and he may have been stimulated by his reading of this book. Darwin's work was largely biometrical or statistical in nature and he certainly renewed enthusiasm in biology. Mendel, too, with his studies of plant hybrids published in 1866, had a biometrical or statistical problem.

Thus in the nineteenth century, the need of a sounder basis for statistics became apparent. Karl Pearson, 1857–1936, initially a mathematical physicist, applied his mathematics to evolution as a result of the enthusiasm in biology engendered by Darwin. Pearson spent nearly half a century in serious statistical research. In addition, he founded the journal *Biometrika* and a school of statistics. The study of statistics gained impetus.

While Pearson was concerned with large samples, large-sample theory was proving somewhat inadequate for experimenters with necessarily small samples. Among these was W. S. Gosset, 1876–1937, a student of Karl Pearson and a scientist of the Guinness firm of brewers. Gosset's mathematics appears to have been insufficient to the task of finding exact distributions of the sample standard deviation, of the ratio of the sample mean to the sample standard deviation, and of the correlation coefficient, statistics with which he was particularly concerned. Consequently he resorted to drawing shuffled cards, computing, and compiling empirical frequency distributions. Papers on the results appeared in *Biometrika* in 1908 under the name of "Student," Gosset's pseudonym while with Guinness. Today "Student's $t$" is a basic tool of statisticians and experimenters, and "studentize" is a common expression in statistics. Now that the use of Student's $t$ distribution is so widespread, it is interesting to note that the German astronomer, Helmert, had obtained it mathematically as early as 1875.

R. A. Fisher, 1890–     , was influenced by Karl Pearson and Student and made numerous and important contributions to statistics. He and his students gave considerable impetus to the use of statistical procedures in many fields, particularly in agriculture, biology, and genetics.

J. Neyman, 1894–     , and E. S. Pearson, 1895–     , presented a theory of testing statistical hypotheses in 1936 and 1938 This theory promoted considerable research and many of the results are of practical use.

In this brief history, we shall mention only one other statistician, Abraham Wald, 1902–1950. His two books, *Sequential Analysis* and *Statistical Decision Functions*, concern great statistical achievements not treated in this text although one application is illustrated in Chap. 18.

It is in this century then, that most of the presently used statistical methods have been developed. The statistics of this text is a part of these methods.

**1.3 Statistics and the scientific method.** It is said that scientists use "scientific method." It would be difficult to define scientific method since the scientist uses any methods or means which he can conceive. However, most of these methods have essential features in common.

Without intending to promote a controversy, let us consider that these are:

1. A review of facts, theories, and proposals with a view to
2. Formulation of a logical hypothesis subject to testing by experimental methods and
3. Objective evaluation of the hypothesis on the basis of the experimental results.

Much could be written about these essential features. How does one arrive at a hypothesis? How does one design an experiment? How does one objectively evaluate a hypothesis? We shall be content with saying very little.

Science is a branch of study which deals with the observation and classification of facts. The scientist must, then, be able to observe an event or set of events as the result of a plan or design. This is the experiment, the substance of the scientific method. Experimental design is a field of statistics.

Objective evaluation of a hypothesis poses problems. Thus it is not possible to observe all conceivable events and, since exact laws of cause and effect are generally unknown, variation will exist among those which are observed. The scientist must then reason from particular cases to wider generalities. This process is one of uncertain inference. It is a process which enables us to disprove hypotheses that are incorrect but does not permit us to prove hypotheses that are correct. The only thing we can muster by way of proof is *proof beyond a reasonable doubt*. Statistical procedures are methods which lead us to this sort of proof.

A part of the possible information necessarily leads only to uncertain inference. Chance is involved in supplying the information and is the cause of the uncertainty. Today's statistician, applying the laws of chance, is able to place a precise and objective measure on the uncertainty of his inference. Actually, this is done for the totality of his inferences rather than for the individual inference. In other words, he follows a procedure that assures him of being right in 9 inferences out of 10, 99 out of 100, or anything short of being right all the time. Why not be right all or very close to all the time? The drawback is one of cost. Cost may be an increase in sample size, the penalty of a wrong decision, or the vagueness of the inference necessary to include the correct answer.

Scientific method is not disjoint hypothesis-experiment-inference sequences which fit into neat compartments. Instead, if the scientist fails to disprove a hypothesis, he may wonder if his theory does not embrace facts beyond the scope of inference of his experiment or if, with modification, it cannot be made to embrace such facts. Thus he is led through the cycle again. On the other hand, he may wonder if all the assumptions involved in his hypothesis are necessary; he formulates his hypothesis with fewer assumptions and starts through the cycle again.

In summary, statistics is a tool applicable in scientific method, for which it was developed. Its particular application lies in the many aspects of the design of an experiment from the initial planning to the collection of the data, and in the analysis of the results from the summarizing of the data to the evaluation of the uncertainty of any statistical inference drawn from them.

**1.4 Studying statistics.** No attempt will be made to make professional statisticians of those who read and study this book. Our aims are to promote clear, disciplined thinking, especially where the collection and interpretation of numerical data are concerned, and to present a considerable number of statistical techniques of general applicability and utility in research. Computations are required in statistics but this is arithmetic, not mathematics nor statistics.

Statistics implies, for most students, a new way of thinking—thinking in terms of uncertainties or improbabilities. Here as elsewhere, students vary in ability and when confronted with statistics for the first time, some may find a mental hurdle which can be emotionally upsetting. We believe we have done everything possible, consistent with our aims, to minimize the problems of learning statistics.

Many students will find that they learn statistics best by direct application to their own problems; few will find, in a course of one or two terms, use for all the material covered in this text. Consequently, many students will need considerable reflection and discussion in order to benefit most from a course based on this text. Questions and exercises are provided to provoke reflection and to offer some opportunity to apply and gain familiarity with techniques.

Finally, it is necessary to keep in mind that statistics is intended to be a tool for research. The research will be in genetics, marketing, nutrition, agronomy, and so on. It is the field of research, not the tools, that must supply the "whys" of the research problem. This fact is sometimes overlooked and the user is tempted to forget that he has to think, that statistics can't think for him. Statistics can, however, help the research worker to design his experiments and to evaluate objectively the resulting numerical data. It is our intention to supply the research worker with statistical tools that will be useful for this purpose.

### References

**1.1** Fisher, R. A.: "Biometry," *Biometrics*, **4:** 217–219 (1948).

**1.2** Fisher, R. A.: "The expansion of statistics," *J. Roy. Stat. Soc., Series A.*, **116:** 1–6 (1953).

**1.3** Fisher, R. A.: "The expansion of statistics," *Am. Scientist*, **42:** 275–282 and 293 (1954).

**1.4** Freeman, Linton C., and Douglas M. More: "Teaching introductory statistics in the liberal arts curriculum," *Am. Stat.*, **10:** 20–21 (1956).

**1.5** Hotelling, Harold: "The teaching of statistics," *Ann. Math. Stat.*, **11:** 1–14 (1940).

**1.6** Hotelling, Harold: "The impact of R. A. Fisher on statistics," *J. Am. Stat. Assoc.*, **46:** 35–46 (1951).

**1.7** McMullen, Launce: Foreword to "*Student's*" *Collected Papers*, edited by E. S. Pearson and John Wishart, *Biometrika* Office, University College, London, 1947.

**1.8** Mahalanobis, P. C.: "Professor Ronald Aylmer Fisher," *Sankhya*, **4:** 265–272 (1938).

**1.9** Mainland, Donald: "Statistics in clinical research; some general principles," *Ann. New York Acad. Sciences*, **52:** 922–930 (1950).

**1.10** Mather, Kenneth: "R. A. Fisher's *Statistical Methods for Research Workers*, an appreciation," *J. Am. Stat. Assoc.*, **46:** 51–54 (1951).

**1.11** Menger, Karl: "The formative years of Abraham Wald and his work in geometry," *Ann. Math. Stat.*, **23:** 13–20 (1952).

**1.12** Pearson, E. S.: "Karl Pearson, an appreciation of some aspects of his life and work, Part I: 1857–1906," *Biometrika*, **28:** 193–257 (1936).

**1.13** Pearson, E. S.: "Karl Pearson, an appreciation of some aspects of his life and work, Part II: 1906–1936," *Biometrika*, **29:** 161–248 (1938).

**1.14** Reid, R. D.: "Statistics in clinical research," *Ann. New York Acad. Science*, **52:** 931–934 (1950).

**1.15** Tintner, G.: "Abraham Wald's contributions to econometrics," *Ann. Math. Stat.*, **23:** 21–28 (1952).

**1.16** Walker, Helen M.: "Bi-centenary of the normal curve," *J. Am. Stat. Assoc.*, **29:** 72–75 (1934).

**1.17** Walker, Helen M.: "Statistical literacy in the social sciences," *Am. Stat.*, **5:** 6–12 (1951).

**1.18** Walker, Helen M.: "The contributions of Karl Pearson," *J. Am. Stat. Assoc.*, **53:** 11–27 (1958).

**1.19** Wolfowitz, J.: "Abraham Wald, 1902–1950," *Ann. Math. Stat.*, **23:** 1–13 (1952).

**1.20** Yates, F.: "The influence of *Statistical Methods for Research Workers* on the development of the science of statistics," *J. Am. Stat. Assoc.*, **46:** 19–34 (1951).

**1.21** Youden, W. J.: "The Fisherian revolution in methods of experimentation," *J. Am. Stat. Assoc.*, **46:** 47–50 (1951).