LOCKHEED MARTIN ENERGY RESEARCH LIBRARIES

3 4456 0515528 4

2

ORNL-4347
UC-32 – Mathematics and Computers

LECTURE SERIES IN STATISTICS AND PROBABILITY

OAK RIDGE NATIONAL LABORATORY
CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION
LIBRARY LOAN COPY
DO NOT TRANSFER TO ANOTHER PERSON
If you wish someone else to see this document, send in name with document and the library will arrange a loan.
UCN-7969
(3 3-67)

OAK RIDGE NATIONAL LABORATORY
operated by
UNION CARBIDE CORPORATION
for the
U.S. ATOMIC ENERGY COMMISSION

Notice
NOV 11 '69

ORNL-4347
Errata

NOV 1 '69

LECTURE SERIES IN STATISTICS AND PROBABILITY

Statistics Group

Corrections to the subject report have been printed on gummed stock and may be cut and glued to the pages as noted.

tes of America. Available from Clearinghouse for Federal
chnical Information, National Bureau of Standards,
ent of Commerce, Springfield, Virginia 22151
Printed Copy $3.00; Microfiche $0.65

LEGAL NOTICE

ccount of Government sponsored work. Neither the United States,
on acting on behalf of the Commission:
esentation, expressed or implied, with respect to the accuracy,
of the information contained in this report, or that the use of
method, or process disclosed in this report may not infringe
respect to the use of, or for damages resulting from the use of
thod, or process disclosed in this report.
cting on behalf of the Commission" includes any employee or
employee of such contractor, to the extent that such employee
, or employee of such contractor prepares, disseminates, or
on pursuant to his employment or contract with the Commission,
ctor.

ORNL-4347
UC-32 – Mathematics and Computers

Contract No. W-7405-eng-26

MATHEMATICS DIVISION

LECTURE SERIES IN STATISTICS AND PROBABILITY

Statistics Group

Edited by J. J. Beauchamp

MARCH 1969

OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee
operated by
UNION CARBIDE CORPORATION
for the
U. S. ATOMIC ENERGY COMMISSION

LOCKHEED MARTIN ENERGY RESEARCH LIBRARIES
3 4456 0515528 4

CONTENTS

LECTURE I: PROBABILITY, RANDOM VARIABLES, AND DISTRIBUTION FUNCTIONS – **Donald A. Gardiner** 1

1. PROBABILITY 1
   - 1.1 Relative Frequency 1
   - 1.2 Axiomatic Probability 2
     - The Three Axioms 2
     - Conditional Probability 3
     - Sets as Events 3
     - Independence 3
   - 1.3 Inductive Probability 4
2. RANDOM VARIABLES 4
   - 2.1 Coin Tossing 5
   - 2.2 Radioactive Decay 5
   - 2.3 A Piscatorial Experiment 6
3. DISTRIBUTION FUNCTIONS 6
   - 3.1 Some Properties of the Distribution Function 7
   - 3.2 The Probability Density Function 8
     - Coin Tossing 8
     - Poisson Distribution 8
     - A Gamma Distribution 9
     - Exponential Distribution 9
   - 3.3 Transformations of Random Variables 9

LECTURE II: CHARACTERIZATIONS AND LIMIT THEOREMS – **V. R. Rao Uppuluri** 11

1. INTRODUCTION AND SUMMARY 11
2. REVIEW 11
3. CHARACTERIZATION THEOREMS 12
   - 3.1 Normal Distribution 13
   - 3.2 Gamma and Exponential Distributions 14
4. STANDARD DISTRIBUTIONS 15
   - 4.1 Chi-Square Distribution 15
   - 4.2 t Distribution 15
   - 4.3 F Distribution 16
5. EXPECTATION 16
   - 5.1 Characteristic Functions 16
   - 5.2 Moments 17
6. CONVERGENCE THEOREMS 18
   - 6.1 Convergence of a Sequence of Random Variables 18
   - 6.2 Laws of Large Numbers 19
   - 6.3 Central Limit Theorem 20

LECTURE III: CONTINGENCY TABLES – **Marvin A. Kastenbaum** 21

1. INTRODUCTION 21

2. STRUCTURE OF 2×2 CONTINGENCY TABLES 21
   - 2.1 One Multinomial, or Two Variates (Responses) 22
   - 2.2 Two Binomial Samples, or One Variate (Response) and One Way of Classification (Factor) 23
   - 2.3 One Hypergeometric, or Two Ways of Classification 23

3. DEGREES OF FREEDOM 24
   - 3.1 One Multinomial 24
   - 3.2 Two Binomials 24
   - 3.3 One Hypergeometric 24

4. STRUCTURE OF $r \times s$ CONTINGENCY TABLES 25
   - 4.1 One Multinomial 25
   - 4.2 r Multinomial Samples 25
   - 4.3 One Hypergeometric 25

5. THE NULL HYPOTHESIS IN THE TWO-WAY TABLE 26
   - 5.1 One Multinomial 26
   - 5.2 r Multinomials 26
   - 5.3 One Hypergeometric 27

6. TESTS OF HYPOTHESES IN THE TWO-WAY TABLE 27
   - 6.1 Yates' Correction for Continuity 29
   - 6.2 Cochran's Recommendations Concerning Analyses of Two-Way Tables 30

7. ALTERNATIVE ANALYSES TO THE TRADITIONAL CHI-SQUARE 30
   - 7.1 The Logit Transformation 30
   - 7.2 The Likelihood-Ratio Test 32

8. PARTITIONING OF CONTINGENCY TABLES 33

9. CONTINGENCY TABLES OF MORE THAN TWO DIMENSIONS 34
   - 9.1 Hypotheses Concerning the Two-Way Marginals of a Three-Way Table 35
   - 9.2 Hypotheses Which Have No Analog in a Two-Way Table 35

10. SPECIAL TOPICS IN CONTINGENCY TABLES 37
    - 10.1 Missing and "Mixed-Up" Values 37
    - 10.2 Consumer Preference 37
    - 10.3 Markov Chains 38
    - 10.4 Measures of Association 38
    - 10.5 Bayesian Analysis 38

RECENT BIBLIOGRAPHY ON CONTINGENCY TABLES 39

INDEX 49

LECTURE IV: CONFIDENCE REGIONS – **Marvin A. Kastenbaum** 53

1. AN EXAMPLE OF STATISTICAL INFERENCE 53
2. CONFIDENCE INTERVALS FOR A MEAN 54
3. CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO MEANS 55
4. CONFIDENCE INTERVALS FOR A VARIANCE 57
5. CONFIDENCE INTERVALS FOR THE RATIO OF TWO VARIANCES 59
6. CONFIDENCE REGIONS FOR A VECTOR OF p MEANS 61
7. CONFIDENCE REGIONS FOR A BIVARIATE VECTOR OF MEANS 62
8. REFERENCES 64
9. APPENDIX 64
   - 9.A The Chi-Square Distribution (χ^2): See Table 3 64
   - 9.B The t Distribution: See Table 2 64
   - 9.C The F Distribution: See Table 4a and b 65

TABLES

- 1a. The Cumulative Normal Distribution Function 66
- 1b. The Cumulative Normal Distribution Function 67
- 2. Distribution of t 68
- 3. Distribution of χ^2 69
- 4a. Distribution of F – Upper 5% Points 70
- 4b. Distribution of F – Upper 1% Points 71

LECTURE V: DISTRIBUTION-FREE TESTS – **W. E. Lever** 72

1. INTRODUCTION 72
2. ADVANTAGES AND DISADVANTAGES OF DISTRIBUTION-FREE TESTS 72
3. SIGN TEST 73
4. SIGNED RANK TEST 75
5. THE RANK SUM TEST 76
6. A K-SAMPLE PROCEDURE FOR UNMATCHED DATA – ONE-WAY CLASSIFICATION 78
7. A K-SAMPLE PROCEDURE FOR MATCHED DATA – TWO-WAY CLASSIFICATION 80
8. CONCLUSION 82

REFERENCES 82

LECTURE VI: LINEAR REGRESSION – **T. L. Hebble** 83

1. INTRODUCTION 83
   - 1.1 Notation 83

2. LINEAR REGRESSION 83
2.1 Definition of Problem 83
3. ESTIMATION OF PARAMETERS 85
3.1 Method of Least Squares 85
3.2 Example 87
3.3 Gauss-Markov Theorem 88
3.4 Parameter Estimation by Method of Maximum Likelihood 89
4. TESTING THE REGRESSION MODEL 90
4.1 Test for Lack of Fit 90
4.2 Comparison of Models Using R^2 92
5. CONFIDENCE INTERVALS 92
5.1 Confidence Intervals About β 92
5.2 Confidence Interval About $E(\mathbf{Y})$ 94
5.3 Confidence Interval About Mean of Future Observations 94
6. SOME COMMENTS ON STEPWISE REGRESSION 94
7. APPENDIX: FUNDAMENTALS OF MATRICES 95
Basic Operations Using Matrices 96
Properties of Matrices 96
LECTURE VII: NONLINEAR ESTIMATION – **John J. Beauchamp** 98
1. INTRODUCTION 98
2. GEOMETRY OF LEAST SQUARES 99
3. METHODS OF ESTIMATION 101
3.1 Gradient or Steepest Descent Method 101
3.2 Gauss-Newton or Linearization Method 103
3.3 Marquardt's Compromise 104
4. ESTIMATES OF VARIANCES 105
5. INITIAL ESTIMATES 105
REFERENCES 107
LECTURE VIII: ANALYSIS OF VARIANCE – **David G. Gosslee** 108
1. INTRODUCTION 108
2. MODELS 108
2.1 Variance Components Model 109
2.2 Fixed Model 109
3. CALCULATIONS 109
4. TESTS OF HYPOTHESES 112

5. ASSUMPTIONS 113

6. DESIGN OF EXPERIMENTS 113

7. TRANSFORMATIONS 114

8. MULTIPLE COMPARISONS 115

9. CONCLUDING REMARKS 115

REFERENCES 115

LECTURE IX: DESIGN OF EXPERIMENTS AND RESPONSE SURFACE METHODOLOGY – **Toby J. Mitchell** 116

1. INTRODUCTION 116

2. BASIC PRINCIPLES OF EXPERIMENTAL DESIGN 117
   - 2.1 Classification of Variables 117
   - 2.2 Randomization 118
   - 2.3 Blocking 120
   - 2.4 "Fixed Variables" 122
   - 2.5 Replication 122
   - 2.6 Sequential Experimentation 123

3. SOME STANDARD DESIGNS 123
   - 3.1 Completely Randomized Design 123
   - 3.2 Randomized Block Design 123
   - 3.3 Latin Square Design 124
   - 3.4 Graeco-Latin Squares 126
   - 3.5 Balanced Incomplete Block Design 126
   - 3.6 Youden Square Design 127

4. FACTORIAL AND FRACTIONAL FACTORIAL DESIGNS 127
   - 4.1 Factorial Designs (General) 127
   - 4.2 2^k Factorial Designs 129
   - 4.3 2^{k-1} Fractional Factorials 131
   - 4.4 2^{k-p} Fractional Factorials 134
     - Resolution III Designs 134
     - Resolution IV Designs 135
   - 4.5 Other Fractions of 2^k Factorials: The Plackett-Burman Designs 137
   - 4.6 Fractions of General Factorial Designs 137
   - 4.7 Blocking Factorial and Fractional Factorial Designs 137

5. THE MODEL-ORIENTED APPROACH TO EXPERIMENTAL DESIGN 137
   - 5.1 Notation 137
   - 5.2 First-Order Designs 138
   - 5.3 Second-Order Designs 140
   - 5.4 Higher-Order Designs and Designs for Nonlinear Models 141

6. RESPONSE SURFACE METHODOLOGY 142
6.1 Viewing the Response as a Surface 142
6.2 An Optimum-Seeking Method: Steepest Ascent 142
REFERENCES 143
LECTURE X: ELEMENTS OF STOCHASTIC PROCESSES – **V. R. Rao Uppuluri** 145
1. DEFINITIONS AND PRELIMINARIES 145
2. EXAMPLES 146
2.1 Brownian Motion Process (or Wiener Process) 146
2.2 Poisson Process 147
3. MARKOV CHAINS 150
3.1 Definitions and Preliminaries 150
3.2 One-Dimensional Random Walks 151
3.3 A Discrete Queueing Markov Chain 152
3.4 Success Runs 153
3.5 Branching Processes 153
3.6 A Genetic Model 154
4. PROPERTIES OF MARKOV CHAINS 154
4.1 n-Step Probabilities 154
4.2 Classification of States 154
4.3 Limit Theorems 155
ACKNOWLEDGMENT 156
REFERENCES 156

INTRODUCTION

This report contains the lecture notes from the series of lectures in statistics and probability given by members of the Statistics Group of the Mathematics Division at ORNL during the period June 10, 1968, to August 12, 1968. Lecture notes were made available at each lecture, and since the completion of this series of lectures the requests from Laboratory personnel for additional sets of the lecture notes have prompted the compilation of these notes into this report. These lectures were held on a weekly basis and were open to all ORNL personnel as well as those interested individuals at Y-12 and K-25. The lectures were well received by the majority of those attending, who ranged from 75 to 150 individuals per lecture. From an examination of the table of contents of this report, which lists the title and lecturer for each lecture, it can be seen that this series covered a wide variety of statistical topics. Each lecture was as self-contained as possible. The notes contained in this report emphasize the important concepts in each lecture.

This series of lectures in statistics and probability was an effort by the Statistics Group to familiarize investigators in the physical and biological sciences with statistical methods and concepts. Since the need for the services of statisticians has been recognized by the Laboratory, one of the primary functions of the Statistics Group of the Mathematics Division has been "to encourage an atmosphere in which new concepts can be explored and fundamental research carried out using statistics and biomathematics."[1]

In line with this function these lecture notes are published to:

1. indicate the tools which are available in statistical methodology;
2. acquaint those interested individuals with the language and terminology of statistics; and
3. illustrate the manner in which statistics may be applied in scientific research.

---

[1] *Math. Div. Ann. Progr. Rept. Dec. 31, 1963,* ORNL-3567, p. 33.

LECTURE I: PROBABILITY, RANDOM VARIABLES, AND DISTRIBUTION FUNCTIONS

Donald A. Gardiner

1. PROBABILITY

Any study, whether intensive or casual, of the subject called statistics must begin with an understanding of probability because probability is the basis of statistics. In fact, at universities where statistics is taught seriously the student is given at least a quarter or semester of probability before he is allowed to take his first real course in statistics. Since this series of lectures is intended to be a survey of some topics in statistics and not a formal course, we need not be so pedantic, but a proper groundwork must be laid. We will begin by outlining the three concepts of probability and by pointing out that the three concepts are not necessarily incompatible.

1.1 Relative Frequency

While interned in Denmark during World War II a South African statistician named Kerrich tossed a coin 10,000 times. He kept careful records, which fortunately he was able to preserve, and these records have become a valuable source for examples and illustrations. Heads turned up 5067 times. One is inclined to conclude that the probability that a head will turn up on a single toss of Kerrich's coin is 0.5067. This is an example of the relative frequency approach to a definition of probability.

Consider a simple experiment the outcome of which cannot be known in advance but in which the possible outcomes may be classified as "favorable" or "unfavorable." Should the experiment be performed a number of times, the ratio

$$\frac{\text{number of favorable outcomes}}{\text{number of trials}}$$

is the relative frequency of favorable outcomes. Intuitively we feel that as the number of trials increases the relative frequency will settle down to some stable value greater than zero and less than unity. The limit of this ratio as the number of trials increases is taken as the relative frequency definition of the probability of a favorable outcome. Inherent in this definition is the assumption that a favorable outcome from one trial is as likely as a favorable outcome from any other trial and that the results of one trial do not affect the results of another trial.

In the example of Kerrich's coin the favorable outcome was the appearance of a head and the experiment was the tossing of the coin. However, 10,000 tosses is not the limit of the number of tosses, even though Kerrich may have wished that it were. The ratio 5067/10,000 is a rather good estimate, but we would be very reluctant to state that the probability of getting a head on Kerrich's coin is 0.5067, because if Kerrich had stopped at some other number of trials the ratio would have been different.

1.2 Axiomatic Probability

Although the relative frequency approach has great appeal on intuitive grounds it has been more fruitful to treat probability in a more formal manner mathematically. Such an approach is the approach based on three simple axioms.

Consider a set $\mathcal{A}$ whose elements are denoted by x. If $\mathcal{A}$ is a set consisting of all the possible outcomes of an experiment, $\mathcal{A}$ is called a sample space and x represents a possible outcome. $\mathcal{A}$ contains subsets, A, B, C, etc., and it will be of interest to consider

1. the probability that $x \in A$, written $P(A)$,
2. the probability that $x \in B$, written $P(B)$,
3. the probability that $x \in \mathcal{A}$, written $P(\mathcal{A})$,
4. other probabilities.

$P(A)$, $P(B)$, $P(\mathcal{A})$, etc. are called probability set functions if they satisfy the three axioms of probability.

Writing ϕ as the null set (the set which contains no elements) and with the understanding that A and B are subsets of $\mathcal{A}$, the three axioms of probability are as given below.

The Three Axioms

1. $P(A \geqq 0$.
2. $P(A \cup B) = P(A) + P(B)$ if A and B are disjoint sets, that is, if $A \cap B = \phi$.
3. $P(\mathcal{A}) = 1$.

Axiom 2 is the familiar "additive law" of probability which applies to mutually exclusive sets.

From these axioms many theorems and corollaries may be deduced, the most important of which are these:

1. If A^c is the complement of A, $P(A^c) = 1 - P(A)$.
2. $P(\phi) = 0$.
3. If $A \subset B$, $P(A) \leqq P(B)$.
4. $0 \leqq P(A) \leqq 1$.
5. Whether or not $A \cap B = \phi$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

The reader may find it amusing and instructive to prove these results from the three axioms. All that is necessary is a knowledge of elementary set theory.

The question naturally arises: "How does one obtain a numerical value for $P(A)$?" This is the question of choosing a probability model, the answer to which, from a strictly mathematical point of view, is purely arbitrary. But from a practical point of view the question is not trivial, and the answer, of course, depends upon the type of experiment that is contemplated and what is known about it. In the coin-tossing experiment one might choose the probability of the appearance of a head to be $\frac{1}{2}$. This could be taken as the definition of a "fair" or "unbiased"

coin. If one assumes that Kerrich's coin was a fair coin, one assumes that the limit of the relative frequency is $\frac{1}{2}$. There is no way of proving or disproving this hypothesis absolutely, but one can see that, in this sense, the relative frequency concept is compatible with the axiomatic concept of probability.

Conditional Probability

Before discussing the third concept of probability it will be convenient to introduce the notion of conditional probability.

Let A and B be two subsets of the sample space $\mathcal{A}$, and let x be an element of $\mathcal{A}$. We inquire now into the question: "If $x \in B$, what is the probability that $x \in A$?"

The probability in question is called a "conditional" probability because it is the probability that x is an element of the set A conditional upon the fact that x is an element of the set B. It is written $P(A|B)$, which is read "the probability of A given B," and is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \; P(B) > 0 .$$

Note that this is a definition and not a result of the three basic axioms. However, $P(A|B)$ does satisfy the three axioms, and therefore the same theorems which hold for "unconditional" probabilities hold for conditional probabilities as well. [In showing this, one should treat B as a sample space so that $P(B|B) = 1$.]

Sets as Events

In statistical parlance an outcome of an experiment is called an *elementary event*, and the collection of all outcomes which belong to a certain subset is called simply an *event*. Thus there is the correspondence

outcome, x: elementary event,
subset, A: event, A.

If, in the performance of an experiment, an outcome $x \in A$ has been observed, we say that the event A has occurred. It will be useful to speak of the event A and its probability $P(A)$ in defining independence.

Independence

Independence in the statistical sense is often called stochastic independence to differentiate it from mathematical independence. We say that the event A is *independent* of the event B if $P(A|B) = P(A)$ and conversely. If the equality does not hold we say the events and their probabilities are dependent.

The definition of independence allows us to state the familiar multiplicative rule of probability, namely that if the events A and B are independent, then

$$P(A \cap B) = P(A) \times P(B) ,$$

which is a direct consequence of the definition of conditional probability.

Frequently in applications $P(A|B)$ is difficult to postulate, whereas $P(A)$ and $P(B)$ may not be so difficult. But $P(A|B) = P(A)$ if the events are independent, and so one is well advised to plan one's experiments to take advantage of the resulting simplification.

1.3 Inductive Probability

The third concept of probability we wish to consider is inductive probability. It has been defined as the degree of belief one is willing to place on a proposition in the light of certain evidence. We see immediately that inductive probability is not the same as the limit of a relative frequency, nor is it the same as the probability that an element belongs to a set of a sample space.

One might say that one is "almost sure" that Kerrich's coin was a fair coin, in which case "almost sure" is a nonquantitative measure of inductive probability. Some might wish to quantify the inductive probability and say, perhaps, "I am 95% certain that Kerrich's coin was fair," but the 95% is still a degree of belief or an inductive probability.

This statement about Kerrich's coin contains a mixture of two kinds of probability: the 95% is a measure of inductive probability, and the word "fair" means that the probability that a head appears is $\frac{1}{2}$.

Another statement that could be made about Kerrich's coin is: "I am 95% confident that the probability that a head turns up is between 0.4967 and 0.5167." This is an example of statistical inference which employs the idea of a confidence interval. Like the previous statement it contains two kinds of probabilities: the 95% is a measure of inductive probability, and the interval (0.4967, 0.5167) is a measure of the probability for a head with Kerrich's coin.

2. RANDOM VARIABLES

Given sufficient ingenuity it is always possible to associate with every possible outcome of an experiment a number on the real line. We shall use lower-case letters to denote these numbers corresponding to the possible outcomes. Then, for example, in a coin-tossing experiment we could set up the correspondence

if a head turns up, $x = 1$,
if a tail turns up, $x = 0$.

Before the experiment is performed we do not know whether $x = 0$ or $x = 1$. Therefore, to represent the outcome of the experiment without a knowledge of how the experiment turns out, we will use

the corresponding capital letter, in this case X. The real number which represents the outcome of an experiment before the outcome of the experiment is known is called a *random variable*. Some examples will be instructive.

2.1 Coin Tossing

In the coin-tossing experiment we could set up the following definitions:

Sample space:	$\mathcal{A}$ = {head, tail}	or $\mathcal{A} = \{x;\ x = 0, 1\}$
Subsets:	A = {head}	or $A = \{x;\ x = 1\}$
	B = {tail}	or $B = \{x;\ x = 0\}$

Then in terms of the random variable X we have

$$P(\mathcal{A}) = \text{Prob}\ \{X = 0 \text{ or } 1\} = 1\ ,$$
$$P(A) = \text{Prob}\ \{X = 1\} = p, \text{ say}\ ,$$
$$P(B) = \text{Prob}\ \{X = 0\} = 1 - p\ .$$

It is an interesting aside to point out that $P(A)$ and $P(B)$ may be summarized in one expression,

$$f(x) = p^x\ (1 - p)^{1-x},\ x = 0, 1,\ 0 \leq p \leq 1\ .$$

Thus if $x = 0$, $f(x) = P(B) = 1 - p$, or if $x = 1$, $f(1) = P(A) = p$. For the moment we will consider this to be merely an ingenious summarization. The point here is to illustrate the use of the random variable to describe an experiment.

2.2 Radioactive Decay

Suppose an experiment consists of placing a radioactive source before a counter for a specified period of time. The possible outcomes of the experiment are all the integers which might register on the counter. In this case the definition of the random variable is straightforward and requires no ingenuity whatsoever. We could set up the following correspondence:

Sample space:	$\mathcal{A}$ = {all nonnegative integers}	or $\mathcal{A} = \{x;\ x = 0, 1, 2, \ldots\}$
Subsets:	A = {integers less than 10}	or $A = \{x;\ x = 0, 1, 2, \ldots, 9\}$
	B = {integers between 40 and 50}	or $B = \{x;\ x = 41, 42, \ldots, 49\}$.

Then in terms of the random variables X we have

$$P(\mathcal{A}) = \text{Prob}\ \{X = 0, 1, 2, \ldots\} = 1\ ,$$
$$P(A) = \text{Prob}\ \{X = 0, 1, 2, \ldots, 9\}\ ,$$
$$P(B) = \text{Prob}\ \{X = 41, 42, \ldots, 49\}\ .$$

If we wished we could postulate a probability model from which we could calculate probabilities for the several sets. Such a model could be

$$\text{Prob}\,\{X = x\} = \frac{e^{-\lambda}\,\lambda^x}{x!},\ x = 0, 1, 2, \ldots,\quad 0 < \lambda\,,$$

but this is purely arbitrary. What we have illustrated is the reduction of the description of an experiment to mathematical terms using the idea of the random variable.

2.3 A Piscatorial Experiment

A sport fisherman would agree that he could not predict the weight of a fish he might catch, and he might also agree that it is reasonable to treat the weight of a fish as a random variable. To describe this phenomenon we could define some sets as follows:

Sample space: $\mathcal{A}$ = {all possible weights} or $\mathcal{A} = \{x;\ 0 < x < \infty\}$

Subsets: A = {weights of fry} or $A = \{x;\ 0 < x < 1\}$

B = {weights of whoppers} or $B = \{x;\ 6 < x < \infty\}$

Although these definitions are quite arbitrary they illustrate the concept of the continuous random variable as opposed to the discrete random variables of the two preceding examples. The probability functions for these sets in terms of the random variables are

$$P(\mathcal{A}) = \text{Prob}\,\{0 < X < \infty\} = 1\,,$$
$$P(A) = \text{Prob}\,\{0 < X < 1\}\,,$$
$$P(B) = \text{Prob}\,\{6 < X < \infty\}\,.$$

It might be reasonable to postulate a function

$$f(x) = 2^{-\alpha} x^{\alpha-1} \exp\{-x/2\}/\Gamma(\alpha),\quad 0 < \alpha < \infty,\quad 0 \leq x < \infty$$
$$= 0 \quad \text{otherwise}$$

to aid in describing the probabilities so that

$$P(A) = \int_0^1 f(x)\,dx\,,$$

$$P(B) = \int_6^\infty f(x)\,dx\,,$$

but again this is purely arbitrary, and some other models could be advanced which might be more realistic.

3. DISTRIBUTION FUNCTIONS

Let X be a random variable defined on an interval of the real line, $a < x < b$. The *distribution function* of the random variable X is defined simply as

$$\text{Prob}\,\{X \leq x\}\,.$$

It is customary to use a capital letter from the first part of the alphabet to represent the distribution function, as, for example,

$$F(x) = \text{Prob}\,\{X \leqq x\}\ .$$

It will be useful to consider the derivative of $F(x)$, which we will denote by the corresponding lower-case letter. Thus

$$f(x) = \frac{dF(x)}{dx}\ .$$

If X is a continuous random variable, $f(x)$ is a derivative in the ordinary sense and it is defined at each point of continuity of x. If X is a discrete random variable, then $f(x)$ is not a derivative in the ordinary sense. In the discrete case we define

$$f(x_i) = F(x_i) - F(x_{i-1})\ ,$$

where we assume a sample space $\mathcal{A}$ such that

$$\mathcal{A} = \{x_1,\ x_2,\ \ldots;\ a \leqq x_1 \leqq x_2 \leqq \ldots \leqq x_{i-1} \leqq x_i \leqq \ldots \leqq b\}\ .$$

Although it is frequently overlooked for convenience, it takes at least three statements to completely specify the distribution function if the range of the variable, $(a,\ b)$, has finite limits. Thus

$$F(x) = \begin{cases} 0, & x < a \\ \displaystyle\int_a^x f(t)\,dt, & a \leqq x < b \\ 1, & b \leqq x \end{cases}$$

If X is a continuous random variable the integral above is an ordinary Riemann integral. But if X is discrete we interpret the integral to mean

$$\sum_A f(t)\ ,$$

in which A is the set of all values of t which are less than or equal to x but greater than or equal to a.

3.1 Some Properties of the Distribution Function

A study of the definition of the distribution function $F(x)$ will reveal that it possesses the following properties:

1. $0 \leqq F(x) \leqq 1$.
2. $F(x)$ is a nondecreasing function of x.
3. $F(\infty) = 1$ and $F(-\infty) = 0$.
4. $F(x)$ is continuous to the right at each point x.

Roughly speaking, property 4 above means that the limit of Prob $\{X \leq x\}$, as one approaches x from values larger than x, exists and this limit is equal to $F(x)$.

From these properties we may deduce that for $x_1 < x_2$

$$\text{Prob}\ \{x_1 < X \leq x_2\} = F(x_2) - F(x_1)\ .$$

If $x_1 = x_2 = x_0$, Prob $\{X = x_0\}$ is equal to the height of the step of $F(x)$ at the point x_0. Thus, if $F(x)$ is continuous at $x = x_0$, Prob $\{X = x_0\} = 0$ because the height of the step at x_0 is 0.

3.2 The Probability Density Function

We have already defined $f(x)$ as the derivative in one sense or another of the distribution function $F(x)$. We call $f(x)$ a *probability density function* (p.d.f.), and we have already seen some examples.

Coin Tossing

In the coin-tossing experiment we had a function which we will rewrite as

$$f(x) = p^x\ (1-p)^{1-x}, \quad x = 0, 1, \quad 0 \leq p \leq 1$$

$$= 0 \quad \text{otherwise}\ .$$

This is a bona fide p.d.f. whose distribution function is

$$F(x) = \begin{cases} 0, & x < 0 \\ 1-p, & 0 \leq x < 1,\ 0 \leq p \leq 1 \\ 1, & x \geq 1 \end{cases}$$

Poisson Distribution

The example about radioactive decay assumed a probability density function

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

$$= 0 \quad \text{otherwise.}$$

Its distribution function is

$$F(x) = \begin{cases} 0, & x < 0 \\ \sum_{t=0}^{x} \dfrac{e^{-\lambda}\lambda^t}{t!} & x = 0, 1, 2, \ldots, \quad 0 < \lambda\ . \end{cases}$$

This is known as the Poisson distribution and is a realistic model for many real situations.

A Gamma Distribution

As an example of a continuous random variable we used the weight of a fish and assumed a p.d.f.

$$f(x) = 2^{-\alpha} x^{\alpha-1} \exp\{-x/2\}/\Gamma(\alpha), \quad 0 \leq x < \infty, \quad 0 < \alpha < \infty$$

$$= 0 \quad \text{otherwise} .$$

The distribution function for this model, which is a special form of the gamma distribution function, may be expressed only as

$$F(x) = \begin{cases} 0, & x < 0 \\ \int_0^x f(t)\, dt, & 0 \leq x < \infty \end{cases}$$

because the integral does not exist in closed form. However, this distribution function is tabled for many values of α.

Exponential Distribution

As an example of a probability model for a random variable whose distribution function does exist in closed form, consider

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\theta x}, & 0 \leq x < \infty, \quad 0 < \theta < \infty . \end{cases}$$

The probability density function for this random variable is

$$f(x) = \theta e^{-\theta x}, \quad 0 \leq x < \infty, \quad 0 < \theta < \infty$$

$$= 0 \quad \text{otherwise.}$$

This model is used extensively in the study of survival of organisms and in life testing.

3.3 Transformations of Random Variables

Very frequently one is interested in the behavior of a random variable that is some function of another random variable. We shall illustrate one method of finding the distribution function (and hence the p.d.f.) of a transformed random variable.

Suppose a random variable X is defined by a distribution function $F(x)$ such that

$$F(x) = \begin{cases} 0, & x < -1 \\ (1 + x)/2, & -1 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$$

and that we are interested in the distribution function of the random variable $Y = X^2$. We have

$$\begin{aligned}\text{Prob}\,\{Y \leqq y\} &= \text{Prob}\,\{X^2 \leqq y\} = \text{Prob}\,\{-\sqrt{y} \leqq X \leqq \sqrt{y}\} \\ &= F(\sqrt{y}) - F(-\sqrt{y})\,,\end{aligned}$$

since Prob $\{X = -\sqrt{y}\} = 0$. Continuing,

$$F(\sqrt{y}) - F(-\sqrt{y}) = \frac{1+\sqrt{y}}{2} - \frac{1-\sqrt{y}}{2} = \sqrt{y}\,.$$

Therefore

$$\text{Prob}\,\{Y \leqq y\} = G(y) = \begin{cases} 0, & y < 0 \\ \sqrt{y}, & 0 \leqq y < 1 \\ 1, & y \geqq 1 \end{cases},$$

which defines the distribution function for Y. The probability density function will be the derivative of $G(y)$ with respect to y or

$$\begin{aligned} g(y) &= \frac{1}{2\sqrt{y}}\,, \quad 0 < y \leqq 1 \\ &= 0 \quad \text{otherwise.}\end{aligned}$$

LECTURE II: CHARACTERIZATIONS AND LIMIT THEOREMS

V. R. Rao Uppuluri

1. INTRODUCTION AND SUMMARY

In this lecture we shall review the fundamental concepts of a probability space, random variable, distribution function, and the probability density function and indicate the tools of analysis that will be essential to probe into these areas. In Sect. 3, we shall indicate the role played by the concept of independence in the context of probability theory by pointing out several characterization theorems of the normal (Gaussian) distribution and the exponential distribution. In Sect. 4 we will give the analytical expressions of several of the standard distributions that are encountered in statistics and indicate how they are derived. In Sect. 5 we will give the formulas for the characteristic function and moments of some of the standard continuous distributions. In the last section we will introduce the concepts of convergence of random variables and distribution functions and give statements of the laws of large numbers and the central limit theorem.

2. REVIEW

In the study of any random phenomena, the concepts of (1) probability space, (2) random variable, (3) distribution function, and (4) probability density function are very fundamental.

In probability theory, the basic sample space and the set of all conceivable events with a given probability structure will be the natural level at which one would work. In other words, one would work with a probability space and make very fine analysis using measure-theoretic approaches. Thus the results established would be theoretical in nature and will tend to be in the nature of existence theorems, as can be illustrated by the strong law of large numbers and the ergodic theorem.

At times, in several physical or natural phenomena it may be possible to directly focus attention on the random variables under consideration. By making a fine analysis of the structure of the phenomena, it may be possible to pin down the relationships between the random variables under consideration. This can be illustrated in the context of regression analysis and analysis of variance, about which we will hear more. The distributional properties and the moment properties of these structured random variables are generally helpful to understand the underlying phenomena. In this context, it may be worth adding a note that the concept of a random variable is that of a continuous function where the domain of the function can be an arbitrary sample space but the range is usually the real line or the Euclidean space. Also, random variables help to bridge the gap between the results proved by a probability theorist and the user. Several types of convergence of random sequences illustrate this point, and an applied person may be satisfied with the weakest type of a convergence theorem, whereas the theorist is restless till he proves the strongest type of convergence theorems.

The next level one may choose to work at may be with the so-called distribution functions associated with random phenomena. A distribution function may be characterized as a real-valued, nonnegative, bounded (less than or equal to 1), and monotone nondecreasing function. A distribution function need not necessarily be continuous, but it will have at most a countable number of discontinuities. It is important to note that if one makes an assumption about the form of a distribution function, it means that one is looking at a subclass of these functions. If one is fastidious about making technical assumptions, like the differentiability of a distribution function, one would work with distribution functions in studying the properties of the underlying random phenomena, and one needs tools like the Lebesgue-Stieltjes integrals. Also, analytically it turns out to be simpler to work with distribution functions, as contrasted to probability density functions, in proving limit theorems like the central limit theorems.

Traditionally, any data that are recorded are grouped and represented as frequency histograms. Then one looks for patterns in these histograms. Some try to find mathematical functions that fit these histograms. If one consults a statistics book, one finds a certain class of neat mathematical functions like (1) the normal probability density function, (2) gamma distribution, (3) beta distribution, (4) uniform distribution, (5) binomial distribution, (6) Poisson distribution, and so on, and one looks for a function that best fits the data. Any nonnegative function which, when integrated over the whole range of permissible values, integrates to unity may be characterized as a probability density function. The data and the phenomenon that is under study generally help to pin down the probability density function that may be appropriate for a particular problem. At times this may be an initial step to formulate further hypotheses in the phenomenon under study and to carry on experimentation for further verification. This leads to the aspect of statistical inference. Though I have described the four concepts in the order (1) probability space, (2) random variable, (3) distribution function, and (4) probability density function, the importance of concept (4) should not be underrated. It should be acknowledged that the growth of some of the ideas in probability theory has its origin in the study of the properties of probability density functions. Since this study requires only advanced calculus as a tool, this is in the reach of many.

3. CHARACTERIZATION THEOREMS

Though it may not be specified, in many practical situations one would work with random phenomena where one can change levels in the above setup. In other words, one would like to believe that one lives in a world free of pathological situations. A situation where one can have a distribution function without necessarily having a (derivative or a) probability density function is an illustration of a pathological situation. For all practical purposes, any further concept introduced at one level will still be valid at all levels. The concept of independence is an illustration of this.

3.1 Normal Distribution

In terms of independence one can characterize the distributional properties of several random phenomena. We will illustrate below a few characterizations of the normal (Gaussian) distribution.

Theorem 3.1: Let $X_1, X_2, \ldots, X_n$ be n independent but not necessarily identically distributed random variables. Suppose that the two linear forms $Y_1 = a_1X_1 + \ldots + a_nX_n$ and $Y_2 = b_1X_1 + \ldots + b_nX_n$ are independently distributed. Then each random variable X_j which has nonzero coefficients in both forms is normally distributed.

The vector analog of the above theorem is also true and may be stated as

Theorem 3.2: Let $X_1, X_2, \ldots, X_n$ be n independent but not necessarily identically distributed p-dimensional random vectors. Suppose that the two random vectors $Y_1 = a_1X_1 + \ldots + a_nX_n$ and $Y_2 = b_1X_1 + \ldots + b_nX_n$ are independently distributed. Then each random vector X_j for which $a_jb_j \neq 0$ has a p-variate normal distribution.

Another characterization of the vector analog may be stated as

Theorem 3.3: Let $X_1, X_2, \ldots, X_n$ be n mutually independent p-dimensional random vectors, and let $A_1, A_2, \ldots, A_n, B_1, B_2, \ldots, B_n$ be nonsingular square matrices of order p. If

$$\sum_{i=1}^{n} A_iX_i'$$

and

$$\sum_{i=1}^{n} B_iX_i'$$

are independently distributed, then each X_i has a p-variate normal distribution.

The next theorem characterizes the normal distribution in terms of the independence of the sample mean and sample variance. This result is useful in deriving the distributions of random variables which are functions of sample mean and sample variance from a normal distribution, such as the Student's t statistic and Fisher's F statistic.

Theorem 3.4: Let $X_1, X_2, \ldots, X_n$ be a sample from a certain population, and denote the sample mean by $\bar{x}$ and the sample variance by s^2. The statistic $\bar{x}$ and s^2 are independent if and only if the population is normal.

It can be easily proved that the distribution of the sum of two independent normal variables is normally distributed. The next theorem tells us about a form of the converse of this statement.

Theorem 3.5: If the sum of two independent random variables is normally distributed, then each summand is also normally distributed.

3.2 Gamma and Exponential Distributions

A nonnegative random variable X is said to have a gamma distribution if its probability density function is given by

$$f(x) = \beta^{\alpha} e^{-x\beta} x^{\alpha-1} / \Gamma(\alpha) , \quad \alpha > 0 . \tag{1}$$

When $\alpha = 1$, X is said to be exponentially distributed; β is referred to as the scale parameter. It can be easily proved that if Z_1 and Z_2 are two independently distributed gamma variables with the same scale parameter, then $Z_1 + Z_2$ and Z_1/Z_2 are independently distributed. The next theorem is a converse of this statement, which together with this remark characterizes the gamma distribution.

Theorem 3.6: If Z_1 and Z_2 are independent nonnegative random variables for which $Z_1 + Z_2$ and Z_1/Z_2 are independent, then both Z_1 and Z_2 have gamma distributions with a common scale parameter.

The next theorem shows that the only distribution which has the property of not having memory is the exponential distribution.

Theorem 3.7: Let T be a nonnegative random variable.

$$P(T > x + y \mid T > x) = P(T > y)$$

if and only if T has an exponential distribution.

The next theorem gives a characterization of the exponential distribution in terms of the minimum and range of two random variables.

Theorem 3.8: Suppose that the random variables X and Y are independent and have absolutely continuous distributions. Then in order that $U = \min(X, Y)$ and $V = X - Y$ be independent, it is necessary and sufficient that both X and Y have exponential distributions.

Let $X_1, X_2, \ldots, X_n$ be n independent identically distributed nonnegative random variables. If we arrange them in order of magnitude as $X_{1,n} \leqq X_{2,n} \leqq \ldots \leqq X_{n,n}$, $X_{k,n}$ is referred to as the kth-order statistic in a sample of size n. Let us define

$$Y_i = (n - i + 1)(X_{i,n} - X_{i-1,n}) , \quad i = 1, 2, \ldots, n ,$$

with the understanding that $X_{0,n} = 0$. The next theorem characterizes the exponential distribution in terms of $Y_1, Y_2, \ldots, Y_n$.

Theorem 3.9: $Y_1, Y_2, \ldots, Y_n$ are independently distributed if and only if X has an exponential distribution.

A random variable U is said to have a uniform distribution on $[0, 1]$ if

$$P(U \leqq u) = u , \quad 0 \leqq u \leqq 1 .$$

The following theorem shows the relationship between a random variable with uniform distribution and a variable with exponential distribution.

Theorem 3.10: Let $0 < U < 1$ and $X = -\ln U$, so that $0 < X < \infty$. X has an exponential distribution with mean 1 if and only if U has a uniform distribution.

4. STANDARD DISTRIBUTIONS

In this section we will introduce the reader to some of the standard distributions used in statistics.

4.1 Chi-Square Distribution

Let X be a random variable distributed normally with mean 0 and variance unity; in other words let the probability density function of X be given by

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

Then it can be easily shown that the variable $Y = X^2$ has a probability density function given by

$$g(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} . \tag{3}$$

This $g(y)$ is referred to as the chi-square distribution with one degree of freedom. This is a special case of the gamma distribution (1) with $\beta = \frac{1}{2}$ and $\alpha = \frac{1}{2}$. The following is an important theorem concerning the gamma variables.

Theorem 4.1: Let $Y_1, Y_2, \ldots, Y_n$ be n independent gamma variables with a common scale parameter β and perhaps different values of α given by $\alpha_1, \alpha_2, \ldots, \alpha_n$. Then $Y_1 + Y_2 + \ldots + Y_n$ has a gamma distribution with the same scale parameter β and with $\alpha = \alpha_1 + \alpha_2 + \ldots + \alpha_n$.

From theorem 4.1 and (3), we can deduce the following useful corollary:

Corollary 4.1: Let $X_1, X_2, \ldots, X_n$ be n independent identically distributed normal variables with mean 0 and variance unity. Then $Y = X_1^2 + \ldots + X_n^2$ has a gamma distribution with $\alpha = n/2$ and $\beta = \frac{1}{2}$ (which is referred to as a chi-square distribution with n degrees of freedom) given by

$$f(y) = \frac{1}{2^{n/2}} \frac{1}{\Gamma(n/2)} y^{(n/2)-1} e^{-y/2} , \quad 0 < y < \infty . \tag{4}$$

4.2 t Distribution

It turns out that the probability density function of $R = \sqrt{Y/n}$ is given by

$$f(r) = \frac{2(n/2)^{n/2}}{\Gamma(n/2)} r^{n-1} e^{-nr^2/2} , \quad 0 < r < \infty , \tag{5}$$

which is referred to as the root-mean-square distribution with n degrees of freedom.

The variate T may be defined as the quotient of two independent variates X and R, where X has a normal distribution with mean 0 and variance unity and R is the root mean square of n other independent identically distributed normal variates with mean 0 and variance unity.

Theorem 4.2: T has the probability density function given by

$$f(t) = \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)\sqrt{n\pi}} \frac{1}{(1+t^2/n)^{(n+1)/2}}, \quad -\infty < t < \infty, \tag{6}$$

which is referred to as the t distribution with n degrees of freedom.

4.3 F Distribution

If Y has a chi-square distribution with n degrees of freedom, then $W = Y/n$ will have the probability density function given by

$$f(w) = \frac{(n/2)^{n/2}}{\Gamma(n/2)} w^{(n/2)-1} e^{-nw/2}, \quad 0 < w < \infty, \tag{7}$$

which is referred to as the mean-square distribution with n degrees of freedom.

The variate F may be defined as the quotient of the respective mean squares of a and b independent identically distributed normal variables with mean 0 and variance unity.

Theorem 4.3: F has the probability density function given by

$$\frac{a^{a/2} b^{b/2}}{\beta(a/2, b/2)} \frac{x^{(a/2)-1}}{(ax+b)^{(a+b)/2}}, \quad 0 < x < \infty. \tag{8}$$

This is referred to as the F distribution with a and b degrees of freedom. This F is also known as the variance ratio with a and b degrees of freedom.

5. EXPECTATION

5.1 Characteristic Functions

In this section we will introduce the notion of the characteristic function of a random variable, which can be used in determining the moments (whenever they exist) and which is helpful in proving the convergence of distribution functions. For brevity, we restrict ourselves to continuous variables. Let X be a random variable with probability density function $f(x)$. Then the characteristic function of the random variable X is defined by

$$\phi(t) = \int e^{itx} f(x)\, dx. \tag{9}$$

The following is a short table of familiar probability density functions and characteristic functions.

Distribution	Probability Density Function	Characteristic Function
Normal	$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\mu^2}$	$e^{i\mu t - \sigma^2 t^2/2}$
Laplace	$\frac{1}{2\sigma} e^{-\lvert x-\mu\rvert/\sigma}$	$e^{i\mu t}(1+\sigma^2 t^2)^{-1}$
Rectangular	$\frac{1}{\theta}\ (a < x < a+\theta)$	$e^{iat}(e^{i\theta t}-1)/it\theta$
Gamma	$\frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}\ (x > 0)$	$[\beta/(\beta - it)]^\alpha$
Beta	$\frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)} x^{m-1}(1-x)^{n-1}\ (0 < x < 1)$	$\frac{\Gamma(m+n)}{\Gamma(m)} \sum_{j=0}^{\infty} \frac{\Gamma(m+j)(it)^j}{\Gamma(m+n+j)\Gamma(j+1)}$
Cauchy	$\frac{\delta}{\pi} \frac{1}{\delta^2 + (x-\mu)^2}$	$e^{i\mu t - \lvert t\delta\rvert}$

5.2 Moments

The kth moment of the random variable X is defined by

$$E(X^k) = \int x^k f(x)\, dx \ . \tag{10}$$

For $k = 1$, $E(X)$ is referred to as the mean of the variable X. The variance of X is defined as

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 \ . \tag{11}$$

An alternate way of obtaining the kth moment of the variable X is to differentiate the characteristic function $\phi(t)$ k times and find its value at $t = 0$, that is,

$$E(X^k) = \phi^{(k)}(0) \ . \tag{12}$$

At times this may turn out to be an easy way of finding the kth moment, but not necessarily always. We will now give some formulas of moments for some well-known distributions.

1. Normal distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}\ ,\quad \sigma > 0\ ,\quad -\infty < x < \infty\ .$$

$$E(X^k) = \begin{cases} 1\cdot 3\cdot 5\cdot \ldots (k-1)\sigma^k & \text{for } k \text{ even} \\ 0 & \text{for } k \text{ odd}\ . \end{cases}$$

2. Rayleigh distribution:

$$f(x) = \frac{x}{\alpha^2} e^{-x^2/2\alpha^2} \quad , \quad 0 < x < \infty \ .$$

$$E(X^k) = \begin{cases} \sqrt{\frac{\pi}{2}}\, 1\cdot 3\cdot 5 \ldots k\alpha^k & \text{for } k \text{ odd} \\ 2^{k/2}\left(\frac{k}{2}\right)!\,\alpha^k & \text{for } k \text{ even} \ . \end{cases}$$

3. Maxwell distribution:

$$f(x) = \frac{\sqrt{2}}{\alpha^3\sqrt{\pi}} x^2 e^{-x^2/2\alpha^2} \quad , \quad 0 < x < \infty \ .$$

$$E(X^k) = \begin{cases} \sqrt{\frac{2}{\pi}}\, 2^{k+1/2}\left(\frac{k+1}{2}\right)!\,\alpha^k & \text{for } k \text{ odd} \\ 1\cdot 3\cdot 5 \ldots (k+1)\alpha^k & \text{for } k \text{ even} \ . \end{cases}$$

4. Gamma distribution:

$$f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad , \quad 0 < x < \infty \ .$$

$$E(X^k) = \alpha(\alpha+1)(\alpha+2)\ldots(\alpha+k-1)\beta^{-k} \ .$$

5. Beta distribution:

$$f(x) = \frac{\Gamma(m+n)}{\Gamma(m)\,\Gamma(n)} x^{m-1}(1-x)^{n-1} \quad , \quad 0 < x < 1 \ .$$

$$E(X^k) = \frac{m(m+1)\ldots(m+k-1)}{(m+n)(m+n+1)\ldots(m+n+k-1)} \ .$$

6. CONVERGENCE THEOREMS

In this section we will introduce the notions of convergence in probability, convergence in quadratic mean, convergence with probability 1, and convergence in distribution. We will point out the connections between these convergences. We will indicate the laws of large numbers and the central limit theorem.

6.1 Convergence of a Sequence of Random Variables

A sequence of random variables $\{X_n\}$, $n = 1, 2, \ldots$, is said to converge to a constant c:

1. in probability or weakly (and written $X_n \xrightarrow{P} c$), if for every given $\epsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - c| > \epsilon) = 0 \; ; \tag{13}$$

2. strongly or almost surely (written $\lim_{n \to \infty} X_n = c$ with probability 1 or $X_n \xrightarrow{\text{a.s.}} c$) if

$$P(\lim_{n \to \infty} X_n = c) = 1 \; , \tag{14}$$

or equivalently

$$\lim_{N \to \infty} P(\sup_{n \geqq N} |X_n - c| > \epsilon) = 0 \text{ for every } \epsilon ;$$

3. in quadratic mean (written $X_n \xrightarrow{\text{q.m.}} c$), if

$$\lim_{n \to \infty} E(X_n - c)^2 = 0 \; . \tag{15}$$

The following theorem shows the relationship among the various types of convergences.

Theorem 6.1:

1. If $X_n \xrightarrow{\text{q.m.}} c$, then $X_n \xrightarrow{P} c$.
2. If $X_n \xrightarrow{\text{a.s.}} c$, then $X_n \xrightarrow{P} c$.
3. If $X_n \xrightarrow{P} c$ and $\sum_{R=1}^{\infty} E(X_n - c)^2 < \infty$, then $X_n \xrightarrow{\text{a.s.}} c$.

6.2 Laws of Large Numbers

Theorem 6.2 (weak law of large numbers): Let $\{X_i\}$, $i = 1, 2, \ldots$, be a sequence of n independent and identically distributed random variables. Then

$$E(X_i) = \mu < \infty, \text{ then } \frac{(X_1 + \ldots + X_n)}{n} \xrightarrow{P} \mu \; .$$

Theorem 6.3 (strong law of large numbers I): Let $\{X_i\}$, $i = 1, 2, \ldots$, be a sequence of n independent random variables such that $E(X_i) = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2$. Then

$$\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty \text{ implies that } \frac{X_1 + \ldots + X_n}{n} - \frac{\mu_1 + \mu_2 + \ldots + \mu_n}{n} \xrightarrow{\text{a.s.}} 0 \; .$$

Theorem 6.4 (strong law of large numbers II): Let $\{X_i\}$, $i = 1, 2, \ldots$, be a sequence of n independent and identically distributed random variables. Then

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{\text{a.s.}} \mu \text{ if and only if } E(X_i) \text{ exists and is equal to } \mu \; .$$

6.3 Central Limit Theorem

Let $F_n(x)$ denote the distribution function of the random variable X_n, that is, $F_n(x) = P(X_n \leqq x)$.

Definition: The sequence of random variables $\{X_n\}$ is said to converge in distribution (or in law) to a random variable X with distribution function $F(x) = P(X \leqq x)$ if $F_n(x) \to F(x)$ as $n \to \infty$ at all points of continuity of F. Such a convergence is expressed as $X_n \xrightarrow{L} X$.

Theorem 6.5: Let $\phi_n(t)$ be the characteristic function of X_n and $\phi(t)$ that of X. Let $\phi(t)$ be continuous at the origin. Then

$$X_n \xrightarrow{L} X \text{ if and only if } \phi_n(t) \longrightarrow \phi(t) .$$

Theorem 6.6 (central limit theorem): Let $\{X_i\}$, $i = 1, 2, \ldots$, be a sequence of n independent and identically distributed random variables such that $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ exist. Then

$$\lim_{n \to \infty} P\left[\frac{\sqrt{n}}{\sigma}\left(\frac{X_1 + \ldots + X_n}{n} - \mu\right) \leqq x\right] = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt .$$

LECTURE III: CONTINGENCY TABLES

Marvin A. Kastenbaum

1. INTRODUCTION

Appended to these notes is a bibliography of papers on the subject of contingency tables that have appeared in the literature primarily during the past decade and a half. This chronological listing is by no means complete. It results from a partial literature search of the statistics journals in the ORNL library and includes only a small number of papers in the social sciences. Therefore, if you know of any well-known papers that have been omitted, kindly call them to my attention so that I may add them to the existing list.

My lecture will begin with some discussion of the underlying sampling distributions which go into the construction and formation of contingency tables. I will briefly review the history and theory of the χ^2 test criterion which is applied for testing hypotheses with contingency tables. I will touch on such questions as partitioning of χ^2, contingency tables of more than two dimensions, measure of association, missing and mixed-up values in contingency tables, contingency tables which come about as a result of consumer preference testing, Markov chains in contingency tables, the Bayesian approach to contingency tables, and alternative analyses to the χ^2 in contingency tables. I will present some illustrative examples at certain points throughout my discussion.

2. STRUCTURE OF 2×2 CONTINGENCY TABLES

The basic structure of contingency tables, especially those which we are most accustomed to encounter, is described with admirable lucidity by G. A. Barnard and E. S. Pearson in two papers which appear in Vol. 34 of *Biometrika* for 1947. Although these authors concentrate primarily on 2×2 contingency tables, their logical approach to the whole problem is stated in such elementary terms that both papers should be read by all statisticians. A systematic elaboration and application of the ideas of Barnard and Pearson appears in a series of papers by Roy, Mitra, and Kastenbaum in 1955 and 1956.

According to Barnard, the theory of statistical significance tests deals with abstraction of experimental results. The fact that the figures in question may happen to be tensile strength of iron bars, counts of radioactive particles, on-the-job accidents, or number of defectives is ignored in carrying out the test. For the purpose of the statistical theory, the experiment in question could just as well be represented by an experiment involving drawing balls from an urn. Once the abstract picture has been formed, the analysis of it is largely a matter of pure mathematics. As Barnard points out, "What distinguishes the statistician from the pure mathematician in this connection should be the statistician's ability to form valid abstract pictures of concrete cases and his clear recognition of the limits of validity of his abstract pictures."

Consider the simplest type of contingency table, namely the 2×2 or fourfold table, as Table 1*a* is sometimes referred to.

Table 1*a*: The 2×2 Contingency Table – Cell Frequencies

	I	II	Total
A	n_{11}	n_{12}	$n_{1.}$
B	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N

Table 1*b*: The 2×2 Contingency Table – Cell Probabilities

	I	II	Total
A	p_{11}	p_{12}	$p_{1.}$
B	p_{21}	p_{22}	$p_{2.}$
Total	$p_{.1}$	$p_{.2}$	1

This table displays a sample of total size N divided into four mutually exclusive and jointly exhaustive cells, so that $n_{11} + n_{12} + n_{21} + n_{22} = N$. Let us consider how such a table might have come about.

2.1 One Multinomial, or Two Variates (Responses)

A and B represent two shops or processes making the same articles. Roman numerals I and II represent defectives and nondefectives respectively. All articles produced in shops A and B are collected in a common bin, each of the articles having a small marking specific to the shop from which it came. On the basis of a sample of size N we wish to determine if the percentage defective in shop A is the same as it is in shop B. The sample of size N is taken from a common bin containing a large number of articles from both shops. If the true proportion of articles in category AI is p_{11} while p_{12}, p_{21}, and p_{22} represent the proportions of the other articles, then the probability associated with the table is given by the general term of the multinomial expansion of $(p_{11} + p_{12} + p_{21} + p_{22})^N$, that is,

$$P = \frac{N!}{n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!} p_{11}^{n_{11}} p_{12}^{n_{12}} p_{21}^{n_{21}} p_{22}^{n_{22}}, \tag{2.1.1}$$

where $p_{11} + p_{12} + p_{21} + p_{22} = 1$.

2.2 Two Binomial Samples, or One Variate (Response) and One Way of Classification (Factor)

Perhaps a more realistic way in which our table may be constructed under a similar set of circumstances would be if $n_{1.}$ articles from shop A and $n_{2.}$ articles from shop B were examined for the proportion of defectives. In this case, of course, the articles which had been manufactured would not be taken from the same bin, but rather shop A would have its bin and shop B would have its bin.

If the probability that shop A will produce defectives is given by p_{11} and the probability that shop B will produce defectives is given by p_{21}, then the probability of the arrangement in the table is equal to the probability of finding n_{11} defectives from a sample of size $n_{1.}$ in shop A and n_{21} defectives in a sample of size $n_{2.}$ in shop B.

This probability may be written as the general term of the product of two binomials, namely

$$P = \frac{n_{1.}!}{n_{11}!\,n_{12}!}\, p_{11}^{n_{11}}\, p_{12}^{n_{12}}\, \frac{n_{2.}!}{n_{21}!\,n_{22}!}\, p_{21}^{n_{21}}\, p_{22}^{n_{22}}\,, \tag{2.2.1}$$

where $p_{12} = 1 - p_{11}$, $p_{22} = 1 - p_{21}$, $n_{1.} = n_{11} + n_{12}$, and $n_{2.} = n_{21} + n_{22}$.

2.3 One Hypergeometric, or Two Ways of Classification

A third way in which the data in our table might occur is perhaps the most artificial of the three structures which we will discuss. Its artificiality comes not from the fact that such a structure does not occur in real life. Rather, such a structure occurs less frequently than the two structures already discussed. The classical example of such a structure occurs in the "lady tasting tea" experiment proposed by Sir Ronald Fisher. In this case we not only fix the total sample size, but we also fix the values in all the marginals. How might this come about?

Here the experimenter has two types of objects of known frequencies, let us say $n_{1.}$ and $n_{2.}$. He tells his subject that there are among the N objects which he is presenting to him $n_{1.}$ of type A and $n_{2.}$ of type B. He then asks his subject to identify these objects accordingly. The probability that the subject will place n_{11} observations in the first cell given that the marginal totals are fixed in this way is given by the underlying hypergeometric distribution as follows:

$$P = \frac{n_{1.}!\,n_{2.}!\,n_{.1}!\,n_{.2}!}{N!\,n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!} = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{21}}}{\binom{N}{n_{.1}}}\,, \tag{2.3.1}$$

where $n_{.1} = n_{1.}$ and $n_{.2} = n_{2.}$.

The three underlying distributions represented by Eqs. (2.1.1), (2.2.1), and (2.3.1) account for almost all contingency tables which are encountered in experimental situations.

3. DEGREES OF FREEDOM

At this point, it might be appropriate to spend a little time discussing the degrees of freedom which are associated with tests of hypotheses concerning the data in contingency tables. The rule for determining the number of degrees of freedom associated with any contingency table of whatever structure is: the number of degrees of freedom is equal to the total number of cells in the table, minus the number of independent linear constraints on the observations, minus the number of free parameters to be estimated from the data. In the case of a 2×2 table consider the situation for the three structures which we discussed above.

3.1 One Multinomial

The total number of cells is 4, the number of independent parameters to be estimated from the data is 2, and the number of linear constraints on the observations is 1. How does this come about? The sum of the p_{ij}'s, $p_{11} + p_{12} + p_{21} + p_{22}$, is 1. This is true. However, these four p_{ij}'s are not independent of one another because they are subject to the two following constraints: $p_{1.} + p_{2.} = 1$ and $p_{.1} + p_{.2} = 1$, so that there are only two independent parameters $p_{1.}$(or $p_{2.}$) and $p_{.1}$ (or $p_{.2}$) to be estimated here. The one linear constraint on the observations is that $n_{11} + n_{12} + n_{21} + n_{22} = N$.

3.2 Two Binomials

In the case of two binomials there is only one independent parameter to be estimated from the data, and there are two linear constraints on the observations. The independent parameter comes from the fact that one set of marginal totals is fixed, leaving no parameters to be estimated here. However, the other set of marginal totals is variable and must be estimated subject to the constraint $p_{.1} + p_{.2} = 1$. The two linear constraints are $n_{11} + n_{12} = n_{1.}$ and $n_{21} + n_{22} = n_{2.}$.

3.3 One Hypergeometric

In this case there are no independent parameters to be estimated from the data. Instead there are four constraints on the data, three of which are independent. The values in the two rows and two columns are constrained by their respective totals, but the fact that both sets of marginal totals are also constrained to add to N reduces the number of independent constraints to 3.

In all three cases we have four cells and a combination of three linear constraints or independent parameters to be estimated from the data. Thus the number of degrees of freedom is equal to *one* in all three cases. This is one very interesting result in the two-way table. Another pertinent and interesting fact which we may refer to now in passing is that in the two-way table all three structures result in the same large-sample χ^2 test of the null hypothesis. These two results, which are perhaps a bit of mathematical good fortune, may be the cause of some of the confusion associated with the interpretation of data in two-way tables.

4. STRUCTURE OF $r \times s$ CONTINGENCY TABLES

Let $i = 1, 2, \ldots, r$ designate rows and $j = 1, 2, \ldots, s$ designate columns.

4.1 One Multinomial

$$\sum_{i=j}^{r} \sum_{j=1}^{s} n_{ij} = N; \quad \sum_{i=1}^{r} \sum_{j=1}^{s} p_{ij} = \sum_{i=1}^{r} p_{i.} = \sum_{j=1}^{s} p_{.j} = 1 .$$

$$P = \frac{N!}{\prod_{i=1}^{r} \prod_{j=1}^{s} n_{ij}!} \prod_{i=1}^{r} \prod_{j=1}^{s} p_{ij}^{n_{ij}} . \tag{4.1.1}$$

Number of degrees of freedom: $rs - [(r-1) + (s-1)] - 1 = (r-1)(s-1)$.

4.2 r Multinomial Samples

$$\sum_{j=1}^{s} n_{ij} = n_{i.} \text{ for all } i; \quad \sum_{i=1}^{r} n_{i.} = N; \quad \sum_{j=1}^{s} p_{.j} = 1 .$$

$$P = \prod_{i=1}^{r} n_{i.}! \prod_{j=1}^{s} \frac{p_{ij}^{n_{ij}}}{n_{ij}!} . \tag{4.2.1}$$

Number of degrees of freedom: $rs - (s-1) - r = (r-1)(s-1)$.

4.3 One Hypergeometric

$$\sum_{j=1}^{s} n_{ij} = n_{i.}, \quad \sum_{i=1}^{r} n_{ij} = n_{.j}, \quad \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} n_{.j} = N .$$

$$P = \frac{\prod_{i=1}^{r} n_{i.}! \prod_{j=1}^{s} n_{.j}!}{N! \prod_{i=1}^{r} \prod_{j=1}^{s} n_{ij}!} . \tag{4.3.1}$$

Number of degrees of freedom: $rs - (r + s - 1) = (r-1)(s-1)$.

5. THE NULL HYPOTHESIS IN THE TWO-WAY TABLE

5.1 One Multinomial

The composite hypothesis we shall be interested in testing is that the two variates are independent, that is, H_0: $p_{ij} = p_{i.}\, p_{.j}$, against the alternative $H \neq H_0$, where $p_{i.}$ and $p_{.j}$ (for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, s$) are arbitrary positive parameters subject to the constraints

$$\sum_i p_{i.} = \sum_i p_{.j} = 1.$$

This is analogous to the hypothesis of no correlation ($\rho_{xy} = 0$) in a bivariate normal population. Under H_0 the likelihood function is given by

$$P_0 = \frac{N!}{\prod_{i=1}^{r} \prod_{j=1}^{s} n_{ij}!} \prod_{i=1}^{r} p_{i.}^{n_{i.}} \prod_{j=1}^{s} p_{.j}^{n_{.j}} \; . \tag{5.1.1}$$

5.2 r Multinomials

We have r independent samples of fixed sizes $n_{1.}, n_{2.}, \ldots, n_{r.}$, with p_{ij}, the probability of an observation in cell (ij), such that

$$\sum_j p_{ij} = p_{i.} = 1.$$

The composite hypothesis of interest is that p_{ij}, for any j, is independent of i; that is, the probability of an observation being in the jth position of row i is the same for all i. This may be specified as

$$H_0\text{:}\ \ p_{ij} = p_{.j} \ \text{ against } H \neq H_0,$$

where $p_{.j}$ ($j = 1, \ldots, s$) is an arbitrary positive parameter subject to the constraint

$$\sum_j p_{.j} = 1.$$

This is referred to as the hypothesis of "homogeneity." It is analogous to and a generalization of the hypothesis of equality of means for r homoscedastic univariate normal populations. For random samples from normal populations, $N(\mu_i, \sigma^2)$, $i = 1, 2, \ldots, r$, the usual hypothesis of interest is:

$$H_0\text{:}\ \ \mu_1 = \mu_2 = \ldots = \mu_r = \mu \ ,$$

$$H\text{:}\ \ \mu_i \neq \mu \ .$$

The standard test of this hypothesis is the F test of equality of means.

The likelihood function under the hypothesis of homogeneity of r multinomial samples is

$$P_0 = \frac{\prod_{i=1}^{r} n_{i.}! \prod_{i=1}^{s} p_{.j}^{n_{.j}}}{\prod_{i=1}^{r} \prod_{j=1}^{s} n_{ij}!} . \tag{5.2.1}$$

5.3 One Hypergeometric

In this case, the likelihood function given by (4.3.1) is unchanged, and the null hypothesis is one of "independence" or "inability to discriminate." The question posed under this structure is specific, as in the case of the "lady tasting tea" experiment.

Note that if we were to start with (4.1.1) and let H_0: $p_{ij} = p_{i.}\, p_{.j}$, and then find the conditional probability under H_0 subject to $n_{i.}$ and $n_{.j}$ being fixed, this probability would be the hypergeometric given by (4.3.1).

6. TESTS OF HYPOTHESES IN THE TWO-WAY TABLE

In the year 1900, Karl Pearson proposed as a criterion for testing hypotheses, such as those mentioned above, the statistic

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - Np_{ij})^2}{Np_{ij}} . \tag{6.1}$$

Pearson suggested that in the limit, as N becomes large, this statistic has the χ^2 distribution with $rs - 1$ degrees of freedom. He further suggested that the number of degrees of freedom remains unchanged when the p_{ij} are estimated from the data. This we now know is wrong. Fisher (1922, 1924) pointed out Pearson's error and went on to give a proof of the limiting distribution of X^2 which avoids most of the mathematical complexities in Pearson's proof. A fully rigorous proof of the limiting distribution of X^2 is given by Cramer (1946). Cramer assumes, among other things, that the p_{ij} are estimated by the method of maximum likelihood.

The maximum likelihood estimates of p_{ij} under the null hypothesis are found to be $p_{ij} = n_{i.}\, n_{.j}/N^2$, so that the usual form of this statistic for contingency tables is

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \dfrac{n_{i.}\, n_{.j}}{N}\right)^2}{\dfrac{n_{i.}\, n_{.j}}{N}} . \tag{6.2}$$

This statistic, we say, is distributed in the limit, as N becomes large, as χ^2 with $(r - 1)(s - 1)$ degrees of freedom.

The most pertinent question at this point is, "Is the same χ^2 test to be used for all three structures?" The answer is, "Yes, when N is large." However, this is not so for small samples. When the sample size N is small, Fisher (1934) recommends that the exact probabilities obtained from the hypergeometric distribution (4.3.1) be used. The greatest objection to Fisher's recommendation is that a loss of power may result if the hypergeometric probabilities are computed when the data actually arise from a single or from several multinomial samples.

K. D. Tocher (1950) has proposed a modification of Fisher's exact test that is most powerful, in the Neyman-Pearson sense, for one-tailed tests with data from any of the three structures. The modification is best illustrated by an example.

Consider the hypothetical example (Table 2) involving the frequency of failure due to cracking of specimens in 30-day tests on 24 large industrial boilers. The observations are classified as cracked and uncracked, and also according to the addition or nonaddition of tannin to the feedwater.

Table 2. Hypothetical Example

	Uncracked	Cracked	Total
+Tannin	9	2	11
−Tannin	7	6	13
Total	16	8	24

The null hypothesis is that the failure of test specimens is not influenced by the addition of tannin to the boiler feedwater. Given these data, we wish to make a one-tailed test at the 5% level. The two possible sets of data which deviate more from the null hypothesis are

$$\begin{array}{ccc} 10 & 1 & 11 \\ 6 & 7 & 13 \\ \hline 16 & 8 & 24 \end{array} \quad \text{and} \quad \begin{array}{ccc} 11 & 0 & 11 \\ 5 & 8 & 13 \\ \hline 16 & 8 & 24 \end{array} .$$

In Fisher's exact test, we add the probabilities of the three tables as computed from the hypergeometric. This gives

$$\frac{8!\ 11!\ 13!\ 16!}{24!}\left[\frac{1}{2!\ 6!\ 7!\ 9!}+\frac{1}{6!\ 7!\ 10!}+\frac{1}{5!\ 8!\ 11!}\right]$$

$$= 0.12833 + 0.02567 + 0.00175$$

$$= 0.15575.$$

This value is regarded as the significance probability.

In Tocher's modification we compute the total probability of all more extreme cases, that is,

$$0.02567 + 0.00175 = 0.02742 .$$

If the numbers 0.15575 and 0.02742 are both *below* the stated significance level 0.05, reject the hypothesis. If they are both *above* 0.05, accept it. If one is above and one is below, as in this example, we calculate

$$\frac{0.05 - 0.02742}{0.12833} = 0.17595 .$$

Now, draw a random number between 0 and 1. If this number is less than 0.27595, we reject; if greater, we accept. (Rationale: If H_0 is rejected only when the two most extreme cases occur, the significance level is 0.02742. The third most extreme case, represented by the data, occurs with probability 0.12833. Tocher's modification declares as "significant" a fraction, 0.17595 of the cases in which the observed data are encountered.) The Pearson chi-square statistic for these data is $X^2 = 2.098$. Inasmuch as the square root of a chi-square variable with one degree of freedom is distributed as a normal variable with zero mean and unit variance, the probability is $P(\chi^2 \geqq 2.098) = 2P(\chi \geqq 1.448) = 0.148$. For comparison with the exact procedure we must consider only one tail of the normal distribution, so that $P(\chi \geqq 1.448) = 0.074$.

6.1 Yates' Correction for Continuity

The distribution of the Pearson X^2 statistic is discontinuous. When all expectations are small the chi-square approximation may be poor. The correction proposed by Yates (1934) amounts to reading the chi-square table not at X_0^2 but at a point between X_0^2 and the value of X^2 immediately below X_0^2 in the discrete series of values. The formula for the 2×2 table is well known and need not be repeated here. If it is applied to the data in Table 2, the resulting statistic is $X_c^2 = 1.028$, and the probability is $P(\chi^2 \geqq 1.028) = 2P(\chi > 1.014) = 0.310$. The corresponding probability for the one-tailed test is $P(\chi > 1.014) = 0.155$. The following table gives comparable probabilities for a one-tailed test of the data in Table 2:

Test	One-Tailed Probability
Fisher exact	0.156
X^2 uncorrected	0.074
X_c^2 corrected	0.155

Caution: If a number of X^2 values, each with one degree of freedom, are added to form a total X^2, the individual values should *not* be corrected for continuity. The total X^2 may be corrected, after it has been obtained, by a procedure given by Cochran (1952). Also see Pasternack (1966) and Grizzle (1967).

6.2 Cochran's Recommendations Concerning Analyses of Two-Way Tables

1. The 2×2 table: For $N < 20$, or $20 < N < 40$ and the smallest expectation less than 5, use Fisher's exact test. For $N > 40$ use X^2, corrected for continuity if the smallest expectation is less than 5.
2. Tables with degrees of freedom between 2 and 60 and *all* expectations less than 5: For small N, use Fisher's exact test. Otherwise use X^2, considering whether the continuity correction is needed.
3. Tables with degrees of freedom more than 60 and *all* expectations less than 5: Try to obtain the exact mean and variance of X^2, and use the normal approximation to the exact distribution. See Haldane (1937).
4. Tables with more than one degree of freedom and *some* expectations greater than 5: Use X^2 uncorrected for continuity.
5. Continuous data: To test goodness of fit, group the data, using enough cells to keep the expectations down to 12 per cell for $N = 200$, 20 per cell for $N = 400$, 30 per cell for $N = 1000$. Pool at the tails so that the minimum expectation is no smaller than 1.

7. ALTERNATIVE ANALYSES TO THE TRADITIONAL CHI-SQUARE

In recent years many authors have discussed analyses of contingency tables other than the traditional Pearson chi-square analysis. These alternatives fall into two major categories, namely, (1) the logit transformation in conjunction with standard least squares and (2) the likelihood ratio test.

7.1 The Logit Transformation

This method was proposed by Woolf (1955) and elaborated on by Plackett (1962), Gart (1962), Goodman (1963, 1964), Lindley (1964), and others. It may be demonstrated for the $2 \times s$ table.

Let n_{ij} be the observation in the ith row and jth column of a $2 \times s$ contingency table, where $i = 1, 2$ and $j = 1, 2, \ldots, s$. If we take

$$x_j = n_{1j}/n_{2j}, \; z_j = \ln x_j, \text{ and } \frac{1}{u_j} = \frac{1}{n_{1j}} + \frac{1}{n_{2j}},$$

then

$$X^2 = \sum_{j=1}^{s} u_j(z_j - \bar{z})^2 \tag{7.1.1}$$

is distributed asymptotically as χ^2 with $s - 1$ degrees of freedom, where

$$\bar{z} = \frac{\sum_{j=1}^{s} u_j z_j}{\sum_{j=1}^{s} u_j} . \tag{7.1.2}$$

Equation (7.1.1) may also be written as

$$X^2 = \sum_{j=1}^{s} u_j z_j^2 - \frac{\left(\sum_{j=1}^{s} u_j z_j \right)^2}{\sum_{j=1}^{s} u_j} . \tag{7.1.3}$$

Example: The data in Table 3 are taken from Table 9.54 on page 292 of O. L. Davies' text *Statistical Methods in Research and Production*. "These data show the number of times piston rings have failed in each leg (North, Centre, South) of two groups of compressors at an I.C.I. factory. The compressors are apparently identical and are oriented the same way in the Compressor House. Each leg consists of two cylinders arranged vertically: the lower cylinder deals with the first stage of compression, and the upper cylinder with the second stage. The South leg is, in every case adjacent to the drive." Is the probability of failure of a piston ring independent of compressors and location in compressor (leg)?

Table 3. Failure of Piston Rings by Location and Compressors

Compressor Group	North Leg	Center Leg	South Leg	Total
1	17	17	12	46
2	36	24	60	120
Total	53	41	72	166

Arithmetic necessary for calculating logit test statistics:

$$x_1 = \frac{17}{36}; \; x_2 = \frac{17}{24}; \; x_3 = \frac{12}{60}.$$

$$z_1 = -0.75031; \; z_2 = -0.34494; \; z_3 = -1.60943 .$$

$$\frac{1}{u_1} = \frac{1}{17} + \frac{1}{36} = 0.08660; \; \frac{1}{u_2} = \frac{1}{17} + \frac{1}{24} = 0.10049; \; \frac{1}{u_3} = \frac{1}{12} + \frac{1}{60} = 0.1 .$$

$$u_1 = 11.54736; \; u_2 = 9.95124; \; u_3 = 10.0 .$$

$$\sum_{j=1}^{3} u_j = 31.499; \quad \sum_{j=1}^{3} u_j z_j = -28.191; \quad \sum_{j=1}^{3} u_j z_j^2 = 33.587 .$$

Logit $X^2 = 8.357$; Pearson $X^2 = 8.760$.

7.2 The Likelihood-Ratio Test

This test, proposed by Wilks (1935), has been investigated by Woolf (1957), Chakravarti and Rao (1959), and Kullback, Kupperman, and Ku (1962). In their 1962 papers, Kullback, Kupperman, and Ku resort to an information theory approach and define a minimum discrimination information statistic (M.D.I.S.) with the following properties:

1. distributed asymptotically as chi-square under the null hypothesis and as noncentral chi-square under the alternative hypothesis, with appropriate degrees of freedom and noncentrality parameter,
2. additivity,
3. convexity.

This statistic is $2\hat{I} = -2 \ln \lambda$, where λ is the likelihood ratio. For the $r \times s$ contingency table,

$$2\hat{I} = \sum_{i=1}^{r} \sum_{i=j}^{s} 2n_{ij} \ln n_{ij} + 2N \ln N - \sum_{i=1}^{r} 2n_{i.} \ln n_{i.} - \sum_{j=1}^{s} 2n_{.j} \ln n_{.j} . \tag{7.2.1}$$

If this appears to be a formidable alternative to the Pearson X^2 statistic, the authors give us reassurance by presenting tables of $2y \ln y$ for values of y from 1 to 10,000.

Applying the likelihood ratio approach to the data in Table 3 yields

$$\sum_{i=1}^{2} \sum_{j=1}^{3} 2n_{ij} \ln n_{ij} = 1154.177; \quad 2N \ln N = 1697.180 ;$$

$$\sum_{i=1}^{2} 2n_{i.} \ln n_{i.} = 1501.233; \quad \sum_{j=1}^{3} 2n_{.j} \ln n_{.j} = 1341.204 .$$

The three comparable test statistics are

$$\begin{aligned} & 2\hat{I} = 8.920, \\ \text{logit} \quad & X^2 = 8.357, \\ \text{Pearson} \quad & X^2 = 8.760. \end{aligned}$$

In each case the probability of observing a value of chi-square greater than X^2 is $P(\chi^2 > X^2) < 0.02$. If we were testing at a 5% significance level, we would reject the hypothesis of independence between compressor groups and location of the leg. That is to say, using any of these procedures we would conclude that the performance of the two compressor groups is not the same at all the legs.

8. PARTITIONING OF CONTINGENCY TABLES

Assume, for one moment, that in the piston-ring example there was reason to believe, prior to experimentation, that the performance of the compressor groups is the same at the north and center leg but different at the south leg. Can we test these hypotheses using the same set of data?

This question often arises in the analysis of variance, where tools are available for partitioning the sum of squares associated with the test of a hypothesis concerning the equality of s means into at most $s - 1$ orthogonal sums of squares. With each of the $s - 1$ sums of squares so derived a hypothesis may be tested. Moreover, the total sum of squares with $s - 1$ degrees of freedom is equal to the sum of all its orthogonal parts.

Applications of an analogous technique to contingency tables are given by Lancaster (1949), Irwin (1949), Kimball (1954), and Kastenbaum (1960). The test statistic proposed by these authors is the standard Pearson X^2. Kullback, Kupperman, and Ku (1962) show how the same additive partitions may be tested with the minimum discrimination information statistic.

Example: Partition of data in Table 3.

Compressor Group	North Leg	Center Leg	Total
1	17	17	34
2	36	24	60
Total	53	41	94

Compressor Group	North and Center Legs	South Leg	Total
1	34	12	46
2	60	60	120
Total	94	72	166

Pearson X^2:

$$X_1^2 = \frac{(166)^2}{(46)(120)}\left[\frac{(17)^2}{53} + \frac{(17)^2}{41} - \frac{(34)^2}{94}\right] = 1.017 \quad (p > 0.3) ,$$

$$X_1^2 = \frac{(166)^2}{(46)(120)}\left[\frac{(34)^2}{94} + \frac{(12)^2}{72} - \frac{(46)^2}{166}\right] = 7.743 \quad (p < 0.01) ,$$

$$1.017 + 7.743 = 8.760 = \text{Pearson's } X^2 .$$

M.D.I.S.:

$$2\hat{I}_1 = 0.879 \quad (p > 0.3), \quad 2\hat{I}_1 = 8.041 \quad (p < 0.01),$$

$$0.879 + 8.041 = 8.920 = 2\hat{I} .$$

In each case we would conclude that performance of the compressor groups was the same for the north and center legs but different at the south leg.

Alternatively, the contingency table may be partitioned into the same component parts, with the standard Pearson X^2 analysis on each part. This procedure will not result in additive X^2. However, this approximate partition is adequate for most tests of significance. Moreover, it has *not* been shown that the additive partition is really preferable to the approximate partition in small samples. This property of additive partitions, in fact, may be the principal motivation for much of the work on contingency tables which has appeared in the literature in the past 15 years. Consider only that statisticians have known for a long time about a random variable whose distribution may be specified by the chi-square probability density function, and that the sum of such random variables is also distributed as chi-square. Add to this the knowledge that a test statistic exists which is, at least asymptotically, distributed as chi-square. Immediately you stimulate the desire to add things up or to separate things into their component parts.

9. CONTINGENCY TABLES OF MORE THAN TWO DIMENSIONS

Except for some brief references in two or three statistical texts, the subject of multidimensional contingency tables was all but ignored until 15 years ago. Indeed, in his text which was published in 1954, O. L. Davies raises and summarily dismisses the three-dimensional case as follows: "Such examples may be treated by an extension of the methods already explained which the reader will have no difficulty in making if he has understood the principles." Unfortunately, things are not as simple as Davies indicates they might be.

The transition from two dimensions to three dimensions necessitates a full understanding of the underlying structure of the data, a clear and concise idea of what null hypotheses are to be tested, and knowledge of the appropriate estimators associated with these null hypotheses in order to calculate the corresponding test statistic. Contrary to Davies' belief, new conceptual problems are posed in going from two-way to three-way tables. On the other hand, the extension from three- to higher-dimensional contingency tables does not pose any new problems.

The theory of multidimensional contingency tables is presented in the many references cited in the bibliography. I shall briefly present some of the highlights and indicate some of the more important references. Notation:

Let n_{ijk} denote the observed frequency and p_{ijk} the probability of having an observation in cell (ijk) of a three-way table, where $i = 1, 2, \ldots, r$ designates rows, $j = 1, 2, \ldots, s$ designates columns, and $k = 1, 2, \ldots, t$ designates layers. Also, let the marginal frequencies be denoted by

$$n_{.jk} = \sum_{i=1}^{r} n_{ijk}; \quad n_{..k} = \sum_{i}^{r}\sum_{j}^{s} n_{ijk} = \sum_{j}^{s} n_{.jk} = \sum_{j}^{r} n_{i.k};$$

$$n_{i.k} = \sum_{j=1}^{s} n_{ijk}; \; n_{.j.} = \sum_{i}^{r}\sum_{k}^{t} = \sum_{i}^{r}\sum_{k}^{t} n_{ijk} = \sum_{i}^{r} n_{ij.} = \sum_{k}^{t} n_{.jk};$$

$$n_{ij.} = \sum_{k=1}^{t} n_{ijk}; \; n_{i..} = \sum_{j}^{s} \sum_{k}^{t} n_{ijk} = \sum_{j}^{s} n_{ij.} = \sum_{k}^{t} n_{i.k};$$

$$N = \sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{t} n_{ijk} .$$

For a single multinomial sample of size N, corresponding summations over the p_{ijk} are similarly denoted, and

$$\sum_{i=1}^{r} \sum_{j=1}^{s} \sum_{k=1}^{t} p_{ijk} = 1 .$$

9.1 Hypotheses Concerning the Two-Way Marginals of a Three-Way Table

Pairwise independence:

$$H_0: \; p_{ij.} = p_{i..} \, p_{.j.} \quad \text{(independence of } I \text{ and } J\text{)} , \tag{9.1.1}$$

$$H_0: \; p_{i.k} = p_{i..} \, p_{..k} \quad \text{(independence of } I \text{ and } K\text{)} , \tag{9.1.2}$$

$$H_0: \; p_{.jk} = p_{.j.} \, p_{..k} \quad \text{(independence of } J \text{ and } K\text{)} . \tag{9.1.3}$$

9.2 Hypotheses Which Have No Analog in a Two-Way Table

Conditional independence (partial independence):

$$H_0: \frac{p_{ijk}}{p_{..k}} = \frac{p_{i.k}}{p_{..k}} \cdot \frac{p_{.jk}}{p_{..k}}, \text{ or } p_{ijk} = \frac{p_{i.k} \, p_{.jk}}{p_{..k}} . \tag{9.2.1}$$

Equation (9.2.1) is the hypothesis of conditional independence of I and J given K. This condition is analogous to the hypothesis of zero partial correlation between I and J in a three-variate normal population. It does not imply the independence of I and K or of J and K. However, if I is independent of K and J is independent of K, and if (9.2.1) also holds, we have the condition of mutual independence:

Mutual independence:

$$H_0: \; p_{ijk} = p_{i..} \, p_{.j.} \, p_{..k} . \tag{9.2.2}$$

Multiple independence:

$$H_0: \; p_{ijk} = p_{ij.} \, p_{..k} \quad \text{(independence of } K \text{ and } IJ\text{)} . \tag{9.2.3}$$

This condition is analogous to the hypothesis of zero multiple correlation in a three-variate normal population. It implies independence between I and K and between J and K. The converse, how-

ever, is not true. That is to say, (9.2.3) implies (9.1.2) and (9.1.3), but (9.1.2) and (9.1.3) do not imply (9.2.3).

It was at this point in late 1954 that Professor Roy and I found a need for additional theoretical concepts. For, in the special case of the multivariate *normal* population, not only does zero multiple correlation imply zero correlation between all pairs of variables, but also conversely. Obviously this did not hold for a three-variate contingency table. Therefore the question we posed to ourselves was, "What set of conditions exists which, when superimposed on the conditions of independence between pairs of variables, will jointly yield the condition of multiple independence?" The answer to this question was:

"*No three-factor interaction*":

$$H_0: \; p_{rst}\, p_{ijt}\, p_{isk}\, p_{rjk} = p_{ist}\, p_{rjt}\, p_{rsk}\, p_{ijk} \,. \tag{9.2.4}$$

Equation (9.2.4) is a generalization of the condition proposed by Bartlett in 1935 for the $2 \times 2 \times 2$ and the $2 \times 2 \times 3$ tables.

There has been considerable discussion in the literature of the past ten years concerning this hypothesis. For a summary of the theory and philosophy concerning some of the hypotheses which may be tested in a three-way table, the reader is referred to the paper by B. N. Lewis (1962). This paper also gives techniques for analysis of data in multidimensional contingency tables. In addition the recent work of Bhapkar and Koch (1961, 1965, 1966) suggests that other hypotheses might be more relevant and appropriate for certain types of contingency tables. For detailed reading on the subject of multiway contingency tables see Bartlett (1935), Norton (1945), Simpson (1951), Lancaster (1951), Roy and Kastenbaum (1956), Roy and Mitra (1956), Kastenbaum and Lamphiear (1959), Lancaster (1960), Darroch (1962), Kullback, Kupperman, and Ku (1962), Plackett (1962), Birch (1963), Goodman (1963), and Goodman (1964, *Ann. Math. Statist.;* 1964, *J. Am. Statist. Assoc.*). For a numerical demonstration of the techniques (other than higher-order interactions) of analyzing data in multidimensional contingency tables, see the paper by P. N. Ries and Harry Smith (1963).

Some of the difficulties which O. L. Davies chose to dismiss in his statement which I quoted earlier may be pointed up in remarks such as this one, extracted from Goodman (1964, *J. Am. Statist. Assoc.*): "In 1951 Lancaster suggested, on heuristic grounds, a rather simple procedure for testing the hypothesis of zero three-factor interaction. Unfortunately, the distribution of the test statistic suggested by Lancaster is not necessarily distributed as supposed (i.e., as chi-square asymptotically) . . . . Lancaster's statement which has been quoted by Kendall and Stuart (Vol. 2, 1961, p. 584) and by Snedecor (1958) that this test and Bartlett's are "asymptotically equal" is in error."

10. SPECIAL TOPICS IN CONTINGENCY TABLES

10.1 Missing and "Mixed-Up" Values

The problem of contingency tables with missing or misclassified values has been considered by Bross (1954) and Watson (1956). In his paper, Watson presents an iterative procedure, similar to the "missing-plot" technique in analysis of variance, for estimating the missing values in the contingency table prior to performing the standard Pearson X^2 test. Kastenbaum (1958) demonstrated that explicit algebraic formulas can be found for certain sets of missing values in contingency tables. This problem has been studied, elaborated on, and generalized by Goodman (1968), Bishop (1968), Bishop and Fienberg (1968), and Caussinus (1964, 1965, 1966). Other recent investigations in this area are Asano (1965) and Mote and Anderson (1965).

10.2 Consumer Preference

An interesting structure for contingency tables is discussed by Anderson (1959) in a problem involving consumer preference studies. One lot of each of three varieties (v_1, v_2, v_3) of snap beans was displayed in retail stores, and each of 123 consumers was asked to rank the beans according to first, second, and third choices. The actual data are given in Table 4. The question is, "Does the usual X^2 test of independence of ranks and varieties with four degrees of freedom apply?" That is, does each variety have the same chance ($\frac{1}{3}$) of receiving a given rank, regardless of rank? This is not the usual problem of a contingency table with fixed border totals, because repeated sampling is not a random rearrangement of 3×123 items subject to border restrictions. For $i = 1, 2, \ldots, r$ varieties and $j = 1, 2, \ldots, r$ ranks, Anderson shows that the appropriate test statistic for n consumers is

$$\frac{r-1}{r} Q^2 = \frac{r-1}{n} \sum_{i=1}^{r} \sum_{j=1}^{r} \left(n_{ij} - \frac{n}{r}\right)^2 , \tag{10.2.1}$$

which is distributed asymptotically as χ^2 with $(r-1)^2$ degrees of freedom.

Table 4. Consumer Rankings of Three Varieties of Snap Beans

Variety	Rank 1	Rank 2	Rank 3	Total
v_1	42	64	17	123
v_2	31	16	76	123
v_3	50	43	30	123
Total	123	123	123	369

10.3 Markov Chains

The structures for contingency tables which we have discussed involve some assumptions of independence of successive sample observations. There frequently arise practical situations in which these assumptions are not valid. One class of such situations involves dependent observations resulting from realizations of states of a simple stationary Markov chain (Billingsley, 1961). In this situation the matrix of transition probabilities is given by a stochastic matrix which is square and whose row totals add to 1. The corresponding frequencies form a square contingency table. Analyses of such data, including tests of hypotheses of a specified matrix of transition probabilities, Markovity, and homogeneity of several realizations of Markov chains are given by Kullback, Kupperman, and Ku (1962). Tests auxiliary to χ^2 in Markov chains are given by Gold (1963).

10.4 Measures of Association

The two areas of study covered by the broad title of "statistical inference" are hypothesis testing and estimation. I have devoted all my time to the former and none to the latter. In contingency tables the lack of independence should give some indication of the degree of association. Measures of association in contingency tables have been proposed for at least as long as tests of hypotheses. Indeed they have been used and abused rather widely in the social sciences.

I shall not go into this area in detail. The series of papers by Kruskal and Goodman (*J. Am. Statist. Assoc.* 1954, 1959, 1963) on measures of association for cross classification and the recent papers by Goodman (1963) on interactions in multidimensional contingency tables provide an excellent history and summary of the subject. The most recent paper on this subject is by Mosteller (1968).

10.5 Bayesian Analysis

Finally, as if not to be outdone, contingency tables have most recently been given consideration by the Bayesian statisticians. Lindley (1964) and Good (1965) describe how data from a multinomial distribution and, in particular, data in the form of a contingency table may be studied using a prior distribution of parameters and expressing the results in the form of a posterior distribution of the parameters. Gart (1966) shows how a Bayesian argument may be used for choosing a critical region for the exact test.

RECENT BIBLIOGRAPHY ON CONTINGENCY TABLES

Compiled by
Marvin A. Kastenbaum

1968

Berkson, J., "Application of Minimum Logit χ^2 Estimate to a Problem of Grizzle with a Notation on the Problem of No Interaction," *Biometrics* **24**(1), 75–96.

Bhapkar, V. P., and Gary G. Koch, "Hypotheses of 'No Interaction' in Multidimensional Contingency Tables," *Technometrics* **10**(1), 107–23.

Bishop, Yvonne, and S. Fienberg, "Incomplete Two-Dimensional Contingency Tables," unpublished manuscript.

Bishop, Yvonne, "Full Contingency Tables, Logits and Split Contingency Tables," unpublished manuscript.

Fienberg, S. E., and Paul W. Holland, "Empty Cells in Contingency Tables," unpublished manuscript.

Goodman, Leo A., "Independence, Quasi-Independence, and Interactions in Contingency Tables With or Without Missing Values," unpublished manuscript.

Hamdan, M. A., "On the Structure of the Tetrachoric Series," *Biometrika* **55**(1), 261–62.

Hamdan, M. A., "Optimum Choice of Classes for Contingency Tables," *J. Am. Statist. Assoc.* **63**(321), 291–97.

Ireland, C. T., and S. Kullback, "Contingency Tables with Given Marginals," *Biometrika* **55**(1), 179–88.

Mosteller, Frederick, "Association and Estimation in Contingency Tables," *J. Am. Statist. Assoc.* **63**(321), 1–28.

1967

Assakul, K., and C. H. Proctor, "Testing Independence in Two Way Contingency Tables with Data Subject to Misclassification," *Psychometrika* **32**(1), 67–76.

Chernoff, Herman, "Degrees of Freedom for Chi-Square," *Technometrics* **9**(3), 489–90.

Cox, D. R., and E. Lauh, "A Note on the Graphical Analysis of Multidimensional Contingency Tables," *Technometrics* **9**(3), 481–88.

Grizzle, James E., "Continuity Correction in the χ^2-Test for 2×2 Tables," *Am. Statistician* **21**(4), 28–32.

Ives, K. H., and J. D. Simmons, "A Correlation Measure for Nominal Data," *Am. Statistician* **21**(5), 16–17.

Mardia, K. V., "Some Contributions to Contingency-Type Bivariate Distributions," *Biometrika* **54**, 235–50.

Naylor, Alfred F., "Small Sample Considerations in Combining 2×2 Tables," *Biometrics* **23**(2), 349–56.

Rudolph, G. J., "A Quasi-Multinomial Type Contingency Table," *S. African Statist. J.* **1**, 59–65.

1966

Armitage, P., "The Chi-Square Test for Heterogeneity of Proportions After Adjustment for Stratification," *J. Roy. Statist. Soc. B* **28**(1), 150–63.

Bennett, B. M., and C. Horst, *Supplement to Tables for Testing Significance in a 2 × 2 Contingency Table*, Cambridge University Press.

Bhapkar, V. P., "A Note on the Equivalence of Two Test Criteria for Hypotheses in Categorical Data," *J. Am. Statist. Assoc.* **61**(313), 228–35.

Bhat, B. R., and S. R. Kulami, "Lamp Tests of Linear and Loglinear Hypotheses in Multinomial Experiments," *J. Am. Statist. Assoc.* **61**(313), 236–45.

Castellan, N. J., "On the Estimation of the Tetrachoric Correlation Coefficient," *Psychometrika* **31**(1), 67–73.

Caussinus, H., "Remarks on Estimation Problems and Testing in Truncated Contingency Tables," *Compt. Rend.* **262,** 293–95.

Caussinus, H., "On the Analysis of Certain Contingency Tables," *Compt. Rend.* **262,** 551–54.

Chapman, D. G., and Rosa C. Meng, "The Power of Chi-Square Tests for Contingency Tables," *J. Am. Statist. Assoc.* **61**(316), 967–74.

Cox, D. R., "A Simple Example of a Comparison Involving Quantal Data," *Biometrika* **53**, 215–20.

Gart, J. J., "Alternative Analyses of Contingency Tables," *J. Roy. Statist. Soc. B* **28**(1), 164–79.

Lancaster, H. O., "Forerunners of the Pearson χ^2," *Australian J. Statist.* **8**, 117–26.

Mantel, N., "Models for Complex Contingency Tables and Polychotomous Dosage Response Curves," *Biometrics* **22**(1), 83–95.

Pasternack, Bernard S., and Nathan Mantel, "A Deficiency in the Summation of Chi Procedure," *Biometrics* **22**(2), 407–8.

Quensel, C. E., "On Tests of Independence and Evaluation of the Magnitude of the Dependence Between Two Categorized Data," *Skand. Aktuarietidskr.*, Haft. 3–4, 199–217.

Slakter, M. J., "Comparative Validity of the Chi-Square and Two Modified Chi-Square Goodness-of-Fit Tests for Small but Equal Expected Frequencies," *Biometrika* **53**, 619–22.

Weiler, H., "A Coefficient Measuring the Goodness of Fit," *Technometrics* **8**(2), 327–34.

1965

Asano, C., "On Estimating Multinomial Probabilities by Pooling Incomplete Samples," *Ann. Inst. Statist. Math.* (Tokyo) **17**(1), 1.

Bhapkar, V. P., and G. G. Koch, "On the Hypothesis of 'No-Interaction' in Three-Way Contingency Tables," *Institute of Statistics, Univ. of North Carolina, Mimeo Series* No. 440.

Bhapkar, V. P., and G. G. Koch, "Hypotheses of No Interaction in Four-Dimensional Contingency Tables," *Institute of Statistics, Univ. of North Carolina, Mimeo Series* No. 449.

Birch, M. W., "The Detection of Partial Association II: The General Case," *J. Roy. Statist. Soc. B* **27**, 111–23.

Caussinus, H., "On Truncated Contingency Tables," *Compt. Rend.* **261,** 5303–6.

Fleiss, J. L., "Estimating the Accuracy of Dichotomous Judgments," *Psychometrika* **30**(4), 469–79.

Good, I. J., *The Estimation of Probabilities; an Essay on Modern Bayesian Methods,* MIT Research Monograph No. 30.

Haynam, G. E., and F. C. Leone, "Analysis of Categorical Data," *Biometrika* **52**, 654–60.

Katti, S. K., and A. N. Sastry, "Biological Examples of Small Expected Frequencies and the Chi-Square Test," *Biometrics* **21**(1), 49–54.

Lancaster, H. O., and T. A. I. Brown, "Sizes of the χ^2 Test in Symmetrical Multinomials," *Australian J. Statist.* **7**, 40–44.

Lewontin, R. C., and J. Felsenstein, "The Robustness of Homogeneity Tests in $2 \times n$ Tables," *Biometrics* **21**(1), 19–33.

Mote, V. L., and R. L. Anderson, "An Investigation of the Effect of Misclassification on the Properties of χ^2-Tests in the Analysis of Categorical Data," *Biometrika* **52**.

Radhakrishna, S., "Combination of Results from Several 2×2 Contingency Tables," *Biometrics* **21**(1), 86–98.

Sachs, L., "The Comparison of Two Percentages – Independence Tests for Multifold Tables," *Biometrische Z.* **7**, 55–60.

Slakter, M. J., "A Comparison of the Pearson Chi-Square and Kolmogorov Goodness-of-Fit Tests with Respect to Validity," *J. Am. Statist. Assoc.* **60**, 854–58.

Watson, G. S., "Some Bayesian Methods Related to χ^2," *Bull. Inst. Intern. Statist.* **41**, book 1, 64–76.

1964

Allison, Harry E., "Computational Forms for Chi-Square," *Am. Statistician* **18**(1), 17–18.

Bennett, B. M., and E. Nakamura, "Tables for Testing Significance in a 2×3 Contingency Table," *Technometrics* **6**(4), 439–58.

Birch, M. W., "The Detection of Partial Association, I: The 2×2 Case," *J. Roy. Statist. Soc. B* **26**, 313–24.

Bross, I. D. J., "Taking a Covariable into Account," *J. Am. Statist. Assoc.* **59**(307), 725–36.

Caussinus, H., "On a Simple Association Coefficient for 2×2 Tables; Comparison Between Several Tables by Means of It," *Compt. Rend.* **259**, 3171–74.

Chew, Victor, "Application of the Negative Binomial Distribution with Probability of Misclassification," *Virginia J. Sci.* **15**(1), 34–40.

Goodman, Leo A., "Simultaneous Confidence Limits for Cross-Product Ratios in Contingency Tables," *J. Roy. Statist. Soc. B* **26**(1), 86–102.

Goodman, Leo A., "Simple Methods for Analyzing Three-Factor Interaction in Contingency Tables ," *J. Am. Statist. Assoc.* **59**, 319–52.

Goodman, Leo A., "Interactions in Multidimensional Contingency Tables," *Ann. Math. Statist.* **35**(2), 632–46.

Goodman, Leo A., "Simultaneous Confidence Intervals for Contrasts Among Multinomial Populations," *Ann. Math. Statist.* **35**(2), 716–25.

Harkness, W. L., and L. Katz, "Comparison of the Power Functions for the Test of Independence in 2×2 Contingency Tables," *Ann. Math. Statist.* **35**(3), 1115–27.

Heuze, G., "Equivalence Relations in Contingency Tables," *Compt. Rend.* **258**, 5349–51.

Lancaster, H. O., and M. A. Hamdan, "Estimation of the Correlation Coefficient in Contingency Tables with Possibly Non-Metrical Characters," *Psychometrika* **29**, 383–91.

Lindley, D. V., "The Bayesian Analysis of Contingency Tables," *Ann. Math. Statist.* **35**(4), 1622–43.

Plackett, R. L., "The Continuity Correction in 2×2 Tables," *Biometrika* **21**, parts 3 and 4, 327–38.

Putter, Joseph, "The χ^2 Goodness-of-Fit Test for a Class of Cases of Dependent Observations," *Biometrika* **51**, 250–52.

Romier, G., "Automorphisms and Algebra of Contingency Tables. Algebraic Characterization of Partially Balanced Block Designs," *Compt. Rend.* **258**, 5345–48.

Somers, R. H., "Simple Measures of Association for the Triple Dichotomy," *J. Roy. Statist. Soc. A* **127**(3), 409–15.

Tallis, G. M., "The Use of Models in the Analysis of Some Classes of Contingency Tables," *Biometrics* **24**(4), 832–39.

Westenberg, J., "Nomograms for Testing Significance in a 2×2 Contingency Table for Sample Sizes of 10 to 500," *Proc. Koninkl. Ned. Akad. Wetenschap.*, Ser. A **67**(IV) 441–66.

1963

Bennett, B. M., and E. Nakamura, "Tables for Testing Significance in a 2×3 Contingency Table," *Technometrics* **5**(4), 501–11.

Birch, M. W., "Maximum Likelihood in Three-Way Contingency Tables," *J. Roy. Statist. Soc. B* **25**(1), 220–33.

Darroch, J. N., and S. D. Silvey, "On Testing More Than One Hypothesis," *Ann. Math. Statist.* **34**(2), 555–67.

Diamond, Earl L., "The Limiting Power of Categorical Data Chi-Square Tests Analogous to Normal Analysis of Variance," *Ann. Math. Statist.* **34**(4), 1432–41.

Edwards, A. W. F., "The Measure of Association in a 2×2 Table," *J. Roy. Statist. Soc. A* **126**(1), 109–14.

Feldman, S. E., and E. Klinger, "Short Cut Calculation of the Fisher-Yates Exact Test," *Psychometrika* **28**(3), 289–91.

Finney, D. J., R. Latscha, B. M. Bennett, and P. Hsu, *Tables for Testing Significance in a 2×2 Contingency Table*, Cambridge University Press.

Gold, Ruth A., "Tests Auxiliary to χ^2 Tests in a Markov Chain," *Ann. Math. Statist.* **34**(1), 56–74.

Good, I. J., "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables," *Ann. Math. Statist.* **34**(3), 911–34.

Goodman, Leo A., "On Methods for Comparing Contingency Tables," *J. Roy. Statist. Soc. A* **126**(1), 94–108.

Goodman, Leo A., "On Plackett's Test for Contingency Table Interactions," *J. Roy. Statist. Soc. B* **25**(1), 179–88.

Goodman, Leo A., and W. H. Kruskal, "Measures of Association for Cross Classification III: Approximate Sampling Theory," *J. Am. Statist. Assoc.* **58**, 310–64.

Ku, H. H., "A Note on Contingency Tables Involving Zero Frequencies and the 21 Tests," *Technometrics* **5**(3), 398–400.

Mantel, Nathan, "Chi-Square Tests with One Degree of Freedom: Extensions of the Mantel-Haenszel Procedure," *J. Am. Statist. Assoc.* **58**, 690–700.

Newell, D. J., "Misclassification in 2×2 Tables," *Biometrics* **19**(1), 187–88.

Okamato, Masashi, "Chi-Square Statistic Based on the Pooled Frequencies of Several Observations," *Biometrika* **50**, 524–28.

Ries, P. N., and Harry Smith, "The Use of Chi-Square for Preference Testing in Multidimensional Problems," *Chem. Eng. Progr., Symp. Ser.* **59**(42), 39–43.

Solari, M. E., "The Distribution of the Chi-Square Test of Fit Statistic," *Statistician* **13**, 263–67.

Walsh, John E., "Loss in Test Efficiency Due to Misclassification for 2×2 Tables," *Biometrics* **19**(1), 158–62.

1962

Bennett, B. M., "Note on an Exact Test for the 2×2 Contingency Table Using the Negative Binomial Model," *Metrika* **5**, 154–57.

Daly, C., "A Simple Test for Trends in a Contingency Table," *Biometrics* **18**(1), 114–19.

Darroch, J. N., "Interactions in Multi-Factor Contingency Tables," *J. Roy. Statist. Soc. B* **24**(1), 251–63.

Fisher, Sir Ronald A., "Confidence Limits for a Cross-Product Ratio," *Australian J. Statist.* **4**(1), 41.

Gart, J. J., "Approximate Confidence Limits for Relative Risks," *J. Roy. Statist. Soc. B* **24**(2), 454–63.

Gart, J. J., "On the Combination of Relative Risks," *Biometrics* **18**(4), 601–10.

Kincaid, W. M., "The Combination of $2 \times m$ Contingency Tables," *Biometrics* **18**(2), 224–28.

Kullback, S., M. Kupperman, and H. H. Ku, "An Application of Information Theory to the Analysis of Contingency Tables with a Table of $2N \ln N$, $N = 1(1)10{,}000$," *J. Res. Natl. Bur. Std., B* **66**, 217–43.

Kullback, S., M. Kupperman, and H. H. Ku, "Tests for Contingency Tables and Markov Chains," *Technometrics* **4**(4), 573–608.

Lewis, B. N., "On the Analysis of Interaction in Multi-Dimensional Contingency Tables," *J. Roy. Statist. Soc. A* **125**(1), 88–117.

Plackett, R. L., "A Note on Interactions in Contingency Tables," *J. Roy. Statist. Soc. B* **24**(1), 162–66.

Tallis, G. M., "The Maximum Likelihood Estimation of Correlation from Contingency Tables," *Biometrics* **18**(3), 342–53.

Young, D. H., "Two Alternatives to the Standard χ^2 Test of the Hypothesis of Equal Cell Frequencies," *Biometrika* **49**, 107–16.

1961

Berger, Agnes, "On Comparing Intensities of Association Between Two Binary Characteristics in Two Different Populations," *J. Am. Statist. Assoc.* **56**, 889–908.

Bhapkar, V. P., "Some Tests for Categorical Data," *Ann. Math. Statist.* **32**(1), 72–83.

Billingsley, P., *Statistical Inference for Markov Processes*, Statistical Research Monographs, vol. 2, The University of Chicago Press.

Claringbold, P. J., "The Use of Orthogonal Polynomials in the Partition of Chi-Square," *Australian J. Statist.* **3**(2), 48–63.

Friedlander, D., "A Technique for Estimating a Contingency Table, Given the Marginal Totals and Some Supplementary Data," *J. Roy. Statist. Soc. A* **124**(3), 412–20.

Gregory, G., "Contingency Tables with a Dependent Classification," *Australian J. Statist.* **3**(2), 42–47.

Grizzle, James E., "A New Method of Testing Hypotheses and Estimating Parameters for the Logistic Model," *Biometrics* **17**(3), 372–85.

Kendall, M. G., and A. Stuart, *The Advanced Theory of Statistics*, vol. 2, London, England: Charles Griffin and Company.

Okamato, M., and G. Ishii, "Test of Independence in Intraclass 2×2 Tables," *Biometrika* **48**, 181–90.

Rogot, E., "A Note on Measurement Errors and Detecting Real Differences," *J. Am. Statist. Assoc.* **56**, 314–19.

Schull, William J., "Some Problems of Analysis of Multi-Factor Tables," *Bull. Inst. Intern. Statist.* **28**, part 3, 259–70.

Yates, F., "Marginal Percentages in Multiway Tables of Quantal Data with Disproportionate Frequencies," *Biometrics* **17**(1), 1–9.

1960

Bennett, B. M., and P. Hsu, "On the Power Function of the Exact Test for the 2×2 Contingency Table," *Biometrika* **47**, 393–98.

Gridgeman, N. T., "Card-Matching Experiments: a Conspectus of Theory," *J. Roy. Statist. Soc. A* **123**(1), 45–49.

Ishii, G., "Intraclass Contingency Tables," *Ann. Inst. Statist. Math.* (Tokyo) **12**, 161–207; corrections, p. 279.

Kastenbaum, M. A., "A Note on the Additive Partitioning of Chi-Square in Contingency Tables," *Biometrics* **16**(3), 416–22.

Kupperman, Morton, "On Comparing Two Observed Frequency Counts," *Appl. Statist.* **9**(1), 37–42.

Lancaster, H. O., "On Tests on Independence in Several Dimensions," *J. Australian Math. Soc.* **1**, 241–54.

Robertson, W. H., "Programming Fisher's Exact Method of Comparing Two Percentages," *Technometrics* **2**(1), 103–7.

1959

Anderson, R. L., "Use of Contingency Tables in the Analysis of Consumer Preference Studies," *Biometrics* **15**(4), 582–90.

Chakravarti, I. M., and C. R. Rao, "Tables for Some Small Sample Tests of Significance for Poisson Distributions and 2×3 Contingency Tables," *Sankhya* **21**, parts 3 and 4, 315–26.

Goodman, Leo A., and W. H. Kruskal, "Measures of Association for Cross Classification II: Further Discussion and References," *J. Am. Statist. Assoc.* **54**, 123–63.

Haldane, J. B. S., "The Analysis of Heterogeneity, I," *Sankhya* **21**, parts 3 and 4, 209–16.

Hoyt, C. J., P. R. Krishnaiah, and E. P. Torrance, "Analysis of Complex Contingency Data," *J. Exptl. Educ.* **27**, 187–94.

Kastenbaum, M. A., and D. E. Lamphiear, "Calculation of Chi-Square to Test the No Three-Factor Interaction Hypothesis," *Biometrics* **15**(1), 107–15.

Kullback, S., *Information Theory and Statistics*, Wiley, New York.

Kupperman, Morton, "A Rapid Significance Test for Contingency Tables," *Biometrics* **15**(4), 625–28.

Nass, C. A. G., "The χ^2 Test for Small Expectations in Contingency Tables, with Special Reference to Accidents and Absenteeism," *Biometrika* **46**, 365–85.

Silvey, S. D., "The Lagrangian Multiplier Tests," *Ann. Math. Statist.* **30**(2), 389–407.

Somers, Robert H., "The Rank Analogue of Product-Moment Partial Correlation and Regression, with Application to Manifold, Ordered Contingency Tables," *Biometrika* **46**, 241–46.

Steyn, H. S., "On χ^2-Tests for Contingency Tables of Negative Binomial Type," *Statist. Neerlandica* **13**, 433–44.

Weiner, Irving B., "A Note of the Use of Mood's Likelihood Ratio Test for Item Analyses Involving 2×2 Tables with Small Samples," *Psychometrika* **24**(4), 371–72.

1958

Blalock, H. M., Jr., "Probabilistic Interpretations for the Mean Square Contingency," *J. Am. Statist. Assoc.* **53**, 102–5.

Kastenbaum, M. A., "Estimation of Relative Frequencies of Four Sperm Types in Drosophila melanogaster," *Biometrics* **14**(2), 223–28.

Mitra, S. K., "On the Limiting Power Function of the Frequency Chi-Square Test," *Ann. Math. Statist.* **29**, 1221–33.

Snedecor, G. W., "Chi-Square of Bartlett, Mood and Lancaster in a 2^3 Contingency Table," *Biometrics* **14**(4), 560–62. (Query)

1957

Bross, Irwin D. J., and Ethel L. Kasten, "Rapid Analysis of 2×2 Tables," *J. Am. Statist. Assoc.* **52**, 18–28.

Corsten, L. C. A., "Partition of Experimental Vectors Connected with Multinomial Distributions," *Biometrics* **13**(4), 451–84.

Edwards, J. H., "A Note on the Practical Interpretation of 2×2 Tables," *Brit. J. Prevent. Social Med.* **11**, 73–78.

Lancaster, H. O., "Some Properties of the Bivariate Normal Distribution Considered in the Form of a Contingency Table," *Biometrika* **44**, 289–92.

Mote, V. L., "An Investigation of the Effect of Misclassification of the Chi-Square Tests in the Analysis of Categorial Data," unpublished Ph.D. dissertation, North Carolina State College, Raleigh, N.C. (also Institute of Statistics Mimeo Series No. 182).

Roy, S. N., *Some Aspects of Multivariate Analysis*, Wiley, New York.

Sakoda, J. M., and B. H. Cohen, "Exact Probabilities for Contingency Tables Using Binomial Coefficients," *Psychometrika* **22**(1), 83–86.

Woolf, Barnet, "The Log Likelihood Ratio Test (The *G*-Test). Methods and Tables for Tests of Heterogeneity in Contingency Tables," *Ann. Human Genet.* **21**, 397–409.

1956

Fishman, J. A., "A Note on Jenkins' 'Improved Method for Tetrachoric *r*,'" *Psychometrika* **20**(3), 305.

Good, I. J., "On the Estimation of Small Frequencies in Contingency Tables," *J. Roy. Statist. Soc. B* **13**(1), 113–24.

Gridgeman, N. T., "A Tasting Experiment," *Appl. Statist.* **15**(2), 106–12.

Leander, E. K., and D. J. Finney, "An Extension of the Use of the χ^2 Test," *Appl. Statist.* **5**(2), 132–36.

Mainland, D., L. Herrera, and M. I. Sutcliffe, "Statistical Tables for Use with Binomial Samples – Contingency Tests, Confidence Limits, and Sample Size Estimates," New York University College of Medicine, New York.

Roy, S. N., and Marvin A. Kastenbaum, "On the Hypothesis of No 'Interaction' in a Multiway Contingency Table," *Ann. Math. Statist.* **27**(3), 749–57.

Roy, S. N., and S. K. Mitra, "An Introduction to Some Non-Parametric Generalizations of Analysis of Variance and Multivariate Analysis," *Biometrika* **43**, parts 3 and 4, 361–76.

Watson, G. S., "Missing and 'Mixed-Up' Frequencies in Contingency Tables," *Biometrics* **12**(1), 47–50.

1955

Armitage, P., "Tests for Linear Trends in Proportions and Frequencies," *Biometrics* **11**(3), 375–86.

Armsen, P., "Tables for Significance Tests of 2×2 Contingency Tables," *Biometrika* **42**, 494–505.

Cochran, W. G., "A Test of Linear Function of the Deviations Between Observed and Expected Numbers," *J. Am. Statist. Assoc.* **50**, 377–97.

Haldane, J. B. S., "Substitutes for χ^2," *Biometrika* **42**, 265–66.

Haldane, J. B. S., "A Problem in the Significance of Small Numbers," *Biometrika* **42**, 266–67.

Jenkins, W. L., "An Improved Method for Tetrachoric *r*," *Psychometrika* **20**(3), 253–58.

Kastenbaum, Marvin A., "Analysis of Data in Multiway Contingency Tables," unpublished doctoral dissertation, North Carolina State College, October 1955.

Leslie, P. H., "A Simple Method of Calculating the Exact Probability in 2×2 Contingency Tables with Small Marginal Totals," *Biometrika* **42**, 522–23.

Mitra, S. K., "Contributions to the Statistical Analysis of Categorical Data," *North Carolina Institute of Statistics Mimeograph Series*, No. 142, December 1955.

Roy, S. N., and Marvin A. Kastenbaum, "A Generalization of Analysis of Variance and Multivariate Analysis to Data Based on Frequencies in Qualitative Categorical or Class Intervals," *North Carolina Institute of Statistics Mimeograph Series*, No. 131, June 1955.

Roy, S. N., and S. K. Mitra, "An Introduction to Some Non-Parametric Generalizations of Analysis of Variance and Multivariate Analysis," *North Carolina Institute of Statistics Mimeograph Series*, No. 139, November 1955.

Sekar, C. Chandra, S. P. Agarivala, and P. N. Chakraborty, "On the Power Function of a Test of Significance for the Difference Between Two Proportions," *Sankhya* **15**, part 4, 381–90.

Stuart, Alan, "A Test of Homogeneity of the Marginal Distributions in a Two-Way Classification," *Biometrika* **42**, 412–16.

Woolf, Barnet, "On Estimating the Relation Between Blood Group and Disease," *Ann. Human Genet.* **19**, 251–53.

Yates, F., "A Note on the Application of the Combination of Probabilities Test to a Set of 2×2 Tables," *Biometrika* **42**, 401–11.

Yates, F., "The Use of Transformations and Maximum Likelihood in the Analysis of Quantal Experiments Involving Two Treatments," *Biometrika* **42**, 382–403.

1954

Bross, Irwin D. J., "Misclassification in 2×2 Tables," *Biometrics* **10**(4), 478–86.

Cochran, W. G., "Some Methods for Strengthening the Common Chi-Square Tests," *Biometrics* **10**(4), 417–51.

Dawson, R. B., "A Simplified Expression for the Variance of the χ^2 Function on a Contingency Table," *Biometrika* **41**, 280.

Goodman, Leo A., and W. H. Kruskal, "Measures of Association for Cross Classification," *J. Am. Statist. Assoc.* **49**, 732–64.

Kimball, A. W., "Short-Cut Formulas for the Exact Partition of Chi-Square in Contingency Tables," *Biometrics* **10**(4), 452–58.

McGill, W. J., "Multivariate Information Transmission," *Psychometrika* **19**(2), 97–116.

1952

Cochran, William G., "The χ^2 Test of Goodness of Fit," *Ann. Math. Statist.* **23**(3), 315–45.

Dyke, G. V., and H. D. Patterson, "Analysis of Factorial Arrangements when the Data Are Proportions," *Biometrics* **8**, 1–12.

1951

Lancaster, H. O., "Complex Contingency Tables Treated by the Partition of Chi-Square," *J. Roy. Statist. Soc. B* **13**, 242–49.

Simpson, C. H., "The Interpretation of Interaction in Contingency Tables," *J. Roy. Statist. Soc. B* **13**, 238–41.

1950

Tocher, K. D., "Extension of the Neyman-Pearson Theory of Tests to Discontinuous Variates," *Biometrika* **37**, 130–44.

1949

Hsu, P. L., "The Limiting Distributions of Functions of Sample Means and Application to Testing Hypotheses," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* (1945, 1946), University of California Press, Berkeley and Los Angeles.

Irwin, J. O., "A Note on the Subdivision of Chi-Square into Components," *Biometrika* **36**, 130–34.

Lancaster, H. O., "The Derivation and Partition of Chi-Square in Certain Discrete Distributions," *Biometrika* **36**, 117–29.

1948

Yates, F., "The Analysis of Contingency Tables with Groupings Based on Quantitative Characters," *Biometrika* **35**, 176–81.

1947

Barnard, G. A., "Significance Tests for 2×2 Tables," *Biometrika* **34**, 123–38.

Pearson, E. S., "The Choice of Statistical Tests Illustrated on the Interpretation of Data Classed in a 2×2 Table," *Biometrika* **34**, 139–67.

1946

Cramer, H., *Mathematical Methods of Statistics*, Princeton University Press, p. 424.

1945

Norton, H. W., "Calculation of Chi-Square for Complex Contingency Tables," *J. Am. Statist. Assoc.* **40**, 251–58.

1937

Haldane, J. B. S., "The Exact Value of the Moments of the Distribution of χ^2 Used as a Test of Goodness of Fit, when Expectations Are Small," *Biometrika* **29**, 133–43.

1935

Bartlett, M. S., "Contingency Table Interactions," *J. Roy. Statist. Soc. Suppl.* **2**, 248–52.

Wilks, S. S., "The Likelihood Test of Independence in Contingency Tables," *Ann. Math. Statist.* **6**, 190–96.

1934

Fisher, R. A., *Statistical Methods for Research Workers*, Edinburgh, 5th and subsequent editions, Oliver and Boyd Ltd., sect. 21.02.

Yates, F., "Contingency Tables Involving Small Numbers and the χ^2 Test," *J. Roy. Statist. Soc. Suppl. 1* **1**, 217–35.

1924

Fisher, R. A., "The Conditions Under Which Chi-Square Measures the Discrepancy Between Observation and Hypothesis," *J. Roy. Statist. Soc.* **87**, 442–50.

1922

Fisher, R. A., "On the Interpretation of Chi-Square from Contingency Tables, and the Calculation of *P*," *J. Roy. Statist. Soc.* **85**, 87–94.

1900

Pearson, Karl, "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such that It Can Be Reasonably Supposed to Have Arisen from Random Sampling," *Phil. Mag.* [5] **50**, 157–72.

INDEX

Agarivala, S. P., C. Chandra Sekar, and P. N. Chakraborty, 1955
Allison, Harry E., 1964
Anderson, R. L., 1959
Anderson, R. L., and V. L. Mote, 1965
Armitage, P., 1955, 1966
Armsen, P., 1955
Asano, C., 1965
Assakul, K., and C. H. Proctor, 1967

Barnard, G. A., 1947
Bartlett, M. S., 1935
Bennett, B. M., 1962
Bennett, B. M., D. J. Finney, R. Latscha, and P. Hsu, 1963
Bennett, B. M., and C. Horst, 1966
Bennett, B. M., and P. Hsu, 1960
Bennett, B. M., and E. Nakamura, 1963, 1964
Berger, Agnes, 1961
Berkson, J., 1968
Bhapkar, V. P., 1961, 1966
Bhapkar, V. P., and G. G. Koch, 1965, 1968
Bhat, B. R., and S. R. Kularni, 1966
Billingsley, P., 1961
Birch, M. W., 1963, 1964, 1965
Bishop, Yvonne, 1968
Bishop, Yvonne, and S. Fienberg, 1968
Blalock, H. M., Jr., 1958
Bross, Irwin D. J., 1954, 1964
Bross, Irwin D. J., and Ethel L. Kasten, 1957
Brown, T. A. I., and H. O. Lancaster, 1965

Castellan, N. J., 1966
Caussinus, H., 1964, 1965, 1966
Chakravarti, I. M., and C. R. Rao, 1959
Chakraborty, P. N., C. Chandra Sekar, and S. P. Agarivala, 1955
Chapman, D. G., and Rosa C. Meng, 1966
Chernoff, Herman, 1967
Chew, Victor, 1964
Claringbold, P. J., 1961
Cochran, W. G., 1952, 1954, 1955
Cohen, B. H., and J. M. Sakoda, 1957
Corsten, L. C. A., 1957
Cox, D. R., 1966
Cox, D. R., and E. Lauh, 1967
Cramer, H., 1946

Daly, C., 1962
Darroch, J. N., 1962
Darroch, J. N., and S. D. Silvey, 1963
Dawson, R. B., 1954
Diamond, Earl L., 1963
Dyke, G. V., and H. D. Patterson, 1952

Edwards, A. W. F., 1963
Edwards, J. H., 1957

Feldman, S. E., and E. Klinger, 1963
Felsentein, J., and R. C. Lewontin, 1965
Fienberg, S., and Yvonne Bishop, 1968
Fienberg, S. E., and Paul W. Holland, 1968
Finney, D. J., R. Latscha, B. M. Bennett, and P. Hsu, 1963
Finney, D. J., and E. K. Leander, 1956
Fisher, Sir Ronald A., 1922, 1924, 1934, 1962
Fishman, J. A., 1956
Fleiss, J. L., 1965
Friedlander, D., 1961

Gart, J. J., 1962, 1966
Gold, Ruth A., 1963
Good, I. J., 1956, 1963, 1965
Goodman, Leo A., 1963, 1964, 1968
Goodman, Leo A., and W. H. Kruskal, 1954, 1959, 1963
Gregory, G., 1961
Gridgeman, N. T., 1956, 1960
Grizzle, James E., 1961, 1967

Haldane, J. B. S., 1937, 1955, 1959
Hamdan, M. A., 1968
Hamdan, M. A., and H. O. Lancaster, 1964
Harkness, W. L., and L. Katz, 1964
Haynam, G. E., and F. C. Leone, 1965
Herrera, L., D. Mainland, and M. I. Sutcliffe, 1956
Heuze, G., 1964
Holland, Paul W., and S. E. Fienberg, 1968
Horst, C., and B. M. Bennett, 1966
Hoyt, C. J., P. R. Krishnaiah, and E. P. Torrance, 1959
Hsu, P. L., 1949
Hsu, P., and B. M. Bennett, 1960
Hsu, P., D. J. Finney, R. Latscha, and B. M. Bennett, 1963

Ireland, C. T., and S. Kullback, 1968
Irwin, J. O., 1949
Ishii, G., 1960
Ishii, G., and M. Okamato, 1961
Ives, K. H., and J. D. Simmons, 1967

Jenkins, W. L., 1955

Kasten, Ethel L., and Irwin D. J. Bross, 1957
Kastenbaum, M. A., 1955, 1958, 1960
Kastenbaum, M. A., and D. E. Lamphiear, 1959
Kastenbaum, M. A., and S. N. Roy, 1955, 1956
Katti, S. K., and A. N. Sastry, 1965
Katz, L., and W. L. Harkness, 1964
Kendall, M. G., and A. Stuart, 1961
Kimball, A. W., 1954
Kincaid, W. M., 1962
Klinger, E., and S. E. Feldman, 1963
Koch, G. G., and V. P. Bhapkar, 1965, 1968
Krishnaiah, P. R., C. J. Hoyt, and E. P. Torrance, 1959
Kruskal, W. H., and Leo A. Goodman, 1954, 1959, 1963

Ku, H. H., 1963
Ku, H. H., S. Kullback, and M. Kupperman, 1962
Kularni, S. R., and B. R. Bhat, 1966
Kullback, S., 1959
Kullback, S., and C. T. Ireland, 1968
Kullback, S., M. Kupperman, and H. H. Ku, 1962
Kupperman, Morton, 1959, 1960
Kupperman, M., S. Kullback, and H. H. Ku, 1962

Lamphiear, D. E., and M. A. Kastenbaum, 1959
Lancaster, H. O., 1949, 1951, 1957, 1960, 1966
Lancaster, H. O., and T. A. I. Brown, 1965
Lancaster, H. O., and M. A. Hamdan, 1964
Latscha, R., D. J. Finney, B. M. Bennett, and P. Hsu, 1963
Lauh, E., and E. R. Cox, 1967
Leander, E. K., and D. J. Finney, 1956
Leone, F. C., and G. E. Haynam, 1965
Leslie, P. H., 1955
Lewis, B. N., 1962
Lewontin, R. C., and J. Felsentein, 1965
Lindley, D. V., 1964

Mainland, D., L. Herrera, and M. I. Sutcliffe, 1956
Mantel, Nathan, 1963, 1966
Mantel, Nathan, and Bernard S. Pasternack, 1966
Mardia, K. V., 1967
McGill, W. J., 1954
Meng, Rosa C., and D. G. Chapman, 1966
Mitra, S. K., 1955, 1958
Mitra, S. K., and S. N. Roy, 1955, 1956
Mosteller, Frederick, 1968
Mote, V. L., 1957
Mote, V. L., and R. L. Anderson, 1965

Nakamura, E., and B. M. Bennett, 1963, 1964
Nass, C. A. G., 1959
Naylor, Alfred F., 1967
Newell, D. J., 1963
Norton, H. W., 1945

Okamato, Masashi, 1963
Okamato, M., and G. Ishii, 1961

Pasternack, Bernard S., and Nathan Mantel, 1966
Patterson, H. D., and G. V. Dyke, 1952
Pearson, E. S., 1947
Pearson, Karl, 1900
Plackett, R. L., 1962, 1964
Proctor, C. H., and K. Assakul, 1967
Putter, Joseph, 1964

Quensel, C. E., 1966

Radhakrishna, S., 1965
Rao, C. R., and I. M. Chakravarti, 1959
Ries, P. N., and Harry Smith, 1963

Robertson, W. H., 1960
Rogot, E., 1961
Romier, G., 1964
Roy, S. N., 1957
Roy, S. N., and Marvin A. Kastenbaum, 1955, 1956
Roy, S. N., and S. K. Mitra, 1955, 1956
Rudolph, G. J., 1967

Sachs, L., 1965
Sakoda, J. M., and B. H. Cohen, 1957
Sastry, A. N., and S. K. Katti, 1965
Schull, William J., 1961
Sekar, C. Chandra, S. P. Agarivala, and P. N. Chakraborty, 1955
Silvey, S. D., 1959
Silvey, S. D., and J. N. Darroch, 1963
Simmons, J. D., and K. H. Ives, 1967
Simpson, C. H., 1951
Slakter, M. J., 1965, 1966
Smith, Harry, and P. N. Ries, 1963
Snedecor, G. W., 1958
Solari, M. E., 1963
Somers, Robert H., 1959, 1964
Steyn, H. S., 1959
Stuart, Alan, 1955
Stuart, A., and M. G. Kendall, 1961
Sutcliffe, M. I., D. Mainland, and L. Herrera, 1956

Tallis, G. M., 1962, 1964
Tocher, K. D., 1950
Torrance, E. P., C. J. Hoyt, and P. R. Krishnaiah, 1959

Walsh, John E., 1963
Watson, G. S., 1956, 1965
Weiler, H., 1966
Weiner, Irving B., 1959
Westenberg, J., 1964
Wilks, S. S., 1935
Woolf, Barnet, 1955, 1957

Yates, F., 1934, 1948, 1955, 1961
Young, D. H., 1962

LECTURE IV: CONFIDENCE REGIONS

Marvin A. Kastenbaum

1. AN EXAMPLE OF STATISTICAL INFERENCE

Let $X_1, X_2, \ldots, X_n$ be a sequence of n independent random variables, X_i having the normal distribution $N(x_i; \mu, \sigma^2)$. If

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) ,$$

then the random variable

$$Z = \sqrt{n}(\overline{X} - \mu)/\sigma$$

has a standardized normal probability density function, $N(z; 0, 1)$. Also if $Z \in \mathcal{A}$, where

$$\mathcal{A} = \{z; -\infty < z < \infty\} ,$$

and

$$A_1 = \{z;\ z < z_1\}$$

and

$$A_2 = \{z;\ z > z_2\}$$

for $z_1 < z_2$, then

$$\text{Prob}\ \{Z \in (A_1 \cup A_2)^c\} = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-t^2/2}\ dt . \tag{1.1}$$

With simple algebraic manipulation it follows that

$$\begin{aligned} \text{Prob}\ \{Z \in (A_1 \cup A_2)^c\} &= \text{Prob}\ \{z_1 \leqq \sqrt{n}(\overline{X} - \mu)/\sigma \leqq z_2\} \\ &\equiv \text{Prob}\ \{z_1\sigma/\sqrt{n} \leqq (\overline{X} - \mu) \leqq z_2\sigma/\sqrt{n}\} \\ &\equiv \text{Prob}\ \{\overline{X} - z_2\sigma/\sqrt{n} \leqq \mu \leqq \overline{X} - z_1\sigma/\sqrt{n}\} . \end{aligned} \tag{1.2}$$

For $z_2 = -z_1 = 1.96$, the tables of the cumulative normal distribution (Tables 1*a* and *b*) show that

$$\begin{aligned} &\text{Prob}\ \{-1.96 \leqq \sqrt{n}\,(\overline{X} - \mu)/\sigma \leqq 1.96\} \\ &\quad \equiv \text{Prob}\ \{\overline{X} - 1.96\ \sigma/\sqrt{n} \leqq \mu \leqq \overline{X} + 1.96\ \sigma/\sqrt{n}\} \\ &\quad = 1 - \int_{-\infty}^{-1.96} f(z)\ dz - \int_{1.96}^{\infty} f(z)\ dz \\ &\quad = 1 - 2\int_{1.96}^{\infty} f(z)\ dz = 1 - 2\int_{-\infty}^{-1.96} f(z)\ dz = 0.95 . \end{aligned} \tag{1.3}$$

If σ is known, each of the random variables $\bar{X} - 1.96\,\sigma/\sqrt{n}$ and $\bar{X} + 1.96\,\sigma/\sqrt{n}$ is a statistic.[1] The interval $[\bar{X} - 1.96\,\sigma/\sqrt{n},\ \bar{X} + 1.96\,\sigma/\sqrt{n}]$ is called a *random interval.*

The probability statement specified by (1.3) may be read as follows: Prior to the performance of an experiment, the probability is 0.95 that the random interval $[\bar{X} - 1.96\,\sigma/\sqrt{n},\ \bar{X} + 1.96\,\sigma/\sqrt{n}]$ includes the unknown fixed point (parameter) μ.

Up to this point the discussion has been a probabilistic one in the sense that the determination of a probability density function for $\bar{X}$ and a random interval are merely exercises in probability. The application of these concepts to the realizations of an experiment can result in statistical inferences.

Suppose an experiment yields $X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_n = x_n$, with a sample value of $\bar{X}$ given by $\bar{x} = (x_1 + x_2 + \ldots + x_n)/n$. If both $\bar{x}$ and σ are known, the interval $[\bar{x} - 1.96\,\sigma/\sqrt{n},\ \bar{x} + 1.96\,\sigma/\sqrt{n}]$ will have known end points. It should be obvious that 0.95 is not the probability that this specific interval includes the parameter μ. In fact, either μ is in this interval or it is not in this interval.

However, the fact that, prior to the performance of the experiment, there was a probability of 0.95 that the random interval $[\bar{X} - 1.96\,\sigma/\sqrt{n},\ \bar{X} + 1.96\,\sigma/\sqrt{n}]$ would include the parameter μ, leads one to have the same degree of faith in the specific interval $[\bar{x} - 1.96\,\sigma/\sqrt{n},\ \bar{x} + 1.96\,\sigma/\sqrt{n}]$. This faith or confidence in the observed interval $[\bar{x} - 1.96\,\sigma/\sqrt{n},\ \bar{x} + 1.96\,\sigma/\sqrt{n}]$ is the reason it is referred to as a *95% confidence interval.* The number 0.95 is called the *confidence coefficient.* The confidence coefficient is equal to the probability that the random interval includes the parameter. One may, of course, obtain 90 or 99% confidence intervals for μ.

If *σ is not known*, the end points of the random interval would not be statistics. Although the probability statement about the random interval remains valid, a sample would not yield an interval with known end points. That is, when σ is unknown a statistical inference about μ does not appear to be feasible. Fortunately, a method does exist for bypassing this obstacle.

2. CONFIDENCE INTERVALS FOR A MEAN

Let $X_1, X_2, \ldots, X_n$ be a sequence of n independent random variables, X_i having the normal distribution $N(x_i;\ \mu,\ \sigma^2)$ with both μ $(-\infty < \mu < \infty)$ and σ^2 $(0 < \sigma^2 < \infty)$ unknown. Let $\bar{X}$ and $S^2 = \sum(X_i - \bar{X})^2/(n-1)$ denote respectively the mean and variance of $X_1, \ldots, X_n$. The problem is to find a confidence interval for μ.

It can be shown (1) that $\sqrt{n}(\bar{X} - \mu)/\sigma$ has the normal distribution $N(0, 1)$, (2) that $(n-1)S^2/\sigma^2$ has a chi-square distribution with $n - 1$ degrees of freedom, and (3) that $\sqrt{n}(\bar{X} - \mu)/\sigma$ and $(n-1)S^2/\sigma^2$ are independent. It can also be shown that a random variable T may be defined in terms of two such variables. In fact,

[1] A statistic is a function of random variables which does not depend upon an unknown parameter.

$$T = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{\dfrac{(n-1)S^2/\sigma^2}{(n-1)}}} = \sqrt{n}(\bar{X} - \mu)/S \tag{2.1}$$

has a t distribution with $n - 1$ degrees of freedom, whatever the value of $\sigma^2 > 0$ (see Appendix, Sect. 9 of this lecture).

For a given positive integer $(n - 1)$ and a probability of 0.95, say, two values a and b $(a < b)$ can be found from a table of the cumulative t distribution (Table 2) such that

$$\text{Prob}\,\{a < T < b\} = \text{Prob}\,\{a < \sqrt{n}(\bar{X} - \mu)/S < b\} = 0.95\ . \tag{2.2}$$

Since the p.d.f. of the random variable T is symmetric about the vertical axis, a and b are conventionally chosen so that $a = -b$, $b > 0$. If, for $a = -b$, the probability of this event is written in the form

$$\text{Prob}\,\{\bar{X} - bS/\sqrt{n} < \mu < \bar{X} + bS/\sqrt{n}\} = 0.95\ , \tag{2.3}$$

then the interval $[\bar{X} - bS/\sqrt{n}\,,\ \bar{X} + bS/\sqrt{n}]$ is a random interval having probability 0.95 of including the unknown fixed point (parameter) μ. If the experimental values of $X_1, X_2, \ldots, X_n$ are $x_1, x_2, \ldots, x_n$, with

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)\ ,$$

where

$$\bar{x} = \sum_{i=1}^{n} x_i/n\ ,$$

then the interval $[(\bar{x} - bs/\sqrt{n}),\ (\bar{x} + bs/\sqrt{n})]$ is a 95% confidence interval for μ, for every $\sigma^2 > 0$ (see Table 2).

Example: If $n = 10$, $\bar{x} = 3.22$, $s = 1.233$, then the interval

$$[3.22 - (2.262)(1.233)/\sqrt{10},\ 3.22 + (2.262)(1.233)/\sqrt{10}]\ ,$$

or (2.34, 4.10), is a 95% confidence interval for μ, where $b = 2.262$.

3. CONFIDENCE INTERVALS FOR THE DIFFERENCE OF TWO MEANS

The random variable T may be used to obtain a confidence interval for the difference $\mu_1 - \mu_2$ between the means of two independent normal distributions, $N(x;\ \mu_1, \sigma^2)$ and $N(y;\ \mu_2, \sigma^2)$, which have the same (but unknown) variance, σ^2.

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two sequences of independent random variables having respectively the probability density functions $N(x;\ \mu_1, \sigma^2)$ and $N(y;\ \mu_2, \sigma^2)$. Denote the

means of the two sequences by $\bar{X}$ and $\bar{Y}$ and the variances by S_1^2 and S_2^2 respectively. These four statistics are mutually independent. Moreover, $\bar{X}$ and $\bar{Y}$ are independently distributed with means μ_1 and μ_2 and variances σ^2/n and σ^2/m respectively. Accordingly, their difference $\bar{X} - \bar{Y}$ is normally distributed with mean $\mu_1 - \mu_2$ and variance $\sigma^2/n + \sigma^2/m$.[2] Then the random variable

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2/n + \sigma^2/m}} \tag{3.1}$$

has the normal distribution $N(0, 1)$. Furthermore $(n-1)S_1^2/\sigma^2$ and $(m-1)S_2^2/\sigma^2$ have independent chi-square distributions with $n-1$ and $m-1$ degrees of freedom respectively, so that their sum

$$\frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2}$$

has a chi-square distribution with $n + m - 2$ degrees of freedom. It follows that the random variable

$$\begin{aligned} T &= \frac{[(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)]/\sqrt{\sigma^2/n + \sigma^2/m}}{\sqrt{\dfrac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2(n+m-2)}}} \\ &= \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n-1)S_1^2 + (m-1)S_2^2}{(n+m-2)}\left(\dfrac{1}{n} + \dfrac{1}{m}\right)}} \end{aligned} \tag{3.2}$$

has a t distribution with $n + m - 2$ degrees of freedom.

As in the earlier part of this discussion, a positive number b can be found such that

$$\text{Prob}\,\{-b < T < b\} = 0.95.$$

If

$$S_d = \sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{(n+m-2)}\left(\frac{1}{n} + \frac{1}{m}\right)}\,, \tag{3.3}$$

this probability may be written

$$\text{Prob}\,\{(\bar{X} - \bar{Y}) - bS_d < (\mu_1 - \mu_2) < (\bar{X} - \bar{Y}) + bS_d\} = 0.95\,.$$

It follows that the random interval $\{(\bar{X} - \bar{Y}) - bS_d, (\bar{X} - \bar{Y}) + bS_d\}$ has a probability 0.95 of including the unknown fixed point $\mu_1 - \mu_2$. As usual, the experimental values of the means and

[2] If X_1 and X_2 are independent random variables, $\text{Var}(X_1 + X_2) = \text{Var}(X_1 - X_2) = \text{Var}\,X_1 + \text{Var}\,X_2$.

variances will provide a 95% confidence interval for $\mu_1 - \mu_2$ *when the variances of the two independent normal distributions are unknown but equal.*

Example:

$$n = 10,\ m = 7,\ \bar{x} = 4.2,\ \bar{y} = 3.4\ ;$$

$$s_1^2 = 54.44,\quad s_2^2 = 37.33,\quad d = \bar{x} - \bar{y} = 0.8;$$

$$s_d = \sqrt{\frac{9(54.44) + 6(37.33)}{15}\left(\frac{1}{10} + \frac{1}{7}\right)} = \sqrt{11.559} = 3.4\ ;$$

degrees of freedom = 15, $\quad b = 1.753$.

The 90% confidence interval for $\mu_1 - \mu_2$ is

$$(-5.16,\ 6.76)\ .$$

We shall not consider the difficulties encountered when the variances are unequal. It will be sufficient to say that exact confidence intervals are difficult to evaluate when this situation prevails. One approach to this problem yields the Behrens-Fisher solution. For $m = n$

$$S_d = \sqrt{\frac{S_1^2 + S_2^2}{n}},$$

and

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(S_1^2 + S_2^2)/n}}$$

has a t distribution with $2(n - 1)$ degrees of freedom.

4. CONFIDENCE INTERVALS FOR A VARIANCE

Consider a sequence of n independent random variables $X_1, X_2, \ldots, X_n$ which have the normal distributions $N(x_i;\ \mu,\ \sigma^2)$. The random variable

$$Y = \sum_{i=1}^{n} (X_i - \mu)^2/\sigma^2$$

has a chi-square probability density function with n degrees of freedom (see Appendix, Sect. 9). For a given probability, say 0.95, and for the fixed positive integer n, values of a and b $(0 < a < b)$ may be determined from the table of the chi-square distribution (Table 3) so that

$$\text{Prob}\ \{a < Y < b\} = 0.95\ ,$$

or

$$\text{Prob}\left\{a < \sum_{i=1}^{n} (X_i - \mu)^2/\sigma^2 < b\right\} = 0.95 \ , \tag{4.1}$$

or

$$\text{Prob}\left\{\sum_{i=1}^{n} (X_i - \mu)^2/b < \sigma^2 < \sum_{i=1}^{n} (X_i - \mu)^2/a\right\} = 0.95 \ .$$

If μ is known, then both $\Sigma(X_i - \mu)^2/b$ and $\Sigma(X_i - \mu)^2/a$ are statistics. Moreover, $[\Sigma(X_i - \mu)^2/b, \Sigma(X_i - \mu)^2/a]$ is a random interval having probability of 0.95 of including the unknown parameter σ^2.

When the random experiment has been run, then the particular interval $[\Sigma(x_i - \mu)^2/b, \Sigma(x_i - \mu)^2/a]$ is a 95% confidence interval for σ^2.

What is the situation, however, *when μ is unknown?*

The fact that $(n - 1)S^2/\sigma^2$ has a chi-square distribution with $n - 1$ degrees of freedom, whatever the value of μ, implies that inferences can be made about the unknown variance σ^2 even though μ is unknown.

With some preassigned probability, say 0.98, a and b $(0 < a < b)$ can be found from tables of the chi-square distribution so that

$$\text{Prob}\ \{a < (n - 1)S^2/\sigma^2 < b\} = 0.98 \ . \tag{4.2}$$

One convention for selecting a and b is to do so in such a way that

$$\int_0^a f(y)\ dy = \int_b^\infty f(y)\ dy = 0.01, \tag{4.3}$$

or

$$\text{Prob}\ \{(n - 1)S^2/\sigma^2 < a\} = 0.01 = \text{Prob}\ \{(n - 1)S^2/\sigma^2 > b\}$$

(see Table 3). Then

$$\text{Prob}\ \{(n - 1)S^2/b < \sigma^2 < (n - 1)S^2/a\} = 0.98 \ , \tag{4.4}$$

so that $[(n - 1)S^2/b, (n - 1)S^2/a]$ is a random interval having probability 0.98 of including the fixed but unknown parameter σ^2. After a random experiment is performed and $X_1 = x_1, \ldots, X_n = x_n$ with

$$s^2 = \Sigma(x_i - \bar{x})^2/(n - 1) \ ,$$

the 98% confidence interval for σ^2 is

$$[(n - 1)s^2/b, (n - 1)s^2/a] \ .$$

Example:

$$n = 9, \quad s^2 = 8.58 .$$

The 90% confidence interval for σ^2 is given by

$$\left[\frac{8(8.58)}{15.507}, \frac{8(8.58)}{2.733}\right],$$

or (4.43, 25.12), where $b = 15.507$ and $a = 2.733$.

5. CONFIDENCE INTERVALS FOR THE RATIO OF TWO VARIANCES

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two sequences of independent random variables having probability density functions $N(x; \mu_1, \sigma_1^2)$ and $N(y; \mu_2, \sigma_2^2)$ respectively. Let the means of the two sequences be denoted by $\bar{X}$ and $\bar{Y}$ and the variances by

$$S_1^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$$

and

$$S_2^2 = \sum_{j=1}^{m} (Y_j - \bar{Y})^2/(m-1) .$$

The random variables $(n-1)S_1^2/\sigma_1^2$ and $(m-1)S_2^2/\sigma_2^2$ are independent and have chi-square distributions with $n-1$ and $m-1$ degrees of freedom respectively. It follows, therefore, that the random variable

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{5.1}$$

has an F distribution with parameters $n-1$ and $m-1$ (see Appendix, Sect. 9). Thus, for given values of n and m, and with a preassigned probability 0.95 (say), two numbers a and b $(0 < a < b)$ can be determined from tables of the F distribution (Table 4a and b), so that

$$\text{Prob} \left\{ a < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < b \right\} = 0.95 , \tag{5.2}$$

or

$$\text{Prob} \{aS_2^2/S_1^2 < \sigma_2^2/\sigma_1^2 < bS_2^2/S_1^2\} = 0.95 .$$

Accordingly, $[aS_2^2/S_1^2,\ bS_2^2/S_1^2]$ is a random interval having probability 0.95 of including the fixed but unknown parameter σ_2^2/σ_1^2.

If the experimental values of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ respectively, and if

$$s_1^2 = \Sigma(x_i - \bar{x})^2/(n-1)$$

and

$$s_2^2 = \Sigma(y_j - \bar{y})^2/(m-1)\ ,$$

then the interval $[as_2^2/s_1^2,\ bs_2^2/s_1^2]$ is a 95% confidence interval for the ratio σ_2^2/σ_1^2 of the two unknown variances.

Example: Find the 90% confidence interval for (σ_2^2/σ_1^2) given that

$$n = 10, \quad m = 5, \quad s_1^2 = 20.0, \quad s_2^2 = 30.0, \quad (s_2^2/s_1^2) = 1.5\ .$$

$$\text{Prob}\left\{a\frac{s_2^2}{s_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < b\frac{s_2^2}{s_1^2}\right\} = 0.90 = \int_a^b g(f)\ df$$

$$= 1 - \int_0^a g(f)\ df - \int_b^\infty g(f)\ df\ , \tag{5.3}$$

where

$$\int_0^a g(f)\ df = \int_b^\infty g(f)\ df = 0.05\ . \tag{5.4}$$

$$\int_0^a g(f)\ df = \text{Prob}\ \{F_{9,4} < a\} = \text{Prob}\left\{\frac{1}{F_{9,4}} > \frac{1}{a}\right\}$$

$$= \text{Prob}\left\{F_{4,9} > \frac{1}{a}\right\} = 0.05\ . \tag{5.5}$$

$$\frac{1}{a} = 3.63, \quad a = \frac{1}{3.63}\ .$$

Similarly $b = 6.00$.

6. CONFIDENCE REGIONS FOR A VECTOR OF p MEANS

Define the p-dimensional random variable $\mathbf{x}'$[3] as the vector

$$\mathbf{x}' = [x_1, x_2, \ldots, x_p]$$

whose elements are jointly distributed as the p-variate normal distribution

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \qquad \text{for } -\infty < \mathbf{x} < \infty, \tag{6.1}$$

where

$$\begin{aligned}\mu' = E(\mathbf{x}') &= [E(x_1), E(x_2), \ldots, E(x_p)] \\ &= [\mu_1, \mu_2, \ldots, \mu_p]\end{aligned} \tag{6.2}$$

and

$$\Sigma = E\{[\mathbf{x} - E(\mathbf{x})][\mathbf{x} - E(\mathbf{x})]'\}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} \cdots \sigma_{1p} \\ \sigma_{12} & \sigma_{22} \cdots \sigma_{2p} \\ \vdots & \vdots \quad \ddots \quad \vdots \\ \sigma_{1p} & \sigma_{2p} \cdots \sigma_{pp} \end{bmatrix} . \tag{6.3}$$

On the basis of a sample of size n from this population, estimate the mean vector μ' by

$$\bar{\mathbf{x}}' = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p]$$

where

$$\bar{x}_i = 1/n \sum_{j=1}^{n} x_{ij}$$

for all $i = 1, 2, \ldots, p$, and the dispersion matrix Σ by

$$S = \begin{bmatrix} s_{11} & s_{12} \cdots s_{1p} \\ s_{12} & s_{22} \cdots s_{2p} \\ \vdots & \vdots \quad \ddots \quad \vdots \\ s_{1p} & s_{2p} \cdots s_{pp} \end{bmatrix} , \tag{6.4}$$

[3]The notational conventions in multivariate analysis are such that the use of lower-case letters for random variables as well as realizations of random variables is almost mandatory. Upper-case letters are used to designate matrices.

where

$$s_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)$$

for all i and $k = 1, 2, \ldots, p$.

The confidence region for μ is the volume enclosed by the p-dimensional hyperellipsoid

$$(\mu - \bar{\mathbf{x}})' S^{-1}(\mu - \bar{\mathbf{x}}) = \frac{(n-1)p}{n(n-p)} F_{(1-\beta;\, p,\, n-p)} , \tag{6.5}$$

where $F_{(1-\beta;\, p,\, n-p)}$ is the upper $100(1-\beta)\%$ point of the F distribution with p and $n-p$ degrees of freedom $(0 \leqq \beta \leqq 1)$. The risk incurred by falsely assigning the point $(\mu_1, \mu_2, \ldots, \mu_p)$ to this region is $100(1-\beta)\%$.

7. CONFIDENCE REGIONS FOR A BIVARIATE VECTOR OF MEANS

When $p = 2$, the random vector $\mathbf{x}' = (x_1, x_2)$ is distributed as a bivariate normal with mean vector $\mu' = (\mu_1, \mu_2)$ and dispersion matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} . \tag{7.1}$$

Estimate μ' by

$$\bar{\mathbf{x}}' = (\bar{x}_1, \bar{x}_2)$$

and Σ by

$$S = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix} , \tag{7.2}$$

where

$$\bar{x}_1 = \frac{1}{n} \sum_{j=1}^{n} x_{1j} ,$$

$$\bar{x}_2 = \frac{1}{n} \sum_{j=1}^{n} x_{2j} ,$$

$$s_{11} = s_1^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_{1j} - \bar{x}_1)^2 , \tag{7.3}$$

$$s_{22} = s_2^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_{2j} - \bar{x}_2)^2 ,$$

$$s_{12} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2) .$$

The confidence region for μ is the area enclosed by the ellipse

$$(\mu - \bar{x})' S^{-1} (\mu - \bar{x}) = \frac{2(n-1)}{n(n-2)} F_{(1-\beta;\ 2,\ n-2)} . \tag{7.4}$$

The risk incurred by assigning the point (μ_1, μ_2) to this region is $100(1 - \beta)\%$.

An example of several such confidence ellipses calculated from data collected on human chromosomes and plotted by the CALCOMP pen and ink plotter is presented in Fig. 1.

ORNL DWG. 65-1411

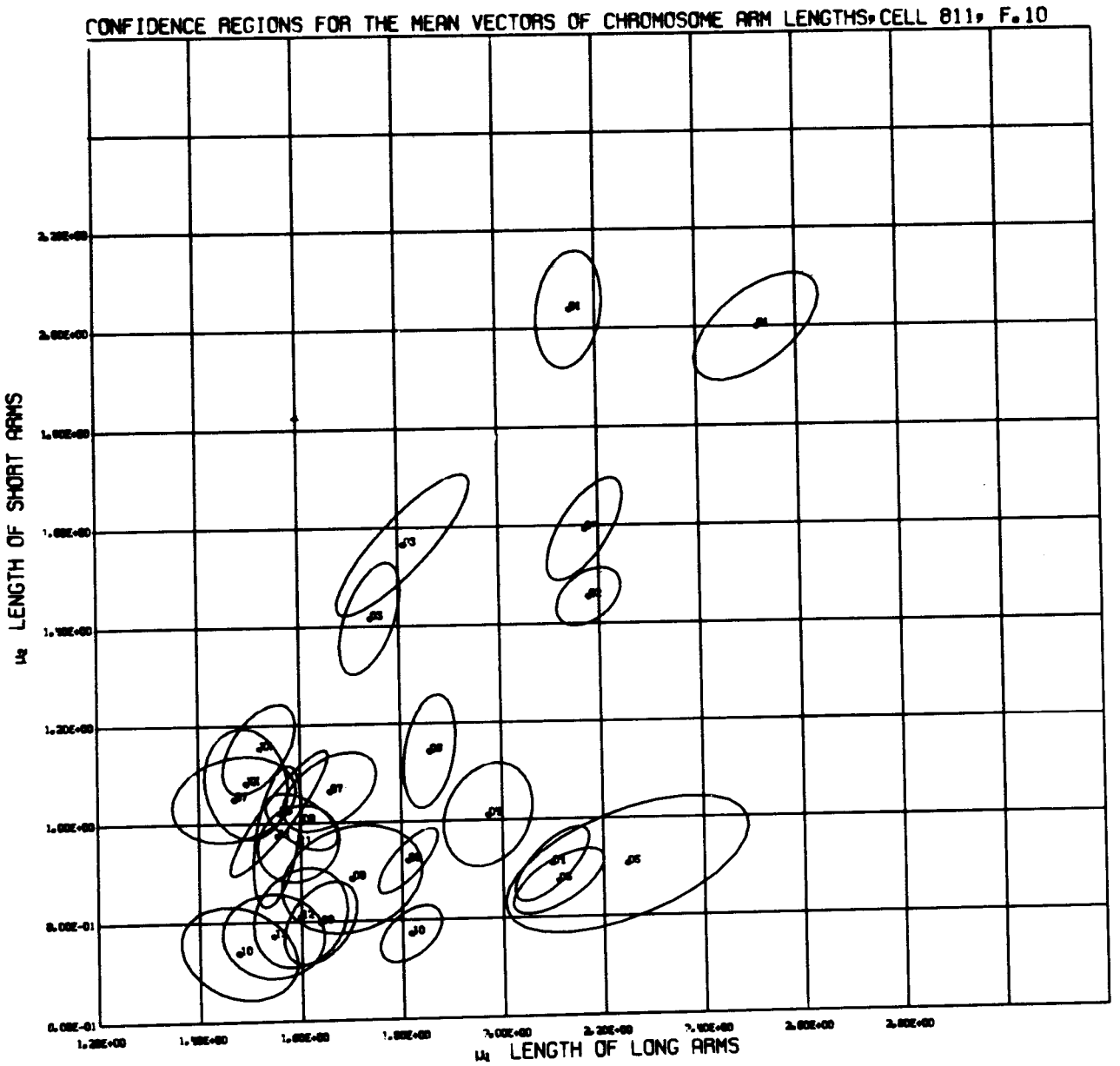


Fig. 1. Confidence Regions for the Mean Vectors of Chromosome Arm Lengths, Cell 811, F.10.

8. REFERENCES

Chew, Victor, "Confidence, Prediction, and Tolerance Regions for the Multivariate Normal Distribution," *J. Am. Statist. Assoc.* **61**, 605–17 (September 1966).

Crow, Edwin L., and Robert S. Gardner, "Confidence Intervals for the Expectation of a Poisson Variable," *Biometrika* **46**, 441–53 (December 1959).

9. APPENDIX

9.A The Chi-Square Distribution (χ^2): See Table 3

Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of n independent random variables each distributed as $N(z;\ 0,\ 1)$. The probability density function of the random variable

$$Y = Z_1^2 + Z_2^2 + \ldots + Z_n^2 \tag{A.1}$$

is

$$g_1(y) = \frac{1}{2^{n/2}\Gamma(n/2)}\, y^{(n/2)-1}\, e^{-y/2} \qquad \text{for } y > 0, \tag{A.2}$$

$$= 0 \qquad \text{for } y \leqq 0.$$

The corresponding cumulative distribution function is

$$G_1(y) = \text{Prob}\,\{Y \leqq y\} = \int_0^y g_1(w)\, dw\ . \tag{A.3}$$

This distribution is known as the χ^2 distribution. It contains one parameter n, which is denoted as the number of degrees of freedom.

9.B The t Distribution: See Table 2

Let Z be a random variable which is $N(z;\ 0,\ 1)$, and let Y be a random variable distributed as chi-square with n degrees of freedom; and let Z and Y be independent. The probability density function of the random variable

$$T = \frac{Z}{\sqrt{Y/n}} \tag{B.1}$$

is

$$g_2(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{(n+1)/2} \qquad \text{for } -\infty < t < \infty. \tag{B.2}$$

The corresponding cumulative distribution function is

$$G_2(t) = \text{Prob}\,\{T \leqq t\} = \int_{-\infty}^{t} g_2\,(w)\,dw\,. \tag{B.3}$$

The t distribution is completely specified by the parameter n, the number of degrees of freedom associated with the random variable Y having the chi-square distribution.

9.C The F Distribution: See Table 4*a* and *b*

Consider two independent chi-square variables U and Y having n_1 and n_2 degrees of freedom respectively. The probability density function of the random variable

$$F = \frac{U/n_1}{Y/n_2} \tag{C.1}$$

is

$$g_3(f) = \frac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)\left(\dfrac{n_1}{n_2}\right)^{n_1/2} f^{(n_1/2)-1}}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)\left(1+\dfrac{n_1 f}{n_2}\right)^{(n_1+n_2)/2}} \qquad \text{for } 0 < f < \infty, \tag{C.2}$$

$$= 0 \qquad \text{elsewhere.} \tag{C.2}$$

The corresponding cumulative distribution function is

$$G_3(f) = \text{Prob}\,\{F \leqq f\} = \int_0^f g_3(w)\,dw\,. \tag{C.3}$$

The F distribution is completely determined by two parameters n_1 and n_2, which are the degrees of freedom associated the two chi-square variables.

TABLE 1a. THE CUMULATIVE NORMAL DISTRIBUTION FUNCTION

$$F(z) = \Pr\{Z \leq z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\,dx \text{ for } -4.99 \leq z \leq 0.00.$$

z	·00	·01	·02	·03	·04	·05	·06	·07	·08	·09
− ·0	·5000	·4960	·4920	·4880	·4840	·4801	·4761	·4721	·4681	·4641
− ·1	·4602	·4562	·4522	·4483	·4443	·4404	·4364	·4325	·4286	·4247
− ·2	·4207	·4168	·4129	·4090	·4052	·4013	·3974	·3936	·3897	·3859
− ·3	·3821	·3783	·3745	·3707	·3669	·3632	·3594	·3557	·3520	·3483
− ·4	·3446	·3409	·3372	·3336	·3300	·3264	·3228	·3192	·3156	·3121
− ·5	·3085	·3050	·3015	·2981	·2946	·2912	·2877	·2843	·2810	·2776
− ·6	·2743	·2709	·2676	·2643	·2611	·2578	·2546	·2514	·2483	·2451
− ·7	·2420	·2389	·2358	·2327	·2297	·2266	·2236	·2206	·2177	·2148
− ·8	·2119	·2090	·2061	·2033	·2005	·1977	·1949	·1922	·1894	·1867
− ·9	·1841	·1814	·1788	·1762	·1736	·1711	·1685	·1660	·1635	·1611
−1·0	·1587	·1562	·1539	·1515	·1492	·1469	·1446	·1423	·1401	·1379
−1·1	·1357	·1335	·1314	·1292	·1271	·1251	·1230	·1210	·1190	·1170
−1·2	·1151	·1131	·1112	·1093	·1075	·1056	·1038	·1020	·1003	·09853
−1·3	·09680	·09510	·09342	·09176	·09012	·08851	·08691	·08534	·08379	·08226
−1·4	·08076	·07927	·07780	·07636	·07493	·07353	·07215	·07078	·06944	·06811
−1·5	·06681	·06552	·06426	·06301	·06178	·06057	·05938	·05821	·05705	·05592
−1·6	·05480	·05370	·05262	·05155	·05050	·04947	·04846	·04746	·04648	·04551
−1·7	·04457	·04363	·04272	·04182	·04093	·04006	·03920	·03836	·03754	·03673
−1·8	·03593	·03515	·03438	·03362	·03288	·03216	·03144	·03074	·03005	·02938
−1·9	·02872	·02807	·02743	·02680	·02619	·02559	·02500	·02442	·02385	·02330
−2·0	·02275	·02222	·02169	·02118	·02068	·02018	·01970	·01923	·01876	·01831
−2·1	·01786	·01743	·01700	·01659	·01618	·01578	·01539	·01500	·01463	·01426
−2·2	·01390	·01355	·01321	·01287	·01255	·01222	·01191	·01160	·01130	·01101
−2·3	·01072	·01044	·01017	·0^{2}9903	·0^{2}9642	·0^{2}9387	·0^{2}9137	·0^{2}8894	·0^{2}8656	·0^{2}8424
−2·4	·0^{2}8198	·0^{2}7976	·0^{2}7760	·0^{2}7549	·0^{2}7344	·0^{2}7143	·0^{2}6947	·0^{2}6756	·0^{2}6569	·0^{2}6387
−2·5	·0^{2}6210	·0^{2}6037	·0^{2}5868	·0^{2}5703	·0^{2}5543	·0^{2}5386	·0^{2}5234	·0^{2}5085	·0^{2}4940	·0^{2}4799
−2·6	·0^{2}4661	·0^{2}4527	·0^{2}4396	·0^{2}4269	·0^{2}4145	·0^{2}4025	·0^{2}3907	·0^{2}3793	·0^{2}3681	·0^{2}3573
−2·7	·0^{2}3467	·0^{2}3364	·0^{2}3264	·0^{2}3167	·0^{2}3072	·0^{2}2980	·0^{2}2890	·0^{2}2803	·0^{2}2718	·0^{2}2635
−2·8	·0^{2}2555	·0^{2}2477	·0^{2}2401	·0^{2}2327	·0^{2}2256	·0^{2}2186	·0^{2}2118	·0^{2}2052	·0^{2}1988	·0^{2}1926
−2·9	·0^{2}1866	·0^{2}1807	·0^{2}1750	·0^{2}1695	·0^{2}1641	·0^{2}1589	·0^{2}1538	·0^{2}1489	·0^{2}1441	·0^{2}1395
−3·0	·0^{2}1350	·0^{2}1306	·0^{2}1264	·0^{2}1223	·0^{2}1183	·0^{2}1144	·0^{2}1107	·0^{2}1070	·0^{2}1035	·0^{2}1001
−3·1	·0^{3}9676	·0^{3}9354	·0^{3}9043	·0^{3}8740	·0^{3}8447	·0^{3}8164	·0^{3}7888	·0^{3}7622	·0^{3}7364	·0^{3}7114
−3·2	·0^{3}6871	·0^{3}6637	·0^{3}6410	·0^{3}6190	·0^{3}5976	·0^{3}5770	·0^{3}5571	·0^{3}5377	·0^{3}5190	·0^{3}5009
−3·3	·0^{3}4834	·0^{3}4665	·0^{3}4501	·0^{3}4342	·0^{3}4189	·0^{3}4041	·0^{3}3897	·0^{3}3758	·0^{3}3624	·0^{3}3495
−3·4	·0^{3}3369	·0^{3}3248	·0^{3}3131	·0^{3}3018	·0^{3}2909	·0^{3}2803	·0^{3}2701	·0^{3}2602	·0^{3}2507	·0^{3}2415
−3·5	·0^{3}2326	·0^{3}2241	·0^{3}2158	·0^{3}2078	·0^{3}2001	·0^{3}1926	·0^{3}1854	·0^{3}1785	·0^{3}1718	·0^{3}1653
−3·6	·0^{3}1591	·0^{3}1531	·0^{3}1473	·0^{3}1417	·0^{3}1363	·0^{3}1311	·0^{3}1261	·0^{3}1213	·0^{3}1166	·0^{3}1121
−3·7	·0^{3}1078	·0^{3}1036	·0^{4}9961	·0^{4}9574	·0^{4}9201	·0^{4}8842	·0^{4}8496	·0^{4}8162	·0^{4}7841	·0^{4}7532
−3·8	·0^{4}7235	·0^{4}6948	·0^{4}6673	·0^{4}6407	·0^{4}6152	·0^{4}5906	·0^{4}5669	·0^{4}5442	·0^{4}5223	·0^{4}5012
−3·9	·0^{4}4810	·0^{4}4615	·0^{4}4427	·0^{4}4247	·0^{4}4074	·0^{4}3908	·0^{4}3747	·0^{4}3594	·0^{4}3446	·0^{4}3304
−4·0	·0^{4}3167	·0^{4}3036	·0^{4}2910	·0^{4}2789	·0^{4}2673	·0^{4}2561	·0^{4}2454	·0^{4}2351	·0^{4}2252	·0^{4}2157
−4·1	·0^{4}2066	·0^{4}1978	·0^{4}1894	·0^{4}1814	·0^{4}1737	·0^{4}1662	·0^{4}1591	·0^{4}1523	·0^{4}1458	·0^{4}1395
−4·2	·0^{4}1335	·0^{4}1277	·0^{4}1222	·0^{4}1168	·0^{4}1118	·0^{4}1069	·0^{4}1022	·0^{5}9774	·0^{5}9345	·0^{5}8934
−4·3	·0^{5}8540	·0^{5}8163	·0^{5}7801	·0^{5}7455	·0^{5}7124	·0^{5}6807	·0^{5}6503	·0^{5}6212	·0^{5}5934	·0^{5}5668
−4·4	·0^{5}5413	·0^{5}5169	·0^{5}4935	·0^{5}4712	·0^{5}4498	·0^{5}4294	·0^{5}4098	·0^{5}3911	·0^{5}3732	·0^{5}3561
−4·5	·0^{5}3398	·0^{5}3241	·0^{5}3092	·0^{5}2949	·0^{5}2813	·0^{5}2682	·0^{5}2558	·0^{5}2439	·0^{5}2325	·0^{5}2216
−4·6	·0^{5}2112	·0^{5}2013	·0^{5}1919	·0^{5}1828	·0^{5}1742	·0^{5}1660	·0^{5}1581	·0^{5}1506	·0^{5}1434	·0^{5}1366
−4·7	·0^{5}1301	·0^{5}1239	·0^{5}1179	·0^{5}1123	·0^{5}1069	·0^{5}1017	·0^{6}9680	·0^{6}9211	·0^{6}8765	·0^{6}8339
−4·8	·0^{6}7933	·0^{6}7547	·0^{6}7178	·0^{6}6827	·0^{6}6492	·0^{6}6173	·0^{6}5869	·0^{6}5580	·0^{6}5304	·0^{6}5042
−4·9	·0^{6}4792	·0^{6}4554	·0^{6}4327	·0^{6}4111	·0^{6}3906	·0^{6}3711	·0^{6}3525	·0^{6}3348	·0^{6}3179	·0^{6}3019

Example: $F(-1.96) = 0.02500$

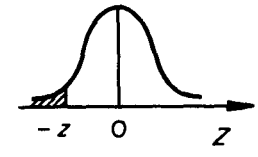


TABLE 1b. THE CUMULATIVE NORMAL DISTRIBUTION FUNCTION

$$F(z) = \Pr\{Z \leq z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx \text{ for } 0.00 \leq z \leq 4.99.$$

z	·00	·01	·02	·03	·04	·05	·06	·07	·08	·09
·0	·5000	·5040	·5080	·5120	·5160	·5199	·5239	·5279	·5319	·5359
·1	·5398	·5438	·5478	·5517	·5557	·5596	·5636	·5675	·5714	·5753
·2	·5793	5832	·5871	·5910	·5948	·5987	·6026	·6064	·6103	·6141
·3	·6179	·6217	·6255	·6293	·6331	·6368	·6406	·6443	·6480	·6517
·4	·6554	·6591	·6628	·6664	·6700	·6736	·6772	·6808	·6844	·6879
·5	·6915	·6950	·6985	·7019	·7054	·7088	·7123	·7157	·7190	·7224
·6	·7257	·7291	·7324	·7357	·7389	·7422	·7454	·7486	·7517	·7549
·7	·7580	·7611	·7642	·7673	·7703	·7734	·7764	·7794	·7823	·7852
·8	·7881	·7910	·7939	·7967	·7995	·8023	·8051	·8078	·8106	·8133
·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	·90147
1·3	·90320	·90490	·90658	·90824	·90988	·91149	·91309	·91466	·91621	·91774
1·4	·91924	·92073	·92220	·92364	·92507	·92647	·92785	·92922	·93056	·93189
1·5	·93319	·93448	·93574	·93699	·93822	·93943	·94062	·94179	·94295	·94408
1·6	·94520	·94630	·94738	·94845	·94950	·95053	·95154	·95254	·95352	·95449
1·7	·95543	·95637	·95728	·95818	·95907	·95994	·96080	·96164	·96246	·96327
1·8	·96407	·96485	·96562	·96638	·96712	·96784	·96856	·96926	·96995	·97062
1·9	·97128	·97193	·97257	·97320	·97381	·97441	·97500	·97558	·97615	·97670
2·0	·97725	·97778	·97831	·97882	·97932	·97982	·98030	·98077	·98124	·98169
2·1	·98214	·98257	·98300	·98341	·98382	·98422	·98461	·98500	·98537	·98574
2·2	·98610	·98645	·98679	·98713	·98745	·98778	·98809	·98840	·98870	·98899
2·3	·98928	·98956	·98983	$\cdot 9^{2}0097$	$\cdot 9^{2}0358$	$\cdot 9^{2}0613$	$\cdot 9^{2}0863$	$\cdot 9^{2}1106$	$\cdot 9^{2}1344$	$\cdot 9^{2}1576$
2·4	$\cdot 9^{2}1802$	$\cdot 9^{2}2024$	$\cdot 9^{2}2240$	$\cdot 9^{2}2451$	$\cdot 9^{2}2656$	$\cdot 9^{2}2857$	$\cdot 9^{2}3053$	$\cdot 9^{2}3244$	$\cdot 9^{2}3431$	$\cdot 9^{2}3613$
2·5	$\cdot 9^{2}3790$	$\cdot 9^{2}3963$	$\cdot 9^{2}4132$	$\cdot 9^{2}4297$	$\cdot 9^{2}4457$	$\cdot 9^{2}4614$	$\cdot 9^{2}4766$	$\cdot 9^{2}4915$	$\cdot 9^{2}5060$	$\cdot 9^{2}5201$
2·6	$\cdot 9^{2}5339$	$\cdot 9^{2}5473$	$\cdot 9^{2}5604$	$\cdot 9^{2}5731$	$\cdot 9^{2}5855$	$\cdot 9^{2}5975$	$\cdot 9^{2}6093$	$\cdot 9^{2}6207$	$\cdot 9^{2}6319$	$\cdot 9^{2}6427$
2·7	$\cdot 9^{2}6533$	$\cdot 9^{2}6636$	$\cdot 9^{2}6736$	$\cdot 9^{2}6833$	$\cdot 9^{2}6928$	$\cdot 9^{2}7020$	$\cdot 9^{2}7110$	$\cdot 9^{2}7197$	$\cdot 9^{2}7282$	$\cdot 9^{2}7365$
2·8	$\cdot 9^{2}7445$	$\cdot 9^{2}7523$	$\cdot 9^{2}7599$	$\cdot 9^{2}7673$	$\cdot 9^{2}7744$	$\cdot 9^{2}7814$	$\cdot 9^{2}7882$	$\cdot 9^{2}7948$	$\cdot 9^{2}8012$	$\cdot 9^{2}8074$
2·9	$\cdot 9^{2}8134$	$\cdot 9^{2}8193$	$\cdot 9^{2}8250$	$\cdot 9^{2}8305$	$\cdot 9^{2}8359$	$\cdot 9^{2}8411$	$\cdot 9^{2}8462$	$\cdot 9^{2}8511$	$\cdot 9^{2}8559$	$\cdot 9^{2}8605$
3·0	$\cdot 9^{2}8650$	$\cdot 9^{2}8694$	$\cdot 9^{2}8736$	$\cdot 9^{2}8777$	$\cdot 9^{2}8817$	$\cdot 9^{2}8856$	$\cdot 9^{2}8893$	$\cdot 9^{2}8930$	$\cdot 9^{2}8965$	$\cdot 9^{2}8999$
3·1	$\cdot 9^{3}0324$	$\cdot 9^{3}0646$	$\cdot 9^{3}0957$	$\cdot 9^{3}1260$	$\cdot 9^{3}1553$	$\cdot 9^{3}1836$	$\cdot 9^{3}2112$	$\cdot 9^{3}2378$	$\cdot 9^{3}2636$	$\cdot 9^{3}2886$
3·2	$\cdot 9^{3}3129$	$\cdot 9^{3}3363$	$\cdot 9^{3}3590$	$\cdot 9^{3}3810$	$\cdot 9^{3}4024$	$\cdot 9^{3}4230$	$\cdot 9^{3}4429$	$\cdot 9^{3}4623$	$\cdot 9^{3}4810$	$\cdot 9^{3}4991$
3·3	$\cdot 9^{3}5166$	$\cdot 9^{3}5335$	$\cdot 9^{3}5499$	$\cdot 9^{3}5658$	$\cdot 9^{3}5811$	$\cdot 9^{3}5959$	$\cdot 9^{3}6103$	$\cdot 9^{3}6242$	$\cdot 9^{3}6376$	$\cdot 9^{3}6505$
3·4	$\cdot 9^{3}6631$	$\cdot 9^{3}6752$	$\cdot 9^{3}6869$	$\cdot 9^{3}6982$	$\cdot 9^{3}7091$	$\cdot 9^{3}7197$	$\cdot 9^{3}7299$	$\cdot 9^{3}7398$	$\cdot 9^{3}7493$	$\cdot 9^{3}7585$
3·5	$\cdot 9^{3}7674$	$\cdot 9^{3}7759$	$\cdot 9^{3}7842$	$\cdot 9^{3}7922$	$\cdot 9^{3}7999$	$\cdot 9^{3}8074$	$\cdot 9^{3}8146$	$\cdot 9^{3}8215$	$\cdot 9^{3}8282$	$\cdot 9^{3}8347$
3·6	$\cdot 9^{3}8409$	$\cdot 9^{3}8469$	$\cdot 9^{3}8527$	$\cdot 9^{3}8583$	$\cdot 9^{3}8637$	$\cdot 9^{3}8689$	$\cdot 9^{3}8739$	$\cdot 9^{3}8787$	$\cdot 9^{3}8834$	$\cdot 9^{3}8879$
3·7	$\cdot 9^{3}8922$	$\cdot 9^{3}8964$	$\cdot 9^{4}0039$	$\cdot 9^{4}0426$	$\cdot 9^{4}0799$	$\cdot 9^{4}1158$	$\cdot 9^{4}1504$	$\cdot 9^{4}1838$	$\cdot 9^{4}2159$	$\cdot 9^{4}2468$
3·8	$\cdot 9^{4}2765$	$\cdot 9^{4}3052$	$\cdot 9^{4}3327$	$\cdot 9^{4}3593$	$\cdot 9^{4}3848$	$\cdot 9^{4}4094$	$\cdot 9^{4}4331$	$\cdot 9^{4}4558$	$\cdot 9^{4}4777$	$\cdot 9^{4}4988$
3·9	$\cdot 9^{4}5190$	$\cdot 9^{4}5385$	$\cdot 9^{4}5573$	$\cdot 9^{4}5753$	$\cdot 9^{4}5926$	$\cdot 9^{4}6092$	$\cdot 9^{4}6253$	$\cdot 9^{4}6406$	$\cdot 9^{4}6554$	$\cdot 9^{4}6696$
4·0	$\cdot 9^{4}6833$	$\cdot 9^{4}6964$	$\cdot 9^{4}7090$	$\cdot 9^{4}7211$	$\cdot 9^{4}7327$	$\cdot 9^{4}7439$	$\cdot 9^{4}7546$	$\cdot 9^{4}7649$	$\cdot 9^{4}7748$	$\cdot 9^{4}7843$
4·1	$\cdot 9^{4}7934$	$\cdot 9^{4}8022$	$\cdot 9^{4}8106$	$\cdot 9^{4}8186$	$\cdot 9^{4}8263$	$\cdot 9^{4}8338$	$\cdot 9^{4}8409$	$\cdot 9^{4}8477$	$\cdot 9^{4}8542$	$\cdot 9^{4}8605$
4·2	$\cdot 9^{4}8665$	$\cdot 9^{4}8723$	$\cdot 9^{4}8778$	$\cdot 9^{4}8832$	$\cdot 9^{4}8882$	$\cdot 9^{4}8931$	$\cdot 9^{4}8978$	$\cdot 9^{5}0226$	$\cdot 9^{5}0655$	$\cdot 9^{5}1066$
4·3	$\cdot 9^{5}1460$	$\cdot 9^{5}1837$	$\cdot 9^{5}2199$	$\cdot 9^{5}2545$	$\cdot 9^{5}2876$	$\cdot 9^{5}3193$	$\cdot 9^{5}3497$	$\cdot 9^{5}3788$	$\cdot 9^{5}4066$	$\cdot 9^{5}4332$
4·4	$\cdot 9^{5}4587$	$\cdot 9^{5}4831$	$\cdot 9^{5}5065$	$\cdot 9^{5}5288$	$\cdot 9^{5}5502$	$\cdot 9^{5}5706$	$\cdot 9^{5}5902$	$\cdot 9^{5}6089$	$\cdot 9^{5}6268$	$\cdot 9^{5}6439$
4·5	$\cdot 9^{5}6602$	$\cdot 9^{5}6759$	$\cdot 9^{5}6908$	$\cdot 9^{5}7051$	$\cdot 9^{5}7187$	$\cdot 9^{5}7318$	$\cdot 9^{5}7442$	$\cdot 9^{5}7561$	$\cdot 9^{5}7675$	$\cdot 9^{5}7784$
4·6	$\cdot 9^{5}7888$	$\cdot 9^{5}7987$	$\cdot 9^{5}8081$	$\cdot 9^{5}8172$	$\cdot 9^{5}8258$	$\cdot 9^{5}8340$	$\cdot 9^{5}8419$	$\cdot 9^{5}8494$	$\cdot 9^{5}8566$	$\cdot 9^{5}8634$
4·7	$\cdot 9^{5}8699$	$\cdot 9^{5}8761$	$\cdot 9^{5}8821$	$\cdot 9^{5}8877$	$\cdot 9^{5}8931$	$\cdot 9^{5}8983$	$\cdot 9^{6}0320$	$\cdot 9^{6}0789$	$\cdot 9^{6}1235$	$\cdot 9^{6}1661$
4·8	$\cdot 9^{6}2067$	$\cdot 9^{6}2453$	$\cdot 9^{6}2822$	$\cdot 9^{6}3173$	$\cdot 9^{6}3508$	$\cdot 9^{6}3827$	$\cdot 9^{6}4131$	$\cdot 9^{6}4420$	$\cdot 9^{6}4696$	$\cdot 9^{6}4958$
4·9	$\cdot 9^{6}5208$	$\cdot 9^{6}5446$	$\cdot 9^{6}5673$	$\cdot 9^{6}5889$	$\cdot 9^{6}6094$	$\cdot 9^{6}6289$	$\cdot 9^{6}6475$	$\cdot 9^{6}6652$	$\cdot 9^{6}6821$	$\cdot 9^{6}6981$

Example: $F(2.57) = .9^{2}4915 = 0.994915$

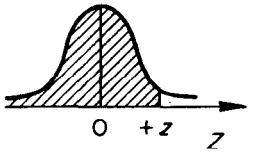


TABLE 2. DISTRIBUTION OF t

P = Probability.

n	·9	·8	·7	·6	·5	·4	·3	·2	·1	·05	·02	·01	·001
1	·158	·325	·510	·727	1·000	1·376	1·963	3·078	6·314	12·706	31·821	63·657	636·619
2	·142	·289	·445	·617	·816	1·061	1·386	1·886	2·920	4·303	6·965	9·925	31·598
3	·137	·277	·424	·584	·765	·978	1·250	1·638	2·353	3·182	4·541	5·841	12·941
4	·134	·271	·414	·569	·741	·941	1·190	1·533	2·132	2·776	3·747	4·604	8·610
5	·132	·267	·408	·559	·727	·920	1·156	1·476	2·015	2·571	3·365	4·032	6·859
6	·131	·265	·404	·553	·718	·906	1·134	1·440	1·943	2·447	3·143	3·707	5·959
7	·130	·263	·402	·549	·711	·896	1·119	1·415	1·895	2·365	2·998	3·499	5·405
8	·130	·262	·399	·546	·706	·889	1·108	1·397	1·860	2·306	2·896	3·355	5·041
9	·129	·261	·398	·543	·703	·883	1·100	1·383	1·833	2·262	2·821	3·250	4·781
10	·129	·260	·397	·542	·700	·879	1·093	1·372	1·812	2·228	2·764	3·169	4·587
11	·129	·260	·396	·540	·697	·876	1·088	1·363	1·796	2·201	2·718	3·106	4·437
12	·128	·259	·395	·539	·695	·873	1·083	1·356	1·782	2·179	2·681	3·055	4·318
13	·128	·259	·394	·538	·694	·870	1·079	1·350	1·771	2·160	2·650	3·012	4·221
14	·128	·258	·393	·537	·692	·868	1·076	1·345	1·761	2·145	2·624	2·977	4·140
15	·128	·258	·393	·536	·691	·866	1·074	1·341	1·753	2·131	2·602	2·947	4·073
16	·128	·258	·392	·535	·690	·865	1·071	1·337	1·746	2·120	2·583	2·921	4·015
17	·128	·257	·392	·534	·689	·863	1·069	1·333	1·740	2·110	2·567	2·898	3·965
18	·127	·257	·392	·534	·688	·862	1·067	1·330	1·734	2·101	2·552	2·878	3·922
19	·127	·257	·391	·533	·688	·861	1·066	1·328	1·729	2·093	2·539	2·861	3·883
20	·127	·257	·391	·533	·687	·860	1·064	1·325	1·725	2·086	2·528	2·845	3·850
21	·127	·257	·391	·532	·686	·859	1·063	1·323	1·721	2·080	2·518	2·831	3·819
22	·127	·256	·390	·532	·686	·858	1·061	1·321	1·717	2·074	2·508	2·819	3·792
23	·127	·256	·390	·532	·685	·858	1·060	1·319	1·714	2·069	2·500	2·807	3·767
24	·127	·256	·390	·531	·685	·857	1·059	1·318	1·711	2·064	2·492	2·797	3·745
25	·127	·256	·390	·531	·684	·856	1·058	1·316	1·708	2·060	2·485	2·787	3·725
26	·127	·256	·390	·531	·684	·856	1·058	1·315	1·706	2·056	2·479	2·779	3·707
27	·127	·256	·389	·531	·684	·855	1·057	1·314	1·703	2·052	2·473	2·771	3·690
28	·127	·256	·389	·530	·683	·855	1·056	1·313	1·701	2·048	2·467	2·763	3·674
29	·127	·256	·389	·530	·683	·854	1·055	1·311	1·699	2·045	2·462	2·756	3·659
30	·127	·256	·389	·530	·683	·854	1·055	1·310	1·697	2·042	2·457	2·750	3·646
40	·126	·255	·388	·529	·681	·851	1·050	1·303	1·684	2·021	2·423	2·704	3·551
60	·126	·254	·387	·527	·679	·848	1·046	1·296	1·671	2·000	2·390	2·660	3·460
120	·126	·254	·386	·526	·677	·845	1·041	1·289	1·658	1·980	2·358	2·617	3·373
∞	·126	·253	·385	·524	·674	·842	1·036	1·282	1·645	1·960	2·326	2·576	3·291

$$
\begin{aligned}
P = 1 - \Pr\{-b < T < b\} &= \Pr\{T < -b\} + \Pr\{T > +b\} \\
&= \int_{-\infty}^{-b} f(t)\,dt + \int_{b}^{\infty} f(t)\,dt \\
&= 1 - \int_{-b}^{b} f(t)\,dt \\
&= 2\int_{b}^{\infty} f(t)\,dt
\end{aligned}
$$

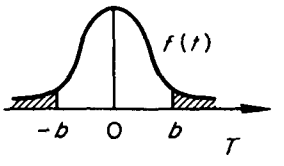


TABLE 3. DISTRIBUTION OF χ^2

P = Probability.

deg of freedom	·99	·98	·95	·90	·80	·70	·50	·30	·20	·10	·05	·02	·01	·001
1	$\cdot 0^3157$	$\cdot 0^3628$	·00393	·0158	·0642	·148	·455	1·074	1·642	2·706	3·841	5·412	6·635	10·827
2	·0201	·0404	·103	·211	·446	·713	1·386	2·408	3·219	4·605	5·991	7·824	9·210	13·815
3	·115	·185	·352	·584	1·005	1·424	2·366	3·665	4·642	6·251	7·815	9·837	11·345	16·268
4	·297	·429	·711	1·064	1·649	2·195	3·357	4·878	5·989	7·779	9·488	11·668	13·277	18·465
5	·554	·752	1·145	1·610	2·343	3·000	4·351	6·064	7·289	9·236	11·070	13·388	15·086	20·517
6	·872	1·134	1·635	2·204	3·070	3·828	5·348	7·231	8·558	10·645	12·592	15·033	16·812	22·457
7	1·239	1·564	2·167	2·833	3·822	4·671	6·346	8·383	9·803	12·017	14·067	16·622	18·475	24·322
8	1·646	2·032	2·733	3·490	4·594	5·527	7·344	9·524	11·030	13·362	15·507	18·168	20·090	26·125
9	2·088	2·532	3·325	4·168	5·380	6·393	8·343	10·656	12·242	14·684	16·919	19·679	21·666	27·877
10	2·558	3·059	3·940	4·865	6·179	7·267	9·342	11·781	13·442	15·987	18·307	21·161	23·209	29·588
11	3·053	3·609	4·575	5·578	6·989	8·148	10·341	12·899	14·631	17·275	19·675	22·618	24·725	31·264
12	3·571	4·178	5·226	6·304	7·807	9·034	11·340	14·011	15·812	18·549	21·026	24·054	26·217	32·909
13	4·107	4·765	5·892	7·042	8·634	9·926	12·340	15·119	16·985	19·812	22·362	25·472	27·688	34·528
14	4·660	5·368	6·571	7·790	9·467	10·821	13·339	16·222	18·151	21·064	23·685	26·873	29·141	36·123
15	5·229	5·985	7·261	8·547	10·307	11·721	14·339	17·322	19·311	22·307	24·996	28·259	30·578	37·697
16	5·812	6·614	7·962	9·312	11·152	12·624	15·338	18·418	20·465	23·542	26·296	29·633	32·000	39·252
17	6·408	7·255	8·672	10·085	12·002	13·531	16·338	19·511	21·615	24·769	27·587	30·995	33·409	40·790
18	7·015	7·906	9·390	10·865	12·857	14·440	17·338	20·601	22·760	25·989	28·869	32·346	34·805	42·312
19	7·633	8·567	10·117	11·651	13·716	15·352	18·338	21·689	23·900	27·204	30·144	33·687	36·191	43·820
20	8·260	9·237	10·851	12·443	14·578	16·266	19·337	22·775	25·038	28·412	31·410	35·020	37·566	45·315
21	8·897	9·915	11·591	13·240	15·445	17·182	20·337	23·858	26·171	29·615	32·671	36·343	38·932	46·797
22	9·542	10·600	12·338	14·041	16·314	18·101	21·337	24·939	27·301	30·813	33·924	37·659	40·289	48·268
23	10·196	11·293	13·091	14·848	17·187	19·021	22·337	26·018	28·429	32·007	35·172	38·968	41·638	49·728
24	10·856	11·992	13·848	15·659	18·062	19·943	23·337	27·096	29·553	33·196	36·415	40·270	42·980	51·179
25	11·524	12·697	14·611	16·473	18·940	20·867	24·337	28·172	30·675	34·382	37·652	41·566	44·314	52·620
26	12·198	13·409	15·379	17·292	19·820	21·792	25·336	29·246	31·795	35·563	38·885	42·856	45·642	54·052
27	12·879	14·125	16·151	18·114	20·703	22·719	26·336	30·319	32·912	36·741	40·113	44·140	46·963	55·476
28	13·565	14·847	16·928	18·939	21·588	23·647	27·336	31·391	34·027	37·916	41·337	45·419	48·278	56·893
29	14·256	15·574	17·708	19·768	22·475	24·577	28·336	32·461	35·139	39·087	42·557	46·693	49·588	58·302
30	14·953	16·306	18·493	20·599	23·364	25·508	29·336	33·530	36·250	40·256	43·773	47·962	50·892	59·703

For larger values of n, the expression $\sqrt{2\chi^2}-\sqrt{2n-1}$ may be used as a normal deviate with unit variance, remembering that the probability for χ^2 corresponds with that of a single tail of the normal curve.

$$P = \Pr\{Y > b\} = \int_b^\infty f(y)\,dy = 1 - \int_0^b f(y)\,dy$$

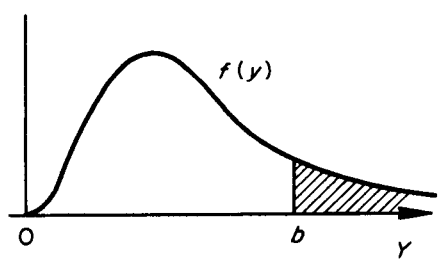
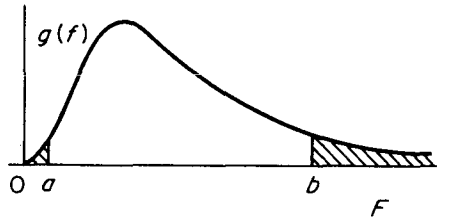


TABLE 4a. DISTRIBUTION OF F.



$$\Pr\{F_{\nu_1,\nu_2} > b\} = \int_b^\infty g(f)\,df$$

Upper 5% points

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161·4	199·5	215·7	224·6	230·2	234·0	236·8	238·9	240·5	241·9	243·9	245·9	248·0	249·1	250·1	251·1	252·2	253·3	254·3
2	18·51	19·00	19·16	19·25	19·30	19·33	19·35	19·37	19·38	19·40	19·41	19·43	19·45	19·45	19·46	19·47	19·48	19·49	19·50
3	10·13	9·55	9·28	9·12	9·01	8·94	8·89	8·85	8·81	8·79	8·74	8·70	8·66	8·64	8·62	8·59	8·57	8·55	8·53
4	7·71	6·94	6·59	6·39	6·26	6·16	6·09	6·04	6·00	5·96	5·91	5·86	5·80	5·77	5·75	5·72	5·69	5·66	5·63
5	6·61	5·79	5·41	5·19	5·05	4·95	4·88	4·82	4·77	4·74	4·68	4·62	4·56	4·53	4·50	4·46	4·43	4·40	4·36
6	5·99	5·14	4·76	4·53	4·39	4·28	4·21	4·15	4·10	4·06	4·00	3·94	3·87	3·84	3·81	3·77	3·74	3·70	3·67
7	5·59	4·74	4·35	4·12	3·97	3·87	3·79	3·73	3·68	3·64	3·57	3·51	3·44	3·41	3·38	3·34	3·30	3·27	3·23
8	5·32	4·46	4·07	3·84	3·69	3·58	3·50	3·44	3·39	3·35	3·28	3·22	3·15	3·12	3·08	3·04	3·01	2·97	2·93
9	5·12	4·26	3·86	3·63	3·48	3·37	3·29	3·23	3·18	3·14	3·07	3·01	2·94	2·90	2·86	2·83	2·79	2·75	2·71
10	4·96	4·10	3·71	3·48	3·33	3·22	3·14	3·07	3·02	2·98	2·91	2·85	2·77	2·74	2·70	2·66	2·62	2·58	2·54
11	4·84	3·98	3·59	3·36	3·20	3·09	3·01	2·95	2·90	2·85	2·79	2·72	2·65	2·61	2·57	2·53	2·49	2·45	2·40
12	4·75	3·89	3·49	3·26	3·11	3·00	2·91	2·85	2·80	2·75	2·69	2·62	2·54	2·51	2·47	2·43	2·38	2·34	2·30
13	4·67	3·81	3·41	3·18	3·03	2·92	2·83	2·77	2·71	2·67	2·60	2·53	2·46	2·42	2·38	2·34	2·30	2·25	2·21
14	4·60	3·74	3·34	3·11	2·96	2·85	2·76	2·70	2·65	2·60	2·53	2·46	2·39	2·35	2·31	2·27	2·22	2·18	2·13
15	4·54	3·68	3·29	3·06	2·90	2·79	2·71	2·64	2·59	2·54	2·48	2·40	2·33	2·29	2·25	2·20	2·16	2·11	2·07
16	4·49	3·63	3·24	3·01	2·85	2·74	2·66	2·59	2·54	2·49	2·42	2·35	2·28	2·24	2·19	2·15	2·11	2·06	2·01
17	4·45	3·59	3·20	2·96	2·81	2·70	2·61	2·55	2·49	2·45	2·38	2·31	2·23	2·19	2·15	2·10	2·06	2·01	1·96
18	4·41	3·55	3·16	2·93	2·77	2·66	2·58	2·51	2·46	2·41	2·34	2·27	2·19	2·15	2·11	2·06	2·02	1·97	1·92
19	4·38	3·52	3·13	2·90	2·74	2·63	2·54	2·48	2·42	2·38	2·31	2·23	2·16	2·11	2·07	2·03	1·98	1·93	1·88
20	4·35	3·49	3·10	2·87	2·71	2·60	2·51	2·45	2·39	2·35	2·28	2·20	2·12	2·08	2·04	1·99	1·95	1·90	1·84
21	4·32	3·47	3·07	2·84	2·68	2·57	2·49	2·42	2·37	2·32	2·25	2·18	2·10	2·05	2·01	1·96	1·92	1·87	1·81
22	4·30	3·44	3·05	2·82	2·66	2·55	2·46	2·40	2·34	2·30	2·23	2·15	2·07	2·03	1·98	1·94	1·89	1·84	1·78
23	4·28	3·42	3·03	2·80	2·64	2·53	2·44	2·37	2·32	2·27	2·20	2·13	2·05	2·01	1·96	1·91	1·86	1·81	1·76
24	4·26	3·40	3·01	2·78	2·62	2·51	2·42	2·36	2·30	2·25	2·18	2·11	2·03	1·98	1·94	1·89	1·84	1·79	1·73
25	4·24	3·39	2·99	2·76	2·60	2·49	2·40	2·34	2·28	2·24	2·16	2·09	2·01	1·96	1·92	1·87	1·82	1·77	1·71
26	4·23	3·37	2·98	2·74	2·59	2·47	2·39	2·32	2·27	2·22	2·15	2·07	1·99	1·95	1·90	1·85	1·80	1·75	1·69
27	4·21	3·35	2·96	2·73	2·57	2·46	2·37	2·31	2·25	2·20	2·13	2·06	1·97	1·93	1·88	1·84	1·79	1·73	1·67
28	4·20	3·34	2·95	2·71	2·56	2·45	2·36	2·29	2·24	2·19	2·12	2·04	1·96	1·91	1·87	1·82	1·77	1·71	1·65
29	4·18	3·33	2·93	2·70	2·55	2·43	2·35	2·28	2·22	2·18	2·10	2·03	1·94	1·90	1·85	1·81	1·75	1·70	1·64
30	4·17	3·32	2·92	2·69	2·53	2·42	2·33	2·27	2·21	2·16	2·09	2·01	1·93	1·89	1·84	1·79	1·74	1·68	1·62
40	4·08	3·23	2·84	2·61	2·45	2·34	2·25	2·18	2·12	2·08	2·00	1·92	1·84	1·79	1·74	1·69	1·64	1·58	1·51
60	4·00	3·15	2·76	2·53	2·37	2·25	2·17	2·10	2·04	1·99	1·92	1·84	1·75	1·70	1·65	1·59	1·53	1·47	1·39
120	3·92	3·07	2·68	2·45	2·29	2·17	2·09	2·02	1·96	1·91	1·83	1·75	1·66	1·61	1·55	1·50	1·43	1·35	1·25
∞	3·84	3·00	2·60	2·37	2·21	2·10	2·01	1·94	1·88	1·83	1·75	1·67	1·57	1·52	1·46	1·39	1·32	1·22	1·00

$F = \frac{s_1^2}{s_2^2} = \frac{S_1}{\nu_1} \Big/ \frac{S_2}{\nu_2}$, where $s_1^2 = S_1/\nu_1$ and $s_2^2 = S_2/\nu_2$ are independent mean squares estimating a common variance σ^2 and based on ν_1 and ν_2 degrees of freedom, respectively.

TABLE 4b. DISTRIBUTION OF F

Upper 1 % points

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999·5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98·50	99·00	99·17	99·25	99·30	99·33	99·36	99·37	99·39	99·40	99·42	99·43	99·45	99·46	99·47	99·47	99·48	99·49	99·50
3	34·12	30·82	29·46	28·71	28·24	27·91	27·67	27·49	27·35	27·23	27·05	26·87	26·69	26·60	26·50	26·41	26·32	26·22	26·13
4	21·20	18·00	16·69	15·98	15·52	15·21	14·98	14·80	14·66	14·55	14·37	14·20	14·02	13·93	13·84	13·75	13·65	13·56	13·46
5	16·26	13·27	12·06	11·39	10·97	10·67	10·46	10·29	10·16	10·05	9·89	9·72	9·55	9·47	9·38	9·29	9·20	9·11	9·02
6	13·75	10·92	9·78	9·15	8·75	8·47	8·26	8·10	7·98	7·87	7·72	7·56	7·40	7·31	7·23	7·14	7·06	6·97	6·88
7	12·25	9·55	8·45	7·85	7·46	7·19	6·99	6·84	6·72	6·62	6·47	6·31	6·16	6·07	5·99	5·91	5·82	5·74	5·65
8	11·26	8·65	7·59	7·01	6·63	6·37	6·18	6·03	5·91	5·81	5·67	5·52	5·36	5·28	5·20	5·12	5·03	4·95	4·86
9	10·56	8·02	6·99	6·42	6·06	5·80	5·61	5·47	5·35	5·26	5·11	4·96	4·81	4·73	4·65	4·57	4·48	4·40	4·31
10	10·04	7·56	6·55	5·99	5·64	5·39	5·20	5·06	4·94	4·85	4·71	4·56	4·41	4·33	4·25	4·17	4·08	4·00	3·91
11	9·65	7·21	6·22	5·67	5·32	5·07	4·89	4·74	4·63	4·54	4·40	4·25	4·10	4·02	3·94	3·86	3·78	3·69	3·60
12	9·33	6·93	5·95	5·41	5·06	4·82	4·64	4·50	4·39	4·30	4·16	4·01	3·86	3·78	3·70	3·62	3·54	3·45	3·36
13	9·07	6·70	5·74	5·21	4·86	4·62	4·44	4·30	4·19	4·10	3·96	3·82	3·66	3·59	3·51	3·43	3·34	3·25	3·17
14	8·86	6·51	5·56	5·04	4·69	4·46	4·28	4·14	4·03	3·94	3·80	3·66	3·51	3·43	3·35	3·27	3·18	3·09	3·00
15	8·68	6·36	5·42	4·89	4·56	4·32	4·14	4·00	3·89	3·80	3·67	3·52	3·37	3·29	3·21	3·13	3·05	2·96	2·87
16	8·53	6·23	5·29	4·77	4·44	4·20	4·03	3·89	3·78	3·69	3·55	3·41	3·26	3·18	3·10	3·02	2·93	2·84	2·75
17	8·40	6·11	5·18	4·67	4·34	4·10	3·93	3·79	3·68	3·59	3·46	3·31	3·16	3·08	3·00	2·92	2·83	2·75	2·65
18	8·29	6·01	5·09	4·58	4·25	4·01	3·84	3·71	3·60	3·51	3·37	3·23	3·08	3·00	2·92	2·84	2·75	2·66	2·57
19	8·18	5·93	5·01	4·50	4·17	3·94	3·77	3·63	3·52	3·43	3·30	3·15	3·00	2·92	2·84	2·76	2·67	2·58	2·49
20	8·10	5·85	4·94	4·43	4·10	3·87	3·70	3·56	3·46	3·37	3·23	3·09	2·94	2·86	2·78	2·69	2·61	2·52	2·42
21	8·02	5·78	4·87	4·37	4·04	3·81	3·64	3·51	3·40	3·31	3·17	3·03	2·88	2·80	2·72	2·64	2·55	2·46	2·36
22	7·95	5·72	4·82	4·31	3·99	3·76	3·59	3·45	3·35	3·26	3·12	2·98	2·83	2·75	2·67	2·58	2·50	2·40	2·31
23	7·88	5·66	4·76	4·26	3·94	3·71	3·54	3·41	3·30	3·21	3·07	2·93	2·78	2·70	2·62	2·54	2·45	2·35	2·26
24	7·82	5·61	4·72	4·22	3·90	3·67	3·50	3·36	3·26	3·17	3·03	2·89	2·74	2·66	2·58	2·49	2·40	2·31	2·21
25	7·77	5·57	4·68	4·18	3·85	3·63	3·46	3·32	3·22	3·13	2·99	2·85	2·70	2·62	2·54	2·45	2·36	2·27	2·17
26	7·72	5·53	4·64	4·14	3·82	3·59	3·42	3·29	3·18	3·09	2·96	2·81	2·66	2·58	2·50	2·42	2·33	2·23	2·13
27	7·68	5·49	4·60	4·11	3·78	3·56	3·39	3·26	3·15	3·06	2·93	2·78	2·63	2·55	2·47	2·38	2·29	2·20	2·10
28	7·64	5·45	4·57	4·07	3·75	3·53	3·36	3·23	3·12	3·03	2·90	2·75	2·60	2·52	2·44	2·35	2·26	2·17	2·06
29	7·60	5·42	4·54	4·04	3·73	3·50	3·33	3·20	3·09	3·00	2·87	2·73	2·57	2·49	2·41	2·33	2·23	2·14	2·03
30	7·56	5·39	4·51	4·02	3·70	3·47	3·30	3·17	3·07	2·98	2·84	2·70	2·55	2·47	2·39	2·30	2·21	2·11	2·01
40	7·31	5·18	4·31	3·83	3·51	3·29	3·12	2·99	2·89	2·80	2·66	2·52	2·37	2·29	2·20	2·11	2·02	1·92	1·80
60	7·08	4·98	4·13	3·65	3·34	3·12	2·95	2·82	2·72	2·63	2·50	2·35	2·20	2·12	2·03	1·94	1·84	1·73	1·60
120	6·85	4·79	3·95	3·48	3·17	2·96	2·79	2·66	2·56	2·47	2·34	2·19	2·03	1·95	1·86	1·76	1·66	1·53	1·38
∞	6·63	4·61	3·78	3·32	3·02	2·80	2·64	2·51	2·41	2·32	2·18	2·04	1·88	1·79	1·70	1·59	1·47	1·32	1·00

$F=\frac{s_1^2}{s_2^2}=\frac{S_1}{\nu_1}\Big/\frac{S_2}{\nu_2}$, where $s_1^2=S_1/\nu_1$ and $s_2^2=S_2/\nu_2$ are independent mean squares estimating a common variance σ^2 and based on ν_1 and ν_2 degrees of freedom, respectively.

LECTURE V: DISTRIBUTION-FREE TESTS

W. E. Lever

1. INTRODUCTION

The purpose of this lecture will be to acquaint you with some of the possible uses of distribution-free tests. Roughly speaking, a distribution-free test is one which does not make any assumptions about the precise form of the populations which are being sampled. However, in many cases the underlying distribution functions must be assumed to be continuous and/or to have the same form or to be symmetrical about the same point.

The reason that we do not need to make extensive assumptions about the distribution functions is that the magnitudes of the observations are not used directly in the test. Instead, we use the ranks, frequency, or some other such attribute of the original observations to provide the information used by the test statistic.

Given these types of measurements, distribution-free tests are usually concerned with medians, ranges, etc., instead of the parameters of the distribution.

However, if the experimenter has definite a priori knowledge of the parametric form of the distribution from which his observations arise, he should definitely use the classical test procedures and not the distribution-free test procedures.

2. ADVANTAGES AND DISADVANTAGES OF DISTRIBUTION-FREE TESTS

Now distribution-free tests have several advantages and disadvantages in relation to classical testing procedures. Some of the advantages are as follows: (1) Distribution-free tests are usually much easier to derive, since they can be derived by use of combinatorial formulas instead of the complex mathematics usually needed to derive the classical tests. (2) The computations needed for distribution-free tests are usually much easier than for the classical tests, and they are usually much faster to apply for small and moderate samples. (3) Since distribution-free tests are based on fewer and less elaborate assumptions than the classical tests, they can be legitimately applied to a much larger class of populations. (4) If the assumption of continuous distributions is violated, both the fact and the degree of the violation become readily apparent from the existence of extensive tied observations in the observed data. Checks of this sort are not available to the experimenter to advise him that a parametric assumption has been violated. (5) Generally, distribution-free tests have a much wider range of possible measurements which can be used.

Now that we have considered some of the good points of distribution-free tests, let us consider some of their disadvantages. (1) When a large number of similar tests are to be conducted using a computer rather than a desk calculator, classical tests will usually be faster to do at all sample sizes. (2) Generally, distribution-free tests have decreasing statistical efficiency relative to classical tests as the sample size becomes large. (3) Problems can arise with the choice of

a rejection region for the test, since the point probability of a distribution-free test statistic does not necessarily increase as the test statistic approaches its most probable value. (4) Also, since the distribution of the test statistic is usually discrete, a value of the test statistic for a preassigned level of significance α may not exist.

The advantages and disadvantages we have just considered are very important factors to be considered when choosing between a distribution-free test and a classical test, but possibly the most important factor is the sample size. This is true for the following reasons: For small samples (e.g., $N \leq 10$), distribution-free tests are easier to apply, as indicated before, and only slightly less efficient even if all the assumptions of the classical test are met. But at these sample sizes, violations of the assumptions for the classical tests generally have their most serious effects and are most likely to go undetected. Thus, unless one has some a priori knowledge that all the assumptions are valid for the classical test, the wise choice would usually appear to be a distribution-free test. For larger samples (e.g., $N > 30$), some distribution-free tests still compare favorably with classical tests, but many are time consuming. Also, in contrast to classical tests whose assumptions are met, their calculated or tabled probabilities may only be approximate. In addition, the violations of the assumptions for the classical tests will have become apparent, and in many cases their effect may have been negligible due to the properties of the central limit theorem. Thus for large samples the wise choice would be the classical test.

In the remainder of this lecture I will try to point out some of the common types of distribution-free tests, so that you will have an indication where they can be applied. In each procedure considered, I will consider the rationale behind the procedure, the null hypothesis (i.e., the hypothesis to be tested), the necessary assumptions, and the treatment of tied observations.

3. SIGN TEST

The sign test is a very simple and easily applied distribution-free test for the equality of the medians of two distributions. It requires that the observations from the two distributions occur in pairs and that we be able to at least observe the sign of the difference of these two observations.

The rationale for the sign test is as follows: Suppose we have two random variables A and B and that n pairs of observations are taken on A and B. Then if the medians of the distributions of A and B are the same, the observed values of the differences $A_i - B_i$ can be either plus or minus with probability $\frac{1}{2}$. Let r be the number of plus or minus signs, whichever is fewer. Then the probability of obtaining r or fewer of the appropriate signs if the medians are the same is

$$q_r = \sum_{i=0}^{r} \binom{n}{i} \left(\frac{1}{2}\right)^n .$$

Thus, if $q_r \leq \alpha/2$, where α is the level of significance, we would reject the hypothesis that the medians are equal.

Formally the null hypothesis is that for every difference $A_i - B_i$,

$$P(A_i > B_i) = P(A_i < B_i) = \tfrac{1}{2} ,$$

that is, the distributions of A_i and B_i have equal medians.

For the null hypothesis to hold, we must assume that $P(A_i = B_i) = 0$, that is, there are no zero differences. Also, the differences $A_i - B_i$ must be assumed to be independent.

However, even if the assumption that $P(A_i = B_i) = 0$ holds, the lack of precise measurements may lead to some zero differences. If this does happen, the most accepted procedure for handling this situation is to assign half of the zero differences plus signs and half minus signs, dropping one zero difference if there is an odd number of them.

For an example of the application of the sign test we will use problem 7.3, p. 201, of Brownlee (1960). The example is as follows:

In a trial of two types of rain gages, 69 of type A and 12 of type B were distributed at random over a certain area. In a certain period 14 storms occurred, and the average amounts of rain found on the two types of gages were as shown in the accompanying table. Now suppose we wish

Storm	Type A	Type B	Sign of $A - B$
1	1.38	1.42	−
2	9.69	10.37	−
3	0.39	0.39	0
4	1.42	1.46	−
5	0.54	0.55	−
6	5.94	6.15	−
7	0.59	0.61	−
8	2.63	2.69	−
9	2.44	2.68	−
10	0.56	0.53	+
11	0.69	0.72	−
12	0.71	0.72	−
13	0.95	0.93	+
14	0.50	0.53	−

to test at the 5% level of significance, by use of the sign test, that the two gages are giving the same results.

To apply the sign test we must be able to assume that the probability of a zero difference is zero and that the differences are independent. Both of these assumptions are reasonable, since the gages would be measuring continuous variables and the measurements are based on different storms. Note that in this case we do not make the assumption that the distribution of possible measurements is symmetrical.

Now there are 2 plus signs, 11 minus signs, and 1 zero difference. Then the probability that we would only find 2 plus signs out of 13 differences (we disregard the one zero difference) is given by

$$q_2 = \sum_{i=0}^{2} \binom{13}{i}\left(\frac{1}{2}\right)^{13} = 0.011 < 0.05/2 .$$

Thus in this case we would reject the hypothesis that the two rain gages give the same results.

Sign tests are also available for testing trend in location, trend in dispersion, and cyclical trend.

4. SIGNED RANK TEST

The next procedure we will consider is again a test based on paired observations, but in this case we will consider the rank of the differences in addition to the sign of the difference. This is done by ranking the absolute values of the differences and then assigning the sign of the difference to the rank.

The rationale behind the signed rank test is as follows: Consider two random variables A and B which are sampled in pairs. Either if the distributions of A and B are the same or if they are just symmetrical about the same point, then there are 2^n possible ways the signs could be assigned to the observed ranks, where n is the number of differences. Thus, if the distributions are the same or are symmetrical about the same point, we would expect the sum of the ranks with plus signs to be about the same as the sum of the ranks with negative signs, in absolute value.

For the test statistic let T be the absolute value of the sum of the ranks with plus or minus signs, whichever is smaller. We will reject the hypothesis that the two populations are the same or that they are symmetrical about the same point if the probability of a value of T less than or equal to the observed T is less than $\alpha/2$, where α is the level of significance. Tables of critical values of T are given by Wilcoxon, Katti, and Wilcox (1963) for $5 \leq n \leq 50$.

Formally, the null hypothesis is that the medians and the means of the two distributions are equal, if the assumptions are true.

The necessary assumptions for this test are as follows: The two distributions have identical form, differing at most in location, or the two distributions are both symmetrical. Also we require that the n differences be independent and that the probability of a zero difference be zero. This latter assumption can be accomplished by assuming that the two distributions are both continuous.

If tied differences occur through a lack of sufficient measurement precision, then the procedure is as follows: If there is an even number m of zero differences, consider them as occupying the m lowest ranks. Then give each the midrank of this group $(m + 1)/2$, and assign half of them a plus sign and half a minus sign. If there is an odd number of zero differences, disregard one of them, and reduce the sample size to $n - 1$.

If nonzero differences are tied in absolute magnitude, the number of each group should be given the midrank of the group and the sign of its corresponding difference.

As an example of the signed rank test, let us again use the example used for the sign test with the additional assumption that both distributions are symmetrical. Then the differences and their ranks are as follows:

Storm	$A - B$	Rank of $A - B$
1	−0.04	−9.5
2	−0.68	−13
3	0	
4	−0.04	−9.5
5	−0.01	−1.5
6	−0.21	−11
7	−0.02	−4
8	−0.02	−4
9	−0.24	−12
10	+0.03	+7
11	−0.03	−7
12	−0.01	−1.5
13	+0.02	+4
14	−0.03	−7

Then the test statistic T is:

$$T = +4 \quad +7 = 11 .$$

Then by checking the Wilcoxon, Katti, and Wilcox tables, it can be seen that the critical value of T is 17; thus

$$P(T \leqq 11) < 0.05 .$$

So again we reject the hypothesis that the gages give equal results.

5. THE RANK SUM TEST

The rank procedure to be considered in this section is again a method of comparing aspects of the distributions of two random variables A and B. But unlike the two previous situations considered, the data are not assumed to arise in pairs. In this procedure we rank the n observed values of A and the m observed values of B (n and m are not necessarily equal) as if they are one population. Then we base our test procedure on the sum of the ranks of the observed values of A or the observed values of B.

The rationale behind this procedure is as follows: If two random samples are drawn from the same population, then they may be regarded as a single random sample of size $m + n$ which has been divided in some random manner into subsamples of sizes n and m. Thus any assignments of n ranks to one population A and m ranks to population B are equally likely if the two samples came from the same population. For a test statistic let T be the sum of ranks from the smaller

sample or, if the samples are of equal size, the smaller sum of ranks for the two populations. Then if the probability that a value of T can occur less than or equal to the observed value of T is less than $\alpha/2$, given that all assignments of n ranks to population A and m ranks to population B are equally likely, we will reject the hypothesis that A and B have the same distribution. Critical values of T can be found in Wilcoxon, Katti, and Wilcox (1963).

Formally, the null hypothesis is that each of the

$$\binom{m+n}{n}$$

pairs of rank assignments are equally likely. This implies the null hypothesis that the distributions of the random variables A and B are identical.

The necessary assumptions for this procedure are: The $m + n$ observations are random and independent. Also, the two distributions must have the same form, and the two distributions are continuous (the probability of ties occurring is zero).

If ties do occur as the result of imprecise measurements, they are only a problem when the group of tied values lies in both samples. When a group of tied observations occurs totally within one sample, they should be arbitrarily ranked as if they were not tied. If the tied group occurs in both samples, the most frequently recommended procedure is to assign each member of the tied group the midrank of the group.

For an example of the application of the rank sum test, let us use problem 7.4, p. 202, of Brownlee (1960). The example is as follows:

A group of mice are allocated to individual cages randomly. The cages are allocated, in equal numbers, randomly, to two treatments, a control A and a certain drug B. All animals are infected, in a random sequence, with tuberculosis. Given that the drug is not toxic, we wish to test the hypothesis that the drug had no effect at the 5% level of significance. The days the mice died and their ranks are as follows (one mouse got lost):

Control, A		Drug, B	
Day of Death	Rank	Day of Death	Rank
5	1	7	4
6	2	8	8.5
7	4	8	8.5
7	4	8	8.5
8	8.5	9	13
8	8.5	9	13
8	8.5	12	15.5
9	13	13	17
12	15.5	14	18
		17	19

Then the statistic T is the sum of the ranks of A, so that

$$T = (1 + 2 + 4 + 4 + 8.5 + 8.5 + 8.5 + 13 + 15.5)$$
$$= 65 .$$

Then by checking the Wilcoxon, Katti, and Wilcox tables, we find that the critical value of T for $m = 9$ and $n = 10$ is 69. Thus $P(T \leq 65) < 0.05$, and we reject the hypothesis that the drug had no effect.

6. A K-SAMPLE PROCEDURE FOR UNMATCHED DATA – ONE-WAY CLASSIFICATION

Up to this point each of the procedures considered has dealt with comparisons of the distributions of two random variables. So now let us consider some applications of distribution-free procedures to the comparison of K distributions. In this section we will consider unmatched data – what is generally called a one-way classification. In the next section we will consider a two-way classification.

The one-way classification procedure which we will consider is an extension of the rank sum test for unmatched data considered in the previous section. Thus the rationale behind it will be similar to the rationale behind the rank sum test. That is, suppose N random and independent observations were taken on C random variables and they were ranked from 1 to N regardless of what population they came from. Then, if the C random variables arise from the same population, the expected sum of ranks for the observations on each of the C random variables would be $n_i(N + 1)/2$, where n_i is the number of observations on the ith random variable and $(N + 1)/2$ is the average rank for all N observations.

Now let R_i be the sum of the observed ranks of the observations on the ith random variable. If the C random variables came from the same population, then the differences

$$R_i - n_i(N + 1)/2$$

should be small. Then, using this idea, two of the statistics which have been proposed to test the hypothesis that the C random variables have arisen from the same population are as follows:

$$S = \sum_{i=1}^{C} [R_i - n_i(N + 1)/2]^2 , \tag{1}$$

and

$$H = \frac{12}{N(N + 1)} \sum_{i=1}^{C} \frac{[R_i - n_i(N + 1)/2]^2}{n_i} . \tag{2}$$

Exact tables for S have been computed by Kruskal and Wallis (1952) and Rijkoort (1952) for $C = 3$, 4, or 5 and for equal sample sizes on each random variable of $n_i = 2$, 3, 4, or 5 for $C = 3$,

$n_i = 2$ or 3 for $C = 4$, and $n_i = 2$ for $C = 5$. Also, exact tables for H have been computed by Kruskal and Wallis (1952 and 1953) for $C = 3$ and $n_i \leqq 5$, the n_i not necessarily equal. Also, for large values of n_i, $i = 1, \ldots, C$, H is distributed approximately as a chi-square random variable with $C - 1$ degrees of freedom.

Formally, the null hypothesis is that the C random variables all come from the same population.

The necessary assumptions for this procedure are that the observations are drawn randomly and independently from continuously distributed populations. The continuous distribution requirement is one way of forcing the probability of a tie to be zero. Also, if the approximate chi-square test is to be used, all the n_i must be large enough for the central limit theorem to apply.

If ties occur, assign each group of tied values the midrank of the tied group. Then if there are t_i tied values in a group, adjust H in the following manner:

$$H' = \frac{H}{1 - \Sigma(t_i^3 - t_i)/(N^3 - N)} .$$

For an example of the application of the K-sample procedure for unmatched data, we will use the example on p. 284 of Bradley (1960). The example is as follows:

Suppose that speed of reading is to be tested under three degrees of illumination, at 5% level of significance. Nine subjects are selected at random from a common population, and three subjects are randomly assigned to each condition of illumination (A, B, C). For some reason, one subject fails to complete the experiment. The results and their ranks are given in the following tables:

Result			Rank		
A	B	C	A	B	C
22	36	39	1	4	6
31	37	44	2	5	7
35		51	3		8
			Sum 6	9	21

Since the subjects have been randomly selected and the response is continuous, the assumptions seem to be satisfied.

Then after ranking the eight observations as one population, we find the rank sums for the three conditions to be 6, 9, and 21 respectively. Also, we find their respective expected rank sums, under the hypothesis of one distribution, to be $3(8 + 1)/2$, $2(8 + 1)/2$, and $3(8 + 1)/2$.

Thus the test statistic H is given by

$$H = \frac{12}{8 \times 9} \left[\frac{(6 - 27/2)^2}{3} + \frac{(9 - 9)^2}{2} + \frac{(21 - 27/2)^2}{3} \right]$$

$$= 6.25 .$$

Then by consulting the Kruskal and Wallis tables, we find that given the null hypothesis:

$$P(H \geqq 6.25) = 0.011 < 0.05 .$$

Thus we would reject the null hypothesis that the speed of reading is the same under the three degrees of illumination.

7. A K-SAMPLE PROCEDURE FOR MATCHED DATA – TWO-WAY CLASSIFICATION

In this next case, instead of each treatment being applied to different subjects, each treatment will be applied to every subject. Thus every observation will depend on two classifications. This method differs from the test considered in the previous section, since that test considered only one classification, the treatment.

In this procedure, the application of a treatment to the subject must be done in such a manner that the response for another treatment applied to the same subject will not be affected. This requirement is needed since we will be interested in seeing whether the responses for the various treatments will be the same over each subject. The rationale behind this procedure is as follows: Suppose that we have m subjects and n treatments and that each of the n treatments is applied to every subject. Now suppose we rank the n treatments over each subject separately; then we will have m sets of ranks from 1 to n.

Then if a subject responds in an equal manner to each treatment, although this response may be different from subject to subject, we would expect the sum of the ranks for each treatment to be $m(n + 1)/2$. The value $(n + 1)/2$ is the average rank for each treatment, averaged within a subject.

Now let R_i be the sum of the ranks for a treatment; then one statistic which has been proposed for testing the effect of the treatment on the subjects is as follows:

$$S = \sum_{i=1}^{n} \left(R_i - m\frac{n+1}{2}\right)^2 .$$

If the hypothesis that the treatments have an equal effect on a subject is true, then we would expect the value of S to be small, so that we would reject the hypothesis for large values of S. For small values of m and n, critical values of S have been tabled by Kendall (1955), and for larger values of m and n, $12S/mn(n - 1)$ has been shown to have approximately a χ^2 distribution with $m - 1$ degrees of freedom.

Formally, the null hypothesis is as follows: For each subject, the random variables which represent the responses of the subject to the treatments have the same distribution. But this does *not* imply that the responses for the treatments on different subjects have the same distribution.

The necessary assumptions for this procedure are as follows: The subjects are independent and the observations on each subject are randomly selected; that is, the treatments are applied in

a random order. Also, the response distribution for each subject is assumed to be continuous. This implies that the probability of tied observations on a subject is zero.

If, through a lack of measurement precision, ties do occur, the following procedure is recommended. For each subject assign the tied group the midrank of that group. A correction for the statistic S when ties occur is as follows: Let t_{ij} be the number of tied observations in the ith group of ties within the jth subject. Then use the corrected statistic

$$S \div \left[\frac{m \cdot n(n+1)}{12} - \frac{\sum_{j=1}^{m} \sum_{i} (t_{ij}^3 - t_{ij})}{12(n-1)} \right] .$$

For an example of the application of the two-way classification procedure we will use the example on p. 292 of Bradley (1960). The example is as follows:

Each of three subjects performs a well-learned task three times, each time under the influence of a different drug. Performance is timed, and the experimenter wishes to test the hypothesis that no subject's performance times were influenced more by one drug than by another, at a 5% level of significance. Tables of the times and of their ranks are as follows:

	Time				Time Rank		
Subject	Drug I	Drug II	Drug III		Drug I	Drug II	Drug III
A	4.76	1.30	7.91		2	1	3
B	14.51	10.27	35.84		2	1	3
C	82.11	82.09	82.14		2	1	3
				Sum	6	3	9

The assumption of continuous response distributions seems valid in this case, since we are measuring time. So if the observations have been selected in the proper manner, the necessary assumptions seem to be valid.

Now after we have ranked the times of each subject separately, we find that the sums of the ranks for the three drugs are 6, 3, and 9. Also we find that the expected rank sum for each drug, under the hypothesis of equal effects, would be

$$m\left(\frac{n+1}{2}\right) = 3\left(\frac{3+1}{2}\right) = 6 .$$

Thus the test statistic S is given by

$$S = (6-6)^2 + (3-6)^2 + (9-6)^2 = 18 .$$

Then by consulting Kendall's exact tables, it can be found that, given the null hypothesis,

$$P(S \geqq 18) = 0.028 < 0.05 .$$

Thus we would reject the hypothesis that the drugs produce equal effects.

8. CONCLUSION

The group of distribution-free test procedures which we have just considered is a small but useful subset of the set of distribution-free procedures.

We also have procedures available for various kinds of trends, runs of events, confidence controls, and several other situations. To read about many distribution-free procedures I suggest you read Wright Air Development Division (WADD) Technical Report 60-661, *Distribution-Free Statistical Tests*, by James V. Bradley, available at the Y-12 Technical Library. This report also contains extensive bibliographies at the end of each chapter.

REFERENCES

Bradley, J. V., "Distribution-Free Statistical Tests," *WADD Technical Report 60-661*, Wright Air Development Division, Air Research and Development Command, United States Air Force, Wright-Patterson Air Force Base, Ohio, 1960.

Brownlee, K. A., *Statistical Theory and Methodology in Science and Engineering*, Wiley, New York, 1960.

Kendall, M. G., *Rank Correlation Methods*, 2d ed., Hafner, New York, 1955.

Kruskal, W. H., and W. A. Wallis, "Use of Rank in One-Criterion Variance Analysis," *J. Am. Statist. Assoc.* **47**, 583–621 (1952); **48**, 907–11 (1953).

Rijkoort, P. J., "A Generalization of Wilcoxon's Test," *Proc. Koninkl. Ned. Akad. Wetenschap.*, Ser. A **55**, 394–404 (1952).

Savage, I. R., *Bibliography of Nonparametric Statistics*, Harvard University Press, Cambridge, Mass., 1962.

Siegel, S., *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York, 1956.

Tate, M. W., and R. C. Clelland, *Nonparametric and Shortcut Statistics*, Publishers, Inc., Danville, Ill., 1957.

Wilcoxon, F., S. K. Katti, and R. A. Wilcox, *Critical Values and Probability Levels for the Wilcoxon Rank Sum Test and the Wilcoxon Signed Rank Test*, American Cyanamid Co. and The Florida State University Pamphlet, Pearl River, N.Y., and Tallahassee, Fla., 1963.

Wilcoxon, F., and R. A. Wilcox, *Some Rapid Approximate Statistical Procedures*, American Cyanamid Co. Pamphlet, Pearl River, N.Y., 1964.

LECTURE VI: LINEAR REGRESSION

T. L. Hebble

1. INTRODUCTION

The concept of linear regression is concerned with the relationship between a measured response Y and a function of known constants $(x_1, \ldots, x_p)$ and unknown parameters $(\beta_1, \ldots, \beta_p)$. This relationship is often written in the form

$$Y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e ,$$

where e is the error in measuring Y. The word "linear" means that Y is a linear function of the β's and not necessarily of the x's. If we know $\beta_1, \ldots, \beta_p$ we could predict the response exactly by the above functional (true) relationship. However, in general, we do not know the β's, and it is desirable to find some function of the observations Y which provides a "good" estimate of $\beta_1, \ldots, \beta_p$.

Examples of linear regression occur frequently in chemical kinetics, thermodynamics, and virtually all areas of physics. The fixed constants are usually specific values of process variables such as temperature, flow rate, and catalyst concentration. Although these variables are not exactly known, the errors associated with measuring them are negligible relative to the errors in measuring the response.

We shall confine ourselves to a discussion of parameter estimation, confidence interval estimation, and some tests of hypotheses associated with linear regression. Since a heavy reliance is placed on the use of matrices, a short review of the fundamentals of matrices is given in the Appendix (Sect. 7 of this lecture).

1.1 Notation

Every effort is made to conform to the notation of previous lectures, but in certain instances tradition dictates that other notation be employed. In addition to capital letters (other than those in boldface) I shall use the letter e to designate random variables. Constants and the realization of random variables are denoted by lower-case letters. Also, parameter estimates which are indicated by a $\wedge$ ("hat") or $\sim$ (tilde) (such as $\hat{\beta}$ and $\tilde{\sigma}^2$) are either random variables or numerical values of random variables.

2. LINEAR REGRESSION

2.1 Definition of Problem

Consider a random variable Y which is functionally related to known constants $x_1, \ldots, x_p$ and unknown parameters $\beta_1, \ldots, \beta_p$ by

$$Y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e ,$$

where e is a random variable representing measurement error. Let the probability density function (p.d.f.) of Y be given by

$$f(y;\ x_1, \ldots, x_p, \beta_1, \ldots, \beta_p) .$$

Assuming that the expected value of e, $E(e)$, is equal to zero, we may write

$$E(Y) = \beta_1 x_1 + \ldots + \beta_p x_p .$$

The x's are process variables or independent variables, and Y is the response or dependent variable. The β's are called the regression coefficients, parameters, or effects.

As an example, we might consider that for some process, efficiency Y is related to temperature according to

$$Y = \beta_1 x_1 + \beta_2 x_2 + e ,$$

where x_1 = temperature, $x_2 = x_1^2$, and e is measurement error. In this model, the unknown parameters β_1 and β_2 are the linear and quadratic effects of temperature on efficiency.

Now expand this idea of a single experiment with two parameters in the model to n experiments with p parameters in the model. That is, we have n equations and p unknowns in which each equation represents an experiment and the Y's and e's represent random variables:

$$\left.\begin{aligned} Y_1 &= \beta_1 x_{11} + \ldots + \beta_p x_{1p} + e_1 \\ Y_2 &= \beta_1 x_{21} + \ldots + \beta_p x_{2p} + e_2 \\ &\ldots\ldots \\ Y_n &= \beta_1 x_{n1} + \ldots + \beta_p x_{np} + e_n \end{aligned}\right\} . \tag{2.1.1}$$

The expected value of the response for the ith experiment is

$$E(Y_i) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \qquad (i = 1, \ldots, n) .$$

It is convenient and certainly less cumbersome to convert to matrix notation. Thus Eqs. (2.1.1) become

$$\mathbf{Y}^{(n\times 1)} = \mathbf{X}^{(n\times p)}\, \boldsymbol{\beta}^{(p\times 1)} + \mathbf{e}^{(n\times 1)} ,$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} .$$

Matrices and vectors are denoted by boldface letters. The superscripts denote order and will be omitted when the context is clear.

The $\mathbf{X}$ matrix may be preselected as prescribed by an experimental design, or the $\mathbf{X}$ matrix may be just a collection of similar experiments. It is important that the values of x are known constants, or at least the errors associated with measuring them are negligible.

3. ESTIMATION OF PARAMETERS

Two general methods of parameter estimation as applied to linear regression are presented. First, the method of least squares is discussed under the assumptions:

$$E(\mathbf{e}) = \mathbf{0}\ , \qquad E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{I}\ ,$$

where $\mathbf{0}$ is an $n \times 1$ vector of zeros and $\mathbf{I}$ is the identity matrix. The prime (′) indicates the transpose. Second, the method of maximum likelihood is discussed under the assumption that $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, that is, that $\mathbf{e}$ has a multivariate normal distribution.

In the second case, each e_i $(i = 1, \ldots, n)$ has a normal distribution with mean 0 and variance σ^2. Also, the variance-covariance matrix $\sigma^2\mathbf{I}$ implies that the e_i are mutually independent (in a stochastic sense), since the covariance elements are 0.

3.1 Method of Least Squares

A technique of parameter estimation which has a strong mathematical and intuitive foundation is the method of least squares. For a given function $f(\mathbf{x}, \beta)$ and a response y, the set of parameters β is found so that the residual sum of squares is a minimum:

$$\text{residual sum of squares} = \sum_{i=1}^{n} [y_i - f_i(\mathbf{x}, \beta)]^2\ . \tag{3.1.1}$$

Here $\mathbf{x}$ is the vector of levels of the independent variables which corresponds to the ith experiment. Equation (3.1.1) represents the sum of squared deviations between the observations and the function.

In terms of our problem, we may write for the residual sum of squares

$$\sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)\ . \tag{3.1.2}$$

Note that in Eq. (3.1.2) we must restrict ourselves to linear functions of β, whereas in Eq. (3.1.1) $f(\mathbf{x}, \beta)$ may be virtually any function. If $\mathbf{e}'\mathbf{e}$ is divided by $n - p$, the degrees of freedom for error, we obtain the residual mean square:

$$\text{residual mean square} = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-p}\,\mathbf{e}'\mathbf{e}\ .$$

Since we have values for y (the outcome of the experiment) and x (the known levels of the independent variables), we need only to find β so that $\mathbf{e}'\mathbf{e}$ is a minimum. To do this we differentiate $\mathbf{e}'\mathbf{e}$ with respect to each β, set the resulting equations equal to zero, and solve for β. Thus:

$$\mathbf{e}'\mathbf{e} = \sum_{i=1}^{n} (Y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2 ,$$

and the partials with respect to the β's are

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \beta_1} = 2 \sum_{i=1}^{n} (Y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})(-x_{i1})$$

$$\cdots\cdots$$

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \beta_p} = 2 \sum_{i=1}^{n} (Y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})(-x_{ip}) .$$

In matrix notation,

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \beta} = -2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta .$$

The value of β for which $\partial \mathbf{e}'\mathbf{e}/\partial \beta = 0$ is denoted $\hat{\beta}$ and is called the least-squares estimate of β. The equations $(\partial \mathbf{e}'\mathbf{e}/\partial \beta) = 0 = \mathbf{X}'\mathbf{X}\hat{\beta} - \mathbf{X}'\mathbf{Y}$ are called the normal equations. Premultiplying both sides of the normal equations by $(\mathbf{X}'\mathbf{X})^{-1}$, we have[1]

$$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad \text{if } |\mathbf{X}'\mathbf{X}| \neq 0 ,$$

and hence

$$\mathbf{I}\hat{\beta} = \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} .$$

The method of least squares does not provide an estimate of σ^2. An estimate of σ^2, denoted $\hat{\sigma}^2$, can be found by substituting $\hat{\beta}$ for β in the equation for the residual mean square:

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-p} \mathbf{e}'\mathbf{e} = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) .$$

It can be shown that this estimate is unbiased, that is, that $E(\hat{\sigma}^2) = \sigma^2$.

[1]The matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is not defined for $|\mathbf{X}'\mathbf{X}| = 0$. To avoid this, we state that $\mathbf{X}$ must be of rank p, which implies that the $p \times p$ $\mathbf{X}'\mathbf{X}$ matrix is also of rank p. A $p \times p$ matrix of rank p is nonsingular.

3.2 Example

Suppose we consider the simple linear model

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \qquad \text{for } i = 1, 2, \ldots, n .$$

There are n observations and $p = 2$ parameters. To generate an intercept β_1, let $x_{i1} = 1$ for all $i = 1, \ldots, n$. Setting $x_{i2} = x_i$, this model may be rewritten as

$$Y_i = \beta_1 + \beta_2 x_i + e_i \qquad \text{for } i = 1, \ldots, n . \tag{3.2.1}$$

Now

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Since

$$(\mathbf{X}'\mathbf{X}) = \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{pmatrix},$$

then

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \dfrac{\Sigma x_i^2}{n\Sigma(x_i - \bar{x})^2} & \dfrac{-\Sigma x_i}{n\Sigma(x_i - \bar{x})^2} \\ -\dfrac{\Sigma x_i}{n\Sigma(x_i - \bar{x})^2} & \dfrac{1}{\Sigma(x_i - \bar{x})^2} \end{pmatrix}.$$

Thus

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Y} = \frac{1}{n\Sigma(x_i - \bar{x})^2} \begin{pmatrix} \Sigma x_i^2 \Sigma y_i - \Sigma x_i \Sigma x_i y_i \\ -\Sigma x_i \Sigma y_i + n\Sigma y_i x_i \end{pmatrix}.$$

The elements of $\hat{\beta}$ can be reduced to

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

and

$$\hat{\beta}_2 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} .$$

The unbiased estimate of σ^2, the residual mean square, is

$$\hat{\sigma}^2 = \frac{1}{n-2}\left\{\Sigma(y_i - \bar{y})^2 - \frac{[\Sigma(x_i - \bar{x})(y_i - \bar{y})]^2}{\Sigma(x_i - \bar{x})^2}\right\} .$$

In the foregoing example the summation over i is understood to go from 0 to n.

For a numerical example, consider the problem presented in Graybill[2] in which we wish to predict the distance that a particle travels in time x. Suppose the time is measured accurately relative to distance and we measure the distance the particle travels at several points in time.

time, x	1	2	3	4	10	12	18
distance, d	9	15	19	20	45	55	78

Assuming that the particle travels at constant velocity, the model [see (3.2.1)] is

$$d_i = \beta_1 + \beta_2 x_i + e_i \qquad (i = 1, 2, \ldots, n) ,$$

where β_1 is the position of the particle at $t = 0$ and β_2 is the velocity of the particle. Thus, if $\mathbf{Y}$ is the vector of distances and $\mathbf{X}$ is the matrix of times,

$$(\mathbf{X}'\mathbf{X}) = \begin{pmatrix} 7 & 50 \\ 50 & 598 \end{pmatrix} , \qquad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.35469 & -0.029656 \\ -0.029656 & 0.0041518 \end{pmatrix} ,$$

$$\mathbf{Y}'\mathbf{Y} = 12{,}201 , \qquad \hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 5.71 \\ 4.02 \end{pmatrix} , \qquad \hat{\sigma}^2 = 2.22 .$$

3.3 Gauss-Markov Theorem

The major properties of $\hat{\beta}$ as an estimator of β are given in an important theorem called the Gauss-Markov theorem. This theorem states that, under the following assumptions:

1. $\mathbf{Y}^{(n\times 1)} = \mathbf{X}^{(n\times p)}\beta^{(p\times 1)} + \mathbf{e}^{(n\times 1)}$, where $\mathbf{X}$ is of rank $p \leq n$,
2. $E(\mathbf{e}) = \mathbf{0}$,
3. $E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{I}$,

the least-squares estimate $\hat{\beta}$ provides the best linear unbiased estimate of β. The word "best" means that the variance of $\hat{\beta}_i$ is a minimum when compared with all other *linear* estimators of β_i. (Note that here "linear" refers to a linear function of the observations.) Since the distribution of the residual $\mathbf{e}$ was not specified, $\hat{\beta}$ cannot be compared with all estimators.

[2]F. A. Graybill, *An Introduction to Linear Statistical Models*, vol. 1, McGraw-Hill, New York, 1961.

The Gauss-Markov theorem may be extended to all linear functions of β. For a given $p \times 1$ vector of constants $\mathbf{r}$, the best linear unbiased estimate of $\mathbf{r}'\beta$ is given by

$$\mathbf{r}'\hat{\beta} = \mathbf{r}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} .$$

3.4 Parameter Estimation by Method of Maximum Likelihood

To employ the method of maximum likelihood as a means of estimating parameters, we must assume that the joint probability density function (p.d.f.) is known. When the p.d.f. is considered to be a function of the parameters, it is called the likelihood equation, and for a given "random" sample, the estimates of the parameters which maximize this equation are called maximum likelihood estimates.

In terms of our regression problem, assume that

$$\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) . \tag{3.4.1}$$

As mentioned earlier, (3.4.1) implies that each e_i $(i = 1, \ldots, n)$ is normally distributed with mean 0 and variance σ^2 and that the e_i are mutually independent. This joint p.d.f. and hence the likelihood equation L may be written as

$$L = f(\mathbf{e}; \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\mathbf{e}'\mathbf{e}/2\sigma^2) . \tag{3.4.2}$$

We now wish to find those estimates of β and σ^2 which maximize (3.4.2). To accomplish this, it will be easier to work with the logarithm of the likelihood function:

$$\begin{aligned} \ln L &= \ln f(\mathbf{e}; \beta, \sigma^2) \\ &= -\frac{(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)}{2\sigma^2} - \frac{n}{2}\ln 2\pi\sigma^2 . \end{aligned} \tag{3.4.3}$$

The values which maximize (3.4.3) also maximize (3.4.2).

To obtain these estimates, the partials of Eq. (3.4.3) with respect to each parameter are found. Thus for β,

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2}[\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\beta] ,$$

and for σ^2,

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{1}{2\sigma^4}[(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) - n\sigma^2] .$$

Setting the above partials equal to zero and solving for the parameters, we have

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

and

$$\tilde{\sigma}^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) .$$

Note that the maximum likelihood estimate of β is the same as the least-squares estimate of β. As indicated earlier,

$$E(\hat{\beta}) = \beta$$

and

$$E\left(\frac{n}{n-p}\,\tilde{\sigma}^2\right) = E(\hat{\sigma}^2) = \sigma^2 .$$

That is, the unbiased estimate of σ^2 is

$$\frac{n}{n-p}\,\tilde{\sigma}^2 = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) .$$

Under the assumptions

1. $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$,
2. $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$, where $\mathbf{X}$ is of rank $p \leq n$,

the estimators $\hat{\beta}$ and $\hat{\sigma}^2$ have several important properties. Among these are:

1. $\hat{\beta} \sim N[\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$ (multivariate normal),
2. minimum variance unbiased,
3. $(n-p)\hat{\sigma}^2/\sigma^2 \sim$ chi-square distribution with $n-p$ degrees of freedom,
4. $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

They also have the properties of being consistent, complete, sufficient, and efficient.

4. TESTING THE REGRESSION MODEL

4.1 Test for Lack of Fit

Regression analysis as presented can also be employed when the relationship is not known or when we wish to approximate a more complicated expression. As mentioned earlier, $\mathbf{e}'\mathbf{e}$ represents the residual (error) sum of squares if the functional relationship is known. However, if it is not known or we wish to approximate a complicated function, $\mathbf{e}'\mathbf{e}$ contains the sum of squares that is due to the inability of the model to fit the data or simply lack of fit. That is,

$$[\text{residual sum of squares (RSS)}] = [\text{error sum of squares (ESS)}] + [\text{lack of fit sum of squares (LFSS)}] .$$

We may test this lack of fit if there are multiple observations at one or more values of x or if there is a separate independent estimate of σ^2. (By "multiple observations" is meant that the experiment is repeated and not just continued by extending a single experiment.)

In terms of matrices,

$$\mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} ,$$

where $\mathbf{Y}'\mathbf{Y}$ is the total sum of squares and $\hat{\beta}'\mathbf{X}'\mathbf{Y}$ is the sum of squares due to regression. Thus

$$\text{LFSS} = \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} - \text{ESS} .$$

The error sum of squares is computed by "pooling" (combining) the sum of squares about the mean at each value of x for which there are multiple observations. For example, if there are k observations made at a particular value x, then the sum of squares about $\bar{y}$, the mean of the response at x, is computed according to the familiar

$$\sum_{j=1}^{k} (y_j - \bar{y})^2 = \left[\sum_{j=1}^{k} y_j^2 \right] - \frac{\left(\sum_{j=1}^{k} y_j \right)^2}{k} ,$$

where

$$\bar{y} = \sum_{i=1}^{n} y_i/k .$$

The number of degrees of freedom is $k - 1$.

If there are ℓ distinct values of x and there are k_i (≥ 1) in the ith group ($i = 1, \ldots, \ell$), then the error sum of squares is obtained by summing over all values. Thus

$$\text{ESS} = \sum_{i=1}^{\ell} \sum_{j=1}^{k_i} (y_{ij} - \bar{y}_i)^2$$

and has $n - \ell$ degrees of freedom. Now, since ESS and LFSS are independent and are distributed χ^2 with $n - \ell$ and $\ell - p$ degrees of freedom, respectively, the ratio

$$\frac{\text{LFSS}/(\ell - p)}{\text{ESS}/(n - \ell)} \sim F[\ell - p, n - \ell] . \tag{4.1.1}$$

For convenience, this test may be outlined by using Table 1.

If the ratio (4.1.1) exceeds the preselected percentage point of the Snedecor F distribution, the model is not an adequate representation of the data, and a new model should be found. If the ratio is not significant, there is no reason to reject the model. On the other hand, there is no reason to assume that the model is correct or even that it is the best model.

The question often arises as to whether the residual mean square can replace the error mean square for testing various hypotheses if the lack of fit is not significant. The purpose of using the residual mean square is to increase the degrees of freedom for error and thereby reduce the value from the F table. There is no clear-cut answer, and the choice is left up to the individual.

Table 1. Analysis of Variance

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio
Total	n	$\mathbf{Y}'\mathbf{Y}$		
Regression	p	$\hat{\beta}'\mathbf{X}'\mathbf{Y}$	$\hat{\beta}'\mathbf{X}'\mathbf{Y}'/p$	
Residual	$n-p$	$\mathbf{Y}'\mathbf{Y}-\hat{\beta}'\mathbf{X}'\mathbf{Y}$	$\frac{1}{n-p}(\mathbf{Y}'\mathbf{Y}-\hat{\beta}'\mathbf{X}'\mathbf{Y})$	
Lack of fit	$\ell-p$	LFSS	LFSS/$(\ell-p)$	Compare [see Eq. (4.1.1)]
Error	$n-\ell$	ESS	ESS/$(n-\ell)$	

4.2 Comparison of Models Using R^2

A second measure of the effectiveness of the regression model is the square of the multiple correlation coefficient R:

$$R^2 = \frac{\hat{\beta}'\mathbf{X}'\mathbf{Y} - n\bar{y}^2}{\mathbf{Y}'\mathbf{Y} - n\bar{y}^2}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$\mathbf{Y}'\mathbf{Y}$ = total sum of squares,

$\hat{\beta}'\mathbf{X}'\mathbf{Y}$ = regression sum of squares.

This is the proportion of the total sum of squares (adjusted for the mean) which is accounted for by the model and should be used to compare different models using the same data. In general, for $n \geq 2p$, a value of $R^2 < 0.85$ indicates a poor model. An $R^2 > 0.90$ indicates a good model, and hopefully $R^2 > 0.95$. It should be recognized that when $p = n$, $R^2 = 1$, and when p is large relative to $n - p$, R^2 is close to 1.

5. CONFIDENCE INTERVALS

5.1 Confidence Intervals About β

Under the assumption that $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, we can place confidence intervals about individual estimates of β or jointly about all β. Earlier it was stated that since

$$\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) ,$$

then

$$\hat{\beta} \sim N[\beta, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2] .$$

Thus the variance of $\hat{\beta}_i$ $(i = 1, \ldots, p)$ is given by $\hat{\sigma}^2 c_{ii}$, where c_{ii} is the ith diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. The $100(1 - \alpha)\%$ confidence interval about β_i is

$$\left(\hat{\beta}_i - t_{(n-p,\,1-\alpha)}\sqrt{c_{ii}\hat{\sigma}^2} \leqq \beta_i \leqq \hat{\beta}_i + t_{(n-p,\,1-\alpha)}\sqrt{c_{ii}\hat{\sigma}^2}\right) \qquad \text{for } i = 1, \ldots, p ,$$

where $t_{(n-p,\,1-\alpha)}$ is the $1 - \alpha$ point of the t distribution with $n - p$ degrees of freedom. However, the confidence interval statement says nothing about the remaining coefficients. It is wrong to compute a similar confidence interval for each coefficient and then consider all simultaneously. The proper joint confidence interval for all β_i $(i = 1, \ldots, p)$ is given by the following inequality:

$$(\beta - \hat{\beta})'\mathbf{X}'\mathbf{X}(\beta - \hat{\beta}) < p\hat{\sigma}^2 F(p, n - p) .$$

A vector of constants, β, is said to lie within the joint interval if it satisfies the above inequality. In a p-dimensional space defined by the p parameters, this inequality generates a p-dimensional hyperellipse.

In terms of the numerical example presented earlier, the individual 95% confidence intervals on β_1 and β_2 (for $n = 7$, $p = 2$, $t_{(5,0.95)} = 2.57$, $\hat{\sigma}^2 = 2.22$) are

$$\left[5.71 - 2.57\sqrt{(2.22)(0.355)} \leqq \beta_1 \leqq 5.71 + 2.57\sqrt{(2.22)(0.355)}\right]$$

and

$$\left[4.02 - 2.57\sqrt{(2.22)(0.00415)} \leqq \beta_2 \leqq 4.02 + 2.57\sqrt{(2.22)(0.00415)}\right]$$

respectively. The joint confidence interval on β_1 and β_2 simultaneously is an ellipse in the space of β_1 and β_2.

The corresponding $100(1 - \alpha)\%$ confidence interval about a linear function of the β_i is given by

$$\left\{\mathbf{r}'\hat{\beta} - t_{(n-p,\,1-\alpha)}\sqrt{\hat{\sigma}^2\mathbf{r}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{r}} \leqq \mathbf{r}'\beta \leqq \mathbf{r}'\hat{\beta} + t_{(n-p,\,1-\alpha)}\sqrt{\hat{\sigma}^2\mathbf{r}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{r}}\right\}, \qquad (5.1.1)$$

where $\mathbf{r}$ is a $p \times 1$ vector of known constants. For example, if

$$\mathbf{r} = \begin{pmatrix} +1 \\ -1 \end{pmatrix},$$

the 95% confidence interval about

$$\mathbf{r}'\beta = (+1 \quad -1)\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \beta_1 - \beta_2$$

is

$$\left[(5.71 - 4.02) - 2.57\sqrt{(2.22)(0.418)} \leqq (\beta_1 - \beta_2) \leqq (5.71 - 4.02) + 2.57\sqrt{(2.22)(0.418)}\right],$$

or

$$[-6.15 \leqq (\beta_1 - \beta_2) \leqq 9.53] .$$

5.2 Confidence Interval About $E(\mathbf{Y})$

The $100(1 - \alpha)\%$ confidence interval about

$$E(\mathbf{Y}) = \sum_{i=1}^{p} \beta_i x_i$$

is similar in form to Eq. (5.1.1):

$$\left\{ \sum_{i=1}^{p} \hat{\beta}_i x_i - t_{(n-p,1-\alpha)}\sqrt{\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \leqq \sum_{i=1}^{p} \beta_i x_i \leqq \sum_{i=1}^{p} \hat{\beta}_i x_i + t_{(n-p,1-\alpha)}\sqrt{\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \right\},$$

where $\mathbf{x}_0$ is a vector of constants representing specific values of the independent variables. This confidence interval gives us a measure of how well we know the mean value of the response at a given set of levels of the independent variable. By plotting the response against one independent variable, the above interval generates a confidence band about the regression curve.

5.3 Confidence Interval About Mean of Future Observations

A confidence interval may be placed about the mean of one or more future observations. Suppose n experiments have been completed and the least-squares estimate of β has been computed. It is now desirable to perform k additional experiments at a particular set of x's, say $\mathbf{x}_0$. The $100(1 - \alpha)\%$ confidence interval about the mean $\bar{y}_0$ of these k *proposed* experiments is given by

$$\left\{ \hat{\beta}'\mathbf{x}_0 - t_{(n-p,1-\alpha)}\sqrt{\hat{\sigma}^2\left[\frac{1}{k} + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right]} \leqq \bar{y}_0 \leqq \hat{\beta}'\mathbf{x}_0 + t_{(n-p,1-\alpha)}\sqrt{\hat{\sigma}^2\left[\frac{1}{k} + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right]} \right\},$$

where $\hat{\beta}$, $\mathbf{X}$, $\hat{\sigma}^2$, and $t_{(n-p,1-\alpha)}$ were defined earlier under Sect. 3.4. When $k = 1$, we have the confidence interval about a single future observation.

6. SOME COMMENTS ON STEPWISE REGRESSION

Most computer programs which generate least-squares estimates of parameters in linear regression belong to a class commonly referred to as stepwise regression programs. In general, stepwise regression programs build models by successively adding single parameters to the pre-

viously determined model according to some scheme. The criterion of such a program might be to add that parameter which causes the greatest reduction in the residual sum of squares. Of course, there must be some ultimate model which is prescribed in the input. If the choice of all models suitable for your data is a subset of the successive models that are built and if the building criterion is compatible with your problem, the concept of stepwise regression may be of some use. However, in most cases, the vast amount of output of such a program clouds the important information to be gained from its use. An excellent account of the various types of regression programs is given in Draper and Smith.[3]

The t tests in the printed output of most programs represent tests made on individual coefficients and do not provide a joint test of all parameters.

The numerical output from one program may differ substantially from the corresponding output of another program. This can be attributed primarily to the use of different matrix inversion subroutines. If the $\mathbf{X}'\mathbf{X}$ matrix is nearly singular, the results may be wild. When in doubt, the best way to check this is to compare $(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{X}$ with the identity matrix. A second reason for differing outputs is in connection with computing the sums of squares associated with each parameter. These values depend on the order in which they were calculated (unless the $\mathbf{X}'\mathbf{X}$ matrix is diagonal). A competent statistician should be consulted before using these sums of squares.

The Statistics Section uses special computer programs which are based on the University of California BIMD series and are altered to meet our specific needs and philosophy. Other programs are equally as good if not better.

7. APPENDIX

Fundamentals of Matrices

An $n \times m$ matrix is a rectangular array of numbers having n rows and m columns (n, $m = 1, 2, \ldots$). Thus

$$\mathbf{X} = [x_{ij}] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ & \cdots & & \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} .$$

The x_{ij} ($i = 1, \ldots, n$; $j = 1, \ldots, m$) are called the elements of the matrix, the first letter (or number) in the subscript referring to the particular row and the second letter (or number) referring to the particular column. The size of a matrix is called the order. Thus $\mathbf{X}$ is of order "n by m"

[3]N. Draper and H. Smith, *Applied Regression Analysis*, Wiley, New York, 1966.

and is written $\mathbf{X}^{(n \times m)}$. The order is usually omitted when the context is clear. If $m = 1$ the matrix is a column vector; if $n = 1$ the matrix is a row vector. A matrix of order 1×1 is a scalar. A matrix is square when $n = m$, and a symmetric matrix is a square matrix in which $x_{ij} = x_{ji}$ $(i, j = 1, \ldots, n)$.

The set of elements $\{x_{ii}\}$ (for $i = 1, \ldots, n$) constitute the main diagonal. A diagonal matrix is one in which only the diagonal elements differ from zero. If all the diagonal elements are equal to 1, such a matrix is called the identity matrix.

Basic Operations Using Matrices

The transpose of $\mathbf{A}$, denoted $\mathbf{A}'$ (or $\mathbf{A}^T$), is the matrix $\mathbf{A}$ with the rows and columns interchanged. Thus $[a_{ij}]' = [a_{ji}]$. Note that $(\mathbf{A}')' = \mathbf{A}$.

The addition of two matrices of the same order is performed by adding to each element of one matrix the corresponding element of the second matrix:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix} .$$

In multiplication, matrix $\mathbf{B}$ is premultiplied by $\mathbf{A}$ if the sequence $\mathbf{AB}$ is preserved. If $\mathbf{BA}$, $\mathbf{B}$ is postmultiplied by $\mathbf{A}$. The matrices $\mathbf{A}$, $\mathbf{B}$ are said to be conformable for multiplication if the number of columns in $\mathbf{A}$ is equal to the number of rows in $\mathbf{B}$ when $\mathbf{A}$ is postmultiplied by $\mathbf{B}$. If $\mathbf{A}$ is premultiplied by $\mathbf{B}$ the number of rows in $\mathbf{A}$ must be equal to the number of columns in $\mathbf{B}$. If $\mathbf{A}$ and $\mathbf{B}$ are of order $n_1 \times m_1$ and $n_2 \times m_2$, respectively, and $m_1 = n_2$, then the product $\mathbf{AB}$ is of order $n_1 \times m_2$. Each element of $\mathbf{C} = \mathbf{AB}$, say c_{ij}, is obtained by multiplying every element of the ith row of $\mathbf{A}$ by the corresponding element of the jth column of $\mathbf{B}$ and summing the resulting m_1 products. Thus

$$c_{ij} = \sum_{k=1}^{m_1} a_{ik} b_{kj} .$$

Note that $\mathbf{AB}$ does not necessarily equal $\mathbf{BA}$.

The inverse of a nonsingular square matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is defined to be that matrix which satisfies

$$(\mathbf{A})(\mathbf{A}^{-1}) = \mathbf{I} \text{ (identity matrix)} .$$

Matrix $\mathbf{A}$ is nonsingular if the determinant, $|\mathbf{A}|$, is not equal to 0.

Properties of Matrices

A matrix of order $m \times n$ is said to be of rank r (for $0 \leq r \leq m \leq n$) if r represents the order of the largest submatrix which is nonsingular.

Let $\mathbf{Y}$ be an $n \times 1$ column vector of constants and $\mathbf{A}$ be an $n \times n$ matrix. Then

$$\mathbf{Y}'\mathbf{A}\mathbf{Y} = \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j a_{ij}$$

is called a quadratic form in the y's. This quadratic form is said to be positive definite (positive semidefinite) if and only if $\mathbf{Y}'\mathbf{A}\mathbf{Y} > 0$ ($\mathbf{Y}'\mathbf{A}\mathbf{Y} \geqq 0$) for *all* vectors $\mathbf{Y} \neq \mathbf{0}$. The matrix $\mathbf{A}$ is also said to be positive definite (positive semidefinite).

A matrix is said to be orthogonal if $\mathbf{A}'\mathbf{A} = \mathbf{I}$, that is, if

$$\sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} a_{ij} a_{ji} = 0 \ .$$

LECTURE VII: NONLINEAR ESTIMATION

John J. Beauchamp

1. INTRODUCTION

In this lecture we present a brief introduction to the subject of nonlinear estimation. This deals with the relationship between an observable random variable and one or more independent (nonrandom) variables, when that relationship can be expressed in a nonlinear functional form. This case arises when definite information is available about the form of the relationship between the random variable or response and the independent variable. When such information is available, we usually prefer to fit the more realistic model rather than a linear model which might be less realistic.

In the lecture on linear regression we discussed the fitting, by least squares, of models which were linear in the parameters of interest, that is, models of the type

$$Y = \beta_0 + \beta_1 Z_1 + \ldots + \beta_p Z_p + \epsilon \ , \tag{1}$$

where Y can represent the observable random variable; $Z_1, \ldots, Z_p$ can represent any functions of the basic independent variables $X_1, \ldots, X_k$; and ϵ can represent the "experimental error" associated with Y. Example:

$$Y = \beta_1 X_1 + \beta_2 X_1^2 \sin X_2 + \epsilon \ .$$

In this case $Z_1 = X_1$ and $Z_2 = X_1^2 \sin X_2$. As we have seen from the earlier lecture the estimation of the parameters $\beta_0, \ldots, \beta_p$ in equations of the form given by (1) is well understood.

Suppose that we assume a functional form which explains the observations apart from experimental error, that is,

$$Y = f(\xi, \theta) + \epsilon \ , \tag{2}$$

or

$$E(Y) = f(\xi, \theta) \ , \tag{3}$$

where

$$\theta = (\theta_1, \ldots, \theta_p)' \tag{4}$$

is the vector of unknown parameters to be estimated and

$$\xi = (\xi_1, \ldots, \xi_k)' \tag{5}$$

is the vector of known or controlled independent variables. There are n observations on the random variable Y corresponding to an observed value of the vector of independent variables ξ. These observations are denoted by

$$Y_u = f(\xi_u, \theta) + \epsilon_u \ , \tag{6}$$

where

$$\xi_u = (\xi_{1u}, \ldots, \xi_{ku})' \tag{7}$$

for $u = 1, 2, \ldots, n$. The actual observed values of Y_u are denoted by y_u, $u = 1, 2, \ldots, n$. Our aim is to obtain the least-squares estimate of θ, that is, obtain a value of θ that minimizes

$$S(\theta) = \sum_{u=1}^{n} [y_u - f(\xi_u, \theta)]^2 \ . \tag{8}$$

When the function f is nonlinear in the elements of θ, then $S(\theta)$ may have multiple or relative minima, whereas if f is of the form given in Eq. (1) then the least-squares estimates of the parameters are unique and also the best linear unbiased estimates for the case when the random variables ϵ_u are assumed to be independent with constant variance. It should be noted that the least-squares estimators of θ are equivalent to the maximum likelihood estimators of θ when the additional assumption is made that the ϵ_u are normally distributed, since minimizing $S(\theta)$ with respect to θ can be shown to be equivalent to maximizing the likelihood function with respect to θ. This can easily be seen by examining the likelihood function, which, for this case, can be written as

$$L(\theta, \sigma^2) = \prod_{u=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-[y_u - f(\xi_u, \theta)]^2/2\sigma^2\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp[-S(\theta)/2\sigma^2] \ .$$

2. GEOMETRY OF LEAST SQUARES

Before we present some of the procedures to estimate the vector θ, we want to examine the geometry of the situation so that we may appreciate some of the difficulties that may arise in nonlinear estimation. The contours defined by $S(\theta)$ = constant may be examined in two different ways. We may examine them in a subspace of the sample space called the estimation space, which consists of all points with coordinates given by

$$\{f(\xi_1, \theta), f(\xi_2, \theta), \ldots, f(\xi_n, \theta)\} \ ; \tag{9}$$

that is, it is the locus of all points with coordinates given by (9) as θ takes on all possible values. If the function f is linear in θ, then the estimation space is a hyperplane with dimension equal to p. In addition, the lines of constant θ are equally spaced and orthogonal. If f is nonlinear in θ, then one or both of these conditions may not be met. In Fig. 1 we have shown the sample space and estimation space for three simple models. Figure 1*a* is a simple linear model, Fig. 1*b* is intrinsically linear since it is linear after a transformation of the parameter, and Fig. 1c displays a model which is intrinsically nonlinear since it is not possible to convert it into a form like Eq. (1). If we let $\psi = e^{\theta}$ in Fig. 1*b* then it becomes like Fig. 1*a*.

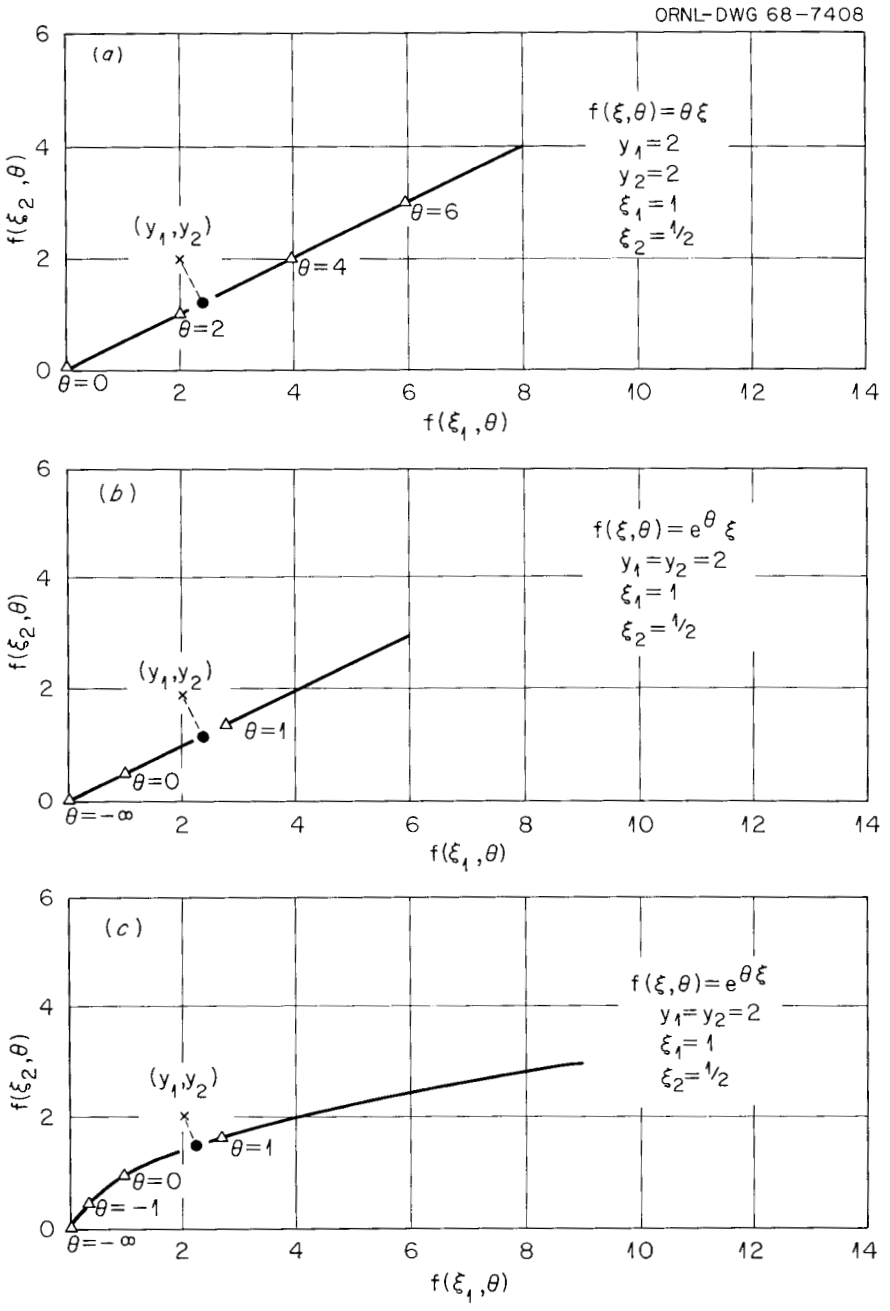


Fig. 1. Sample Space for Three Different Regression Models.

Another related method of examining contours of constant $S(\theta)$ is to examine these contours in the parameter space or θ space. If the function f is linear, then the contours of constant $S(\theta)$ in the parameter space consist of concentric ellipses. If the function f is nonlinear, then the contours are often elongated or banana-shaped. In addition, the contours for the nonlinear model may have multiple loops surrounding many stationary values which provide alternative minima for $S(\theta)$.

Example: Consider the model

$$Y = \frac{\theta_1 e^{-\theta_2 \xi} - \theta_2 e^{-\theta_1 \xi}}{\theta_2 - \theta_1} + \epsilon \ . \tag{10}$$

If θ_1 and θ_2 are interchanged, this leaves the model unaltered. Therefore, if the minimum for $S(\theta)$ is attained at $(\theta_1, \theta_2) = (\hat{\theta}_1, \hat{\theta}_2)$, the same minimum value for $S(\theta)$ is given at $(\theta_1, \theta_2) = (\hat{\theta}_2, \hat{\theta}_1)$, so that a double solution exists.

3. METHODS OF ESTIMATION

We will now discuss some methods which may be used to find the least-squares estimate of the vector θ. These methods make use of one or both of the following: (1) the gradient or steepest descent method and (2) the Gauss-Newton or linearization method. Both of these methods are iterative and involve the choice of a vector $\hat{\theta}^{(0)}$ of initial estimates of the parameters. These initial estimates may be intelligent guesses or preliminary estimates based on some available information. Hopefully, these initial values will be improved upon in the successive iterations.

3.1 Gradient or Steepest Descent Method

We will need the following definition before discussing the gradient or steepest descent method:

Definition: Let $g(\mathbf{x})$ be a real-valued function, where $\mathbf{x}$ is a vector in p-dimensional Euclidean space. The gradient of g is then defined as

$$\nabla g = \left(\frac{\partial g}{\partial x_1}, \frac{\partial g}{\partial x_2}, \ldots, \frac{\partial g}{\partial x_p} \right)' \ . \tag{11}$$

Hence we note that ∇g is vector valued and indicates a direction. If $\mathbf{x}^{(0)}$ is some fixed point in the p-dimensional space, then we can show that $\nabla g(\mathbf{x}^{(0)})$ points in the direction from $\mathbf{x}^{(0)}$ in which g increases most rapidly. Therefore $-\nabla g(\mathbf{x}^{(0)})$ points in the direction from $\mathbf{x}^{(0)}$ in which g decreases most rapidly.

For our problem, let $\hat{\theta}^{(\nu)}$ be the estimate of the vector θ at the νth iteration; then the gradient method would find a new estimate of θ by

$$\hat{\theta}^{(\nu+1)} = \hat{\theta}^{(\nu)} - \lambda \nabla S(\hat{\theta}^{(\nu)}) \ , \tag{12}$$

where λ is a positive constant. This procedure is repeated with $\hat{\theta}^{(\nu+1)}$ taking the place of $\hat{\theta}^{(\nu)}$. The calculation proceeds in this way until it is halted by some arbitrary rule. Some stop when the residual sum of squares is sufficiently small, some stop when the elements of $\theta^{(\nu+1)} - \theta^{(\nu)}$ are sufficiently small, and some stop after a fixed number of iterations. However, there is no generally accepted stopping rule.

The gradient method has the advantage that it is easy to compute and it gives rapid improvement in reducing $S(\theta)$ when $\hat{\theta}^{(\nu)}$ is far from the minimum. This method also has some disadvantages, though, and some of these disadvantages are: (1) the step size from iteration to iteration must be determined; that is, the value of the constant λ in (12) must be determined, and this has produced many variations of the method; (2) convergence can be very slow near the minimum value of $S(\theta)$; and (3) the direction of steepest descent is not unique under a change of scale.

The third disadvantage may be understood by considering the following example:

Example: The two models whose regression equations are given by

$$E(Y_u) = \theta_1 e^{-\theta_2 \xi_u} = f_1(\xi_u, \theta)$$

and

$$E(Y_u) = \theta_1 e^{-3\theta_2 \xi_u} = f_2(\xi_u, \theta)$$

for $i = 1, 2, \ldots, n$, differ only in a change of scale of the parameter θ_2. Let

$$S_1(\theta) = \sum_{u=1}^{n} [y_u - f_1(\xi_u, \theta)]^2$$

and

$$S_2(\theta) = \sum_{u=1}^{n} [y_u - f_2(\xi_u, \theta)]^2 \ ;$$

then by comparing the vectors $-\nabla S_1(\theta)$ and $-\nabla S_2(\theta)$ it can easily be seen that these two vectors do not have the same direction in the θ space. For example, let $n = 3$, $y_1 = 2$, $y_2 = 1$, $y_3 = 2$, $\xi_1 = 1.5$, $\xi_2 = 1.0$, $\xi_3 = 0.5$, $\theta_1 = 5$, and $\theta_2 = 1$; then

$$-\nabla S_1(\theta) = (-1.47562, 3.25976)$$

and

$$-\nabla S_2(\theta) = (0.51263, -5.05388) \ ,$$

which do not have the same direction in the θ space.

3.2 Gauss-Newton or Linearization Method

The Gauss-Newton or linearization method of estimating the parameters in nonlinear regression uses the results of linear least squares in a succession of stages and has been referred to by some as linear regression for nonlinear models. The first step in this method involves the expansion of $f(\xi_u, \theta)$ in Eq. (6) in a Taylor series expansion about the point $\theta^{(0)}$ of initial estimates of the elements of θ and curtailing this expansion after the linear terms in the elements of θ. That is, the function $f(\xi_u, \theta)$ is approximated by the following linear function:

$$f(\xi_u, \theta^{(0)}) + \sum_{i=1}^{p} \left[\frac{\partial f(\xi_u, \theta)}{\partial \theta_i} \right]_{\theta = \theta^{(0)}} (\theta_i - \theta_i^{(0)}) \ . \tag{13}$$

Let

$$f_u^{(0)} = f(\xi_u, \theta^{(0)}) \ ,$$

$$\beta_i^{(0)} = \theta_i - \theta_i^{(0)} \ ,$$

$$Z_{iu}^{(0)} = \left[\frac{\partial f(\xi_u, \theta^{(0)})}{\partial \theta_i} \right]_{\theta = \theta^{(0)}} \ ;$$

then it can be seen that the model in Eq. (6) is of the form, approximately,

$$Y_u - f_u^{(0)} = \sum_{i=1}^{p} \beta_i Z_{iu} + \epsilon_u \ . \tag{14}$$

That is, it is of the linear form given in Eq. (1) to the selected order of approximation. Then, near $\theta^{(0)}$, it should be true that $\dot{S}(\theta) \approx S(\theta)$, where $\dot{S}(\theta)$ is $S(\theta)$ when $f(\xi_u, \theta)$ is approximated by (13), namely,

$$\begin{aligned} \dot{S}(\theta) &= \sum_{u=1}^{n} (y_u - f_u^{(0)} - \sum_{i=1}^{p} Z_{iu}^{(0)} \beta_i^{(0)})^2 \\ &= (\mathbf{r}^{(0)} - \mathbf{Z}^{(0)} \beta^{(0)})' (\mathbf{r}^{(0)} - \mathbf{Z}^{(0)} \beta^{(0)}) \ , \end{aligned} \tag{15}$$

where

$$\mathbf{r}^{(0)} = (y_1 - f_1^{(0)} , \ldots , y_n - f_n^{(0)})' \ ,$$

$$\mathbf{Z}^{(0)} = \begin{pmatrix} Z_{11}^{(0)} & \ldots & Z_{p1}^{(0)} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ Z_{1n}^{(0)} & \ldots & Z_{pn}^{(0)} \end{pmatrix} = \{Z_{iu}^{(0)}\} \ ,$$

$$\beta^{(0)} = (\beta_1^{(0)} , \ldots , \beta_p^{(0)})' = (\theta_1 - \theta_1^{(0)} , \ldots , \theta_p - \theta_p^{(0)})' \ .$$

Since $\dot{S}(\theta)$ is a quadratic form in θ, it can be minimized with

$$\hat{\beta}^{(0)} = (\mathbf{Z}^{(0)\prime}\,\mathbf{Z}^{(0)})^{-1}\,(\mathbf{Z}^{(0)\prime}\mathbf{r}^{(0)})\ , \tag{16}$$

if the inverse of $\mathbf{Z}^{(0)\prime}\mathbf{Z}^{(0)}$ exists. A new estimate of the vector θ is found from

$$\hat{\theta}^{(1)} = \hat{\theta}^{(0)} + \hat{\beta}^{(0)}\ . \tag{17}$$

Using $\hat{\theta}^{(1)}$ in the place of $\hat{\theta}^{(0)}$, the above steps are repeated again. Let $\hat{\theta}^{(\nu)}$ be the vector of estimates of θ on the νth iteration; then

$$\hat{\theta}^{(\nu+1)} = \hat{\theta}^{(\nu)} + \hat{\beta}^{(\nu)}\ . \tag{18}$$

The above procedure is repeated in this way until it is halted by some arbitrary stopping rule. As was mentioned in the discussion on the gradient method, there are numerous stopping rules that may be adopted. The iterative linearization method for estimating the elements of θ has the advantage that (1) it can solve our estimation problem in one iteration if f is linear and (2) it usually converges rapidly when we are close to the minimum, since that is where the quadratic approximation to $S(\theta)$ should be rather good. However, this procedure may also have some drawbacks in that (1) it may converge slowly, (2) it may oscillate widely, and (3) it may fail to converge at all when it attempts to take too large a step in the parameter space. To combat some of these deficiencies there have been numerous modifications proposed to circumvent these problems. One of these modifications takes only a fraction of the $\hat{\beta}^{(\nu)}$ vector and adds to $\hat{\theta}^{(\nu)}$, that is, takes very short steps in the parameter space. For some problems this has been quite effective in achieving convergence to a minimum of $S(\theta)$.

3.3 Marquardt's Compromise

D. W. Marquardt has developed a modification of the linearization method that also has connections with the gradient method. Marquardt's method represents a compromise between the linearization method and the steepest descent method and appears to combine the best features of both while avoiding their limitations. Since the gradient and linearization methods each give a correction vector for the vector of initial estimates of the parameters, the Marquardt algorithm provides a method for interpolating between these two vectors and also for obtaining a suitable step size. The basic idea behind this algorithm is that if the linearization approximation yields a function $S(\theta)$ which is a good approximation to $S(\theta)$ only in a certain neighborhood of the current iterate on θ, then $\dot{S}(\theta)$ should be minimized only within that neighborhood and not globally. Intuitively, this should be a better procedure than merely continually halving the correction vector found by the linearization procedure, since this vector may be pointing nearly 90° away from the optimum local direction, which is the direction found by the gradient or steepest descent method. The complete reference on Marquardt's algorithm is given in the list of references. There have also been numerous nonlinear estimation procedures developed for cases when f is of a particular form, for example, if f is a sum of exponential terms. However, since it would be difficult to

discuss all of them, we will only give a reference to an extensive bibliography on the general subject of nonlinear estimation that covers many of these special techniques. This bibliography is found at the end of the chapter on nonlinear estimation in the book by Draper and Smith which is also listed in the references.

4. ESTIMATES OF VARIANCES

In order to obtain estimates of the variances and covariances of our least-squares estimates of the elements of the vector θ, which we denote by $\hat{\theta}^{(\nu)}$, some approximation must be used, since it is generally impossible to obtain exact expressions for these variances and covariances. If we assume that the observations y_u, $u = 1, 2, \ldots, n$, are observed values of n independent random variables each having a constant variance σ^2, then the covariance matrix of $\hat{\theta}^{(\nu)}$ may be approximated by

$$(\mathbf{Z}^{(\nu)\prime}\mathbf{Z}^{(\nu)})^{-1}\sigma^2 \ , \tag{19}$$

where $\mathbf{Z}^{(\nu)}$ is the $n \times p$ matrix of partial derivatives of the function f with respect to the elements of θ evaluated at $\hat{\theta}^{(\nu)}$, which has been defined earlier. In order to evaluate this covariance matrix the constant σ^2 is estimated by

$$\hat{\sigma}^2 = S(\theta^{(\nu)})/(n - p) \ , \tag{20}$$

where $S(\hat{\theta}^{(\nu)})$ is the residual sum of squares evaluated at $\hat{\theta}^{(\nu)}$. By substituting $\hat{\sigma}^2$ into (19), the estimated variance of $\hat{\theta}_i^{(\nu)}$ is found as the ith diagonal element of the resulting matrix.

5. INITIAL ESTIMATES

The iterative procedures that have been discussed above all require initial estimates of the parameters $\theta_1, \ldots, \theta_p$. It is important that as much prior information as possible be used to make these starting values as reliable as possible. With good initial values it will often be possible for an iterative method to converge to a solution faster than would otherwise be possible. In addition, good starting values may protect against the convergence of the iterative scheme to an unwanted stationary point of $S(\theta)$ when multiple minima exist or if several local minima exist, one of which is the absolute minimum. Parameter values which may be physically impossible or which do not provide the true minimum of $S(\theta)$ may result from such unwanted points. One way to help avoid such difficulties is to set up a grid of points in the parameter space where $S(\theta)$ is evaluated. From this set of values of $S(\theta)$ it will usually be possible to see what the form of the sum of squares surface is, and it may also reveal that multiple minima are possible. In any case, the grid point at which $S(\theta)$ is smallest can be used as a starting point of an iterative procedure, or a reduced grid can be examined in the best neighborhood, to obtain a better starting point.

Frequently it is possible to obtain initial estimates of the parameters in a nonlinear regression model by a graphical or visual examination approach. In order to demonstrate how the approach can be used we consider the model given by

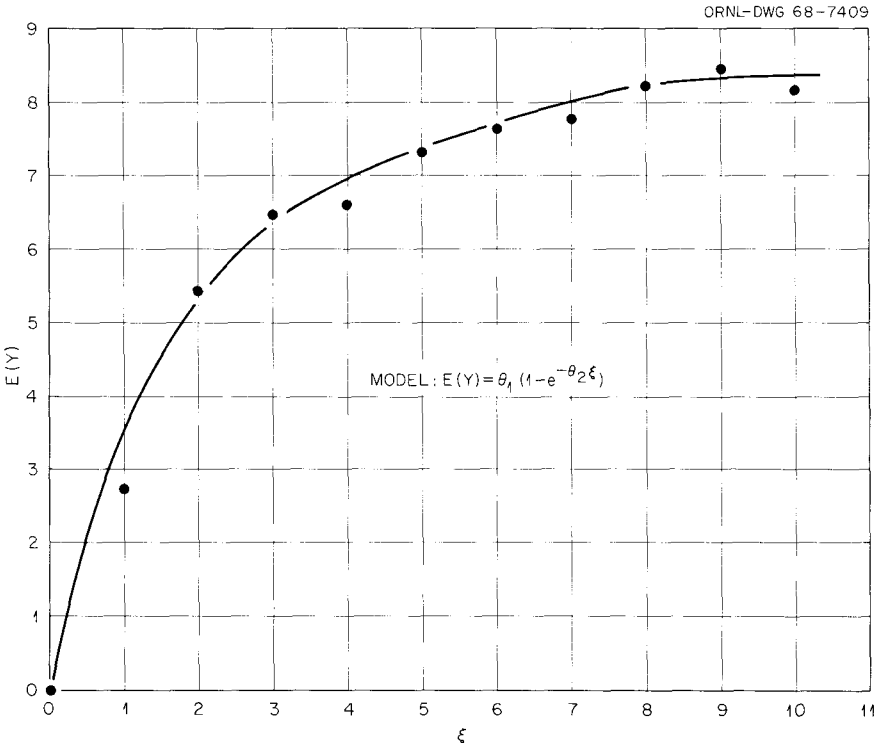


Fig. 2. Regression Curve and Data Used to Obtain Initial Estimates of θ_1 and θ_2.

$$E(Y_u) = \theta_1(1 - e^{-\theta_2 \xi_u}) \tag{21}$$

for $u = 0, 1, 2, \ldots, 10$. In Fig. 2 a plot of the observed data points (ξ_u, y_u) is given along with a smooth hand-drawn curve to approximate the function $E(Y)$ in Eq. (21). The hand-drawn curve appears to approach an asymptote of about 8.5 as ξ increases. Therefore an initial estimate of θ_1 could be taken as $\hat{\theta}_1^{(0)} = 8.5$. In order to obtain an initial estimate of θ_2 we use that point where $E(Y)$ achieves about one-half of its maximum value and then read off the value of ξ where this occurs, which is denoted by ξ^*, that is, find ξ when $E(Y) \approx 4.25$. Many other points could have been used for this step, but this point was chosen since $E(Y)$ is equal to $\theta_1/2$ at this point, which implies that $e^{-\theta_2 \xi^*} = 0.5$. We then estimate θ_2 by

$$\hat{\theta}_2^{(0)} = \frac{-\ln 0.5}{\xi^*} .$$

From Fig. 2, $\xi^* \approx 1.25$ and $\hat{\theta}_2^{(0)} \approx 0.55$. Although we are unable to make any statements about the statistical properties of $\hat{\theta}_1^{(0)}$ and $\hat{\theta}_2^{(0)}$, they are easily found and can help in choosing a particular region of the θ space where we evaluate $S(\theta)$. In addition, these graphical estimates could be used as initial estimates in the iterative estimation procedures described earlier.

REFERENCES

Draper, N. R., and H. Smith, *Applied Regression Analysis,* Wiley, New York, 1966.

Marquardt, D. W., "An Algorithm for Least-Squares Estimation of Nonlinear parameters," *J. Soc. Ind. Appl. Math.* **2**, 431–41 (1963).

LECTURE VIII: ANALYSIS OF VARIANCE

David G. Gosslee

1. INTRODUCTION

Repeated observations generally exhibit variation due to changing conditions, observers, material, and other factors. These factors can seldom, if ever, be controlled easily and inexpensively. Thus many factors are controlled by the experimental design, and other factors are controlled in a probabilistic sense by randomization and replication. A properly designed experiment allows efficient analysis of the variation due to specific factors and unbiased estimation of the effects of factors controlled by design.

Sir Ronald Fisher (1925) originally defined analysis of variance as "separation of variance ascribable to one group of causes from the variance ascribable to other groups." The analysis of variance is a statistical method that has been widely used for many years since its development by Fisher. Thus many books and journal papers contain the theory, method, designs, principles, calculations, and examples necessary for intelligent use of the method. Many programs for computers exist to perform the calculations for a variety of designs.

In previous lectures in this series the analysis of discrete observations and distribution-free tests were developed and described. The analysis of variance is a method of analyzing continuous observations assuming that the error term distribution is Gaussian. The design factors can be quantitative or qualitative.

The development of linear regression also was given in a previous lecture using the general linear hypothesis theory. The analysis of variance can be derived from the theory, although a difficulty arises since the design matrix is singular and not of full rank. This can be overcome by reparameterization or by use of the generalized inverse (Kempthorne 1952, Graybill 1961, and Searle 1966).

I will speak of single observations on an experimental unit such as the diameter of a reactor fuel pellet or of "one at a time" analyses of multiple measurements which could include, for example, other measurements on the pellet. Methods of analyzing multiple measurements include the analysis of covariance and multivariate analysis of variance.

2. MODELS

The analysis of variance is used to analyze observations in experiments and surveys in social, physical, and biological sciences. The variation might be considered in some investigations as variance per se, while in others it might be considered as the variation caused by effects of factors on the mean levels. The models for the two cases are called variance components and fixed models respectively. The mixed model which contains both fixed and random effects will not be discussed.

2.1 Variance Components Model

Interest in the relative variation or variance due to different sources is typical in genetic studies. The variation due to inheritance (genotypic variation, σ_g^2) is compared with the variation due to the interaction of inherited traits with the environment (phenotypic variation, σ_p^2). The calculation of σ_g^2 and σ_p^2 will be described in Sect. 3. Other examples are common, such as the investigation of the precision of manufacturing operations.

The levels of a factor are assumed to be random samples from a population of levels, and we are interested in making an inference about the variance among levels from the sample to the population. In the genetic study it might be assumed that herds of cows were sampled randomly and that cows within herds were sampled randomly. Likewise, it might be assumed in a production study that types of machines were randomly sampled and that machines of a given type were randomly sampled.

The analysis of variance for random models is used to answer questions such as:

1. What proportion of variability is due to inherited traits? That is, what is the estimate of $\sigma_g^2/(\sigma_g^2 + \sigma_p^2)$?
2. In how many generations will this ratio stabilize?
3. How much variation is caused by nonhomogeneous raw material relative to the variation caused by differences among machines?
4. What is the optimum allocation of effort on material, machines, operators, and training to reduce the variation in product?

2.2 Fixed Model

The fixed model is assumed if we are interested in the effect due to a treatment compared with a control or in the differences among several treatment effects. These differences can be expressed as variances and analyzed by the analysis of variance. If we trained operators by several different methods we might wish to test whether differences among the programs were significant in order to test hypotheses on the effectiveness of training programs. That is, we are primarily interested in testing differences among mean levels.

3. CALCULATIONS

The analysis of variance is obtained by decomposing the sum of squares of the observations into meaningful component sums of squares. Each sum of squares is divided by the corresponding degrees of freedom to obtain a variance, usually termed a mean square.

Given $y_1, y_2, \ldots, y_n$ as realizations of the random variable Y, we wish to subdivide the total sum of squares, Σy_i^2, $i = 1, 2, \ldots, n$, where the y's represent measurements on a random sample of experimental units. The treatments are allocated to the units by a restricted randomization procedure. The particular set of restrictions determines the experimental design. The model is determined by the design and the assumptions.

Table 1. Summary of Measurements in a One-Way Classification[a]

	Group 1	Group 2	...	Group i	...	Group t	
	y_{11}	y_{21}	...	y_{i1}	...	y_{t1}	
	y_{12}	y_{22}	...	y_{i2}	...	y_{t2}	
	y_{13}	y_{23}	...	y_{i3}	...	y_{t3}	
	⋮	⋮	⋮	⋮	⋮	⋮	
	y_{1j}	y_{2j}	...	y_{ij}	...	y_{tj}	
	⋮	⋮	⋮	⋮	⋮	⋮	
	y_{1r}	y_{2r}	...	y_{ir}	...	y_{tr}	
Group mean	$\bar{y}_{1.}$	$\bar{y}_{2.}$	...	$\bar{y}_{i.}$	...	$\bar{y}_{t.}$	
Grand mean							$\bar{y}_{..}$

[a]The table naturally hides the randomization scheme. The experimental units should be assigned to the groups randomly.

For example, given n observations from t groups with r units in each group, we let y_{ij} represent the jth measurement in the ith group (Table 1). The following algebraic identity demonstrates a meaningful decomposition:

$$\sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}^2 = n\bar{y}_{..}^2 + r\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij} - \bar{y}_{i.})^2 .$$

The corresponding model for this one-way classification is:

$$E(Y) = \mu + \tau_i ,$$

$$y_{ij} = \mu + \tau_i + e_{ij} .$$

Thus a measurement is represented as a linear combination of the mean μ, the effect of the ith group τ_i, and the error e_{ij}. The summary analysis of variance is shown in Table 2. The expected mean squares indicate that the statistics s_T^2 and s_E^2 can be used to test hypotheses and make estimates on the parameters σ_τ^2 and τ_i.

In a previous lecture we learned that the ratio of two independent random variables each having a chi-square distribution is a random variable having an F distribution. In this analysis of variance

$$(t-1)s_T^2/(\sigma^2 + r\sigma_\tau^2)$$

and

$$t(r-1)s_E^2/\sigma^2$$

Table 2. Summary of Analysis of Variance Calculations

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Expected Mean Squares	
				Random Model	Fixed Model
Total	$n = rt$	$\sum_{j=1}^{r} \sum_{i=1}^{t} y_{ij}^2$			
Mean	1	$n\,\bar{y}_{..}^2$			
Among treatments	$t - 1$	$r \sum_{i=1}^{t} (\bar{y}_{i.} - \bar{y}_{..})^2$	s_T^2	$\sigma^2 + r\sigma_\tau^2$	$\sigma^2 + r \sum_{i=1}^{t} \tau_i^2/(t-1)$
Within treatments	$t(r - 1)$	$\sum_{j=1}^{r} \sum_{i=1}^{t} (y_{ij} - \bar{y}_{i.})^2$	s_E^2	σ^2	σ^2

are independently distributed as chi-square with $t - 1$ and $t(r - 1)$ degrees of freedom, respectively, and

$$F = \frac{s_T^2/(\sigma^2 + r\sigma_\tau^2)}{s_E^2/\sigma^2}.$$

The test of the hypothesis that $\sigma_\tau^2 = 0$ is easily formulated since F then reduces to s_T^2/s_E^2. However, if the ratio of interest is $\sigma_\tau^2/(\sigma^2 + \sigma_\tau^2)$ and not $\sigma_\tau^2/(\sigma^2 + r\sigma_\tau^2)$, then the development of confidence limits on the ratio is more complex.

The mean squares in Table 2 for the random model can be related to the genetic example by considering that the variability among herds represents genetic variation and the variability among cows within herds represents phenotypic variation. It is assumed that a random sample of t herds and r cows within herds was selected. The ratio

$$\frac{\sigma_\tau^2}{\sigma^2 + \sigma_\tau^2}$$

can be estimated by

$$\frac{s_T^2 - s_E^2}{s_T^2 + (r - 1)s_E^2}.$$

For the fixed model we can consider the treatments to be the training methods mentioned earlier. Then $F = s_T^2/s_E^2$ is a statistic that is obviously sensitive to variation in training methods since s_T^2 is an unbiased estimate of

$$\sigma^2 + r\ \Sigma\tau^2/(t-1)\ .$$

Furthermore, the distribution of F is the Fisher-Snedecor distribution when $\Sigma\tau_i^2 = 0$. Thus we can test the hypothesis that $\tau_1 = \tau_2 = \ldots = \tau_t = 0$ using F.

4. TESTS OF HYPOTHESES

Given a properly designed and conducted experiment we can obtain efficient and unbiased estimates of the parameters and linear functions of the parameters in the model. In addition, each linear combination that is estimable has a corresponding unique mean square that can be used to test the hypothesis that the linear function of the parameters is zero.

For example, if

$$\lambda = \tau_1 - \tau_2\ ,$$

$$\ell = \bar{y}_{1.} - \bar{y}_{2.}\ ,$$

and $E(\ell) = \lambda$, then λ is estimable and $s_T^2 = r\ell^2/2$ is a unique mean square. A test of the hypothesis that $\tau_1 - \tau_2 = 0$ (i.e., that $\tau_1 = \tau_2$) is obtained by computing $F = s_T^2/s_E^2$ and comparing it with the $100(1-\alpha)$ percentile of the F distribution with 1 and $t(r-1)$ degrees of freedom, $F_\alpha[1,\ t(r-1)]$. With two treatments, one degree of freedom, an equivalent test can be made by calculating a t statistic which will be the square root of the F statistic. In this case,

$$t = (\bar{y}_{1.} - \bar{y}_{2.})/\sqrt{2s_E^2/r}\ .$$

For three treatments and two degrees of freedom we can construct many pairs of independent contrasts such that the sum of squares for each pair is the same. For example, for one pair we can choose

$$\ell_1 = \bar{y}_{1.} - \bar{y}_{3.} \qquad \text{and} \qquad \ell_2 = \bar{y}_{1.} - 2\bar{y}_{2.} + \bar{y}_{3.}$$

and for another pair

$$\ell_3 = \bar{y}_{1.} - \bar{y}_{2.} \qquad \text{and} \qquad \ell_4 = \bar{y}_{1.} + \bar{y}_{2.} - 2\bar{y}_{3.}\ .$$

Then

$$s_1^2 = r\ell_1^2/2\ , \qquad s_2^2 = r\ell_2^2/6$$

and

$$s_3^2 = r\ell_3^2/2\ , \qquad s_4^2 = r\ell_4^2/6$$

are single degree of freedom mean squares associated with the corresponding contrasts. It can be shown that

$$s_1^2 + s_2^2 = s_3^2 + s_4^2 = r\ \Sigma(\bar{y}_{i.} - \bar{y}_{..})^2 ,$$

which is equal to the among treatments sum of squares in Table 2 when $t = 3$.

Furthermore, each contrast can be used to test a hypothesis using the associated mean square. Thus the most meaningful pair can be chosen.

This procedure can be generalized to more than three treatments and demonstrates the manner in which the F test on the equality of several means is a generalization of the Student t test on two means.

5. ASSUMPTIONS

Let me introduce this section with a quote from Mainland (1968): "Until the very last step, when we take the variance ratio (F) to a published table, the method has no connection with the Gaussian or any other particular frequency distribution. It represents a neat, exact and rather remarkable relationship that exists in any group of numbers that is divided into two or more subgroups (classes) either equal or unequal in size."

As a first step, the analysis of variance table is a meaningful way to summarize the observations. If, secondly, we choose to estimate parameters in the model or perform tests of hypotheses, in a probability sense, then we need to assume that the random components are independently and normally distributed about zero mean and with common variance.

Normality is often justified after examining data. Repeated observations on identical material can be analyzed to test the assumption of normality. Also, the residuals from the model can be analyzed in a similar manner. The residuals are the differences between the observed values and the corresponding estimated values. For example, given $\hat{\mu}$ and $\hat{\tau}_i$ to be maximum likelihood estimates of the parameters μ and τ_i, then

$$e_{ij} = y_{ij} - \hat{\mu} - \hat{\tau}_i$$

are the n residuals for the model previously stated. Graphical methods are particularly useful in examining residuals.

6. DESIGN OF EXPERIMENTS

The proper decomposition depends on the design. In some cases poor planning or loss of experimental units and observations causes difficulty, and a statistician is sought to "salvage something" from the experiment. A statistician should be consulted to aid in planning the experiment.

A good design will ensure unbiased estimates of the effects of factors and an unbiased estimate of the error variance. The design can be chosen either to obtain the most sensitive tests for a given effort or to minimize the effort for a given sensitivity.

The principles of experimental design are also directed toward ensuring that the assumptions are valid. Thus randomization is essential to allow us to assume independence of error terms, obtain an unbiased estimate of the error variance, and to make valid inferences.

One restriction on randomization is a type of sampling, termed systematic, in which the experimental units are chosen in a cyclic manner so that every kth unit is chosen. For example, every 5th unit on a production line might be chosen for testing, every 3d plot in an agricultural experiment might be selected for treatment, or every 12th name on a list is chosen for interview in a survey.

Systematic sampling is often less expensive, and the means are sometimes estimated more accurately. However, the estimate of the variance is often biased. In the examples mentioned above, such bias would exist if there was a trend in quality of the item under production, if there was a fertility gradient in the agricultural trial, and if the list of persons had been ordered by a variable correlated with a variable measured in the survey. Obviously, the means could be biased if a cycle exists in the sequence of units, for example, if the units on the production line came from five machines and always in the same sequence.

7. TRANSFORMATIONS

If normality cannot be assumed or if the variance is not homogeneous, a mathematical transformation of the observations can cause the assumptions to be more nearly valid. Since investigators are accustomed to the concept that measuring devices and experimental units often perform transformations on the variable of interest, the idea of a mathematical transformation should seem somewhat natural.

The investigator very often measures a variable that is functionally related to the variable of interest. The function may be known or unknown. An example from several divisions of the laboratory is the measurement of the diameter of an experimental unit, the diameter then being transformed into surface, volume, or mass, depending on the investigation. The statistical analysis is made on the values most nearly validating the assumptions.

Transformations are used in analysis of variance most frequently to correct for nonhomogeneity of the error variance. The need might be discovered by plotting residuals or by plotting means and variances for subgroups of the data. Also, specific transformations are proposed for certain types of data on a theoretical or a common usage basis.

Some variance-stabilizing transformations based on statistical theory are: (1) square root if the variance is proportional to the mean, (2) logarithm if the standard deviation is proportional to the mean, and (3) inverse sine for proportions.

The square root transformation is common for counts less than 100 and also for proportions less than 0.2.

A transformation also affects the additivity of effects and the distribution of the error term. For example, the logarithmic transformation will change a multiplicative model to an additive model. A transformation that stabilizes the error variance is expected to make the distribution more nearly normal. As we learned from a previous lecturer, the normal distribution is the only distribution in which the variance is independent of the mean.

8. MULTIPLE COMPARISONS

In the analysis of variance situation the investigator usually has some specific tests planned and also wishes to search for significant effects which could not be anticipated. Thus the standard procedure for testing a few carefully posed hypotheses needed to be, and was, modified to include multiple comparisons. Even if some of the specific hypotheses could be included in a general multiple comparison procedure they should be considered separately in order to make the overall procedure more sensitive or powerful. This is an example of providing more powerful tests by including more information in the procedure.

For example, a specific comparison of the mean of a control group with the mean of several treated groups or comparison of the control with each treatment individually is more powerful than a multiple comparisons test among all the group means.

A multiple comparisons test is a procedure for performing all possible pairwise comparisons. If the standard t test is used, too many significant results will be declared. That is, the actual level of significance is larger than the nominal level and increases rapidly as the number of contrasts increases. The new methods allow multiple tests to be performed and maintain a chosen level of significance.

The text by Steel and Torrie (1960) has an excellent discussion of these procedures.

9. CONCLUDING REMARKS

The assumptions are realistic in many experimental situations, and the analysis of variance method is a widely used tool for the following reasons:

1. In many situations the errors, which are often a composite of several sources of error, are random variables which are approximately Gaussian, independent, and additive.

2. The analysis of variance tests and estimation procedures are robust. That is, they are not highly sensitive to departures from normality. This is particularly true for tests of hypotheses on means. Caution is in order for tests of hypotheses on variances.

3. Transformations can be made easily on observations to make the error terms approximately normal.

REFERENCES

Fisher, R. A., *Statistical Methods for Research Workers*, Hafner, 1925.

Goulden, C. H., *Methods of Statistical Analysis*, Wiley, 1952.

Graybill, F., *An Introduction to Linear Statistical Models*, vol. I, McGraw-Hill, 1961.

Kempthorne, O., *Design and Analysis of Experiments*, Wiley, 1952.

Mainland, Donald, *Notes on Biometry in Medical Research*, Note 21, VA Monograph 10–1, Suppl. 3, March 1968.

Searle, S. R., *Matrix Algebra for the Biological Sciences*, Wiley, 1966.

Steel, R. G. D., and J. H. Torrie, *Principles and Procedures of Statistics*, McGraw-Hill, 1960.

LECTURE IX: DESIGN OF EXPERIMENTS AND RESPONSE SURFACE METHODOLOGY

Toby J. Mitchell

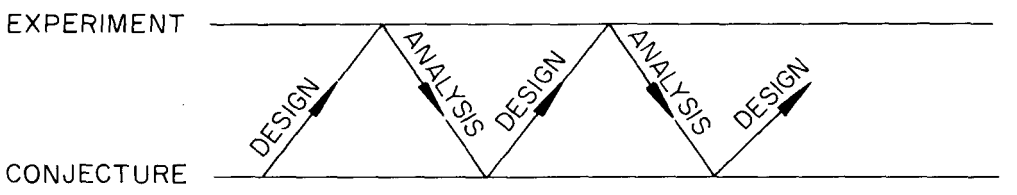
1. INTRODUCTION

In its widest sense, experimental design can be defined briefly and directly as the planning of experiments. In this lecture, we shall center our attention on experimental design from the statistical point of view. That is, we shall view the results of the experiment in terms of some sort of probabilistic mathematical model, and we shall see what principles of experimental design emerge.

Why has the statistician become involved in this area? Generally he is not knowledgeable in the subject field in which the experiment is being conducted, and it is fair to ask why he feels able to contribute anything at all. The answer is based to some degree on the fact that virtually all inferences concerning the results of experiments are, implicitly or explicitly, probabilistic in nature. In addition, there is an increasing tendency to use probabilistic models to describe the actual results of the experiment itself. This dual use of probability to describe not only the results of experiments but the inferences which can be made from these results has brought the statistician, whose business is applied probability, into the middle of scientific experimentation.

Many experimenters, recognizing this, dutifully (and sometimes reluctantly) take their data to a statistician from time to time. Far from being grateful, the statistician is often heard to grumble because he was not consulted when the experiment was still in the planning stage. Disagreeable or not, the statistician has a point, and it is this: In order to carry out a reasonably valid statistical analysis of the data, it is desirable first that the data be collected in such a way that certain probabilistic assumptions are satisfied. If this is not possible, then the data should be collected in such a way that failure to satisfy such assumptions will not foul up the inferences made from the results. In any case, the manner of collecting the data, that is, the design of the experiment, is important to the validity of the analysis. If the statistical principles of experimental design are ignored, it is quite possible to collect an enormous set of data for which there are no known methods of making valid inferences about the results.

Experimental design is, in other words, an integral part of the whole iterative process which is involved in scientific experimentation. Professor George Box of the University of Wisconsin often illustrates this point by means of the following diagram:



This diagram illustrates how the prior knowledge of the experimenter is first used to formulate a *conjecture* about the underlying physical mechanism of the system he is studying. He uses this conjecture to *design* an *experiment,* the results of which are then *analyzed.* The results of the analysis help him formulate a new conjecture, which in turn leads to a new design, and so forth. This iterative process is continued either until the goals of the project have been reached or, perhaps more realistically, until there are no more funds available.

Just as the statistician is often quite ignorant of the subject field of the experimenter, so the experimenter often lacks a deep understanding of statistics. Even if he recognizes the importance of statistical experimental design, he may feel so intimidated by the prospect of getting involved with statistics that he essentially ignores the formal aspects of experimental design altogether. Fortunately, the common sense of a good experimenter often leads him to behave just as he would if he knew all the statistical rules. Most of the basic principles of experimental design are, as we shall see, almost obvious, and simple adherence to them will generally be enough to keep the experimenter out of trouble. In some situations, however, especially those in which an explicit mathematical model is conjectured, a more sophisticated approach is necessary if the experimenter is to squeeze as much information as possible out of his data. In any case, the better his experiment is designed, the more useful information he can expect to get out of it.

In this lecture, we shall first discuss a few of the basic "common sense" principles, which will be presented in connection with some of the older "standard" designs. We shall then consider the designs which have arisen more recently in connection with response surface methodology.

2. BASIC PRINCIPLES OF EXPERIMENTAL DESIGN

2.1 Classification of Variables

Before beginning a discussion of basic design principles, it will be helpful to consider the different types of variables present in an experiment. We shall classify these as follows:

1. Response variables. These are the variables which correspond to the results of the experiment. They are represented by the observed data and are not directly under the control of the experimenter. There are generally many response variables corresponding to each experimental run, though only a few may be of sufficient interest to be measured.
2. Controlled, measurable variables.
   a) Variables of interest. The purpose of the experiment is generally to discover what effect changes in these variables have on the response variables. Therefore the levels of these variables, which can be controlled by the experimenter, are deliberately altered during the course of the experiment.

   b) Variables which are not of interest. The main reason for considering these is to control them in such a way that their effects on the response variables will not bias conclusions about the effects of the variables of interest.
3. Uncontrolled variables. These variables also affect the response variables and are, in fact, the cause of the "random error" associated with the response.

Most statistical models consider an observed response y to be composed of two parts:

1. the "true value" η of that response, where η depends only on the levels of the controlled variables, and
2. the "random error" ϵ caused by the uncontrolled variables.

Often these two parts are assumed to act additively, so the basic mathematical model is of the form:

$$y = \eta + \epsilon \ , \tag{2.1.1}$$

where η is some function of the controlled variables and ϵ is presumed to be a random variable.

The usual goal of a good experimental design is to perform the experiment in such a way that the nature of the relationship between η and the controlled variables will become clear in spite of the uncontrolled variation.

2.2 Randomization

One of the most important principles of experimental design is *randomization.* Probably the simplest application of this technique decides the order in which to perform a set of experimental runs, after all other aspects of the design have been settled. If there are n runs in the experiment, this sort of randomization can be accomplished by writing the integers from 1 to n on separate pieces of paper and then drawing them out of a hat, one by one. The order in which the numbers happen to be drawn specifies the order in which the runs are to be performed. (Equivalently, a table of random numbers could be used.)

What is the purpose of this? First, we must recognize that no experiment is under perfect control, hence the inclusion of the random error term (ϵ) in the basic model (2.1.1). We can write this model in terms of the individual observations:

$$y_i = \eta_i + \epsilon_i \ , \qquad i = 1, 2, \ldots, n \ , \tag{2.2.1}$$

where n is the number of runs in the experiment. If we now assume that each ϵ_i depends only on the state of the uncontrolled variables, which in turn depends only on the time of the ith run, then we can regard the ϵ's as being "fixed" by the choice of times at which the experimental runs are to be performed. By choosing the *order* of the runs we are, in effect, "assigning" the η's to the fixed ϵ's. There are, in fact, $n!$ possible assignments which could be made. Even though we do not know the magnitudes of the errors, we can at least give each experimental arrangement an *equal chance* by drawing the arrangement, as from a hat, in such a way that the probability, prior to the draw, of obtaining any particular arrangement is $1/n!$.

The effect of this randomization procedure is to ensure a "fair game." For example, suppose the ϵ's, unknown to us, increase as the experiment progresses. If we fail to randomize the order of the runs and choose instead to do the experiment in the most convenient way, we shall unwittingly inflate all the readings y_i corresponding to the sets of conditions which were convenient for us to run toward the end of the experiment. Of course, we can never give each

experimental run an equal random error. What we can do through randomization, however, is to give each experimental run an equal chance of being associated with any particular error from the set at hand.

It should be noted here that we have not, in the above, made any assumptions about the distribution of the ϵ's – in fact, we have not even assumed that they have a distribution. There are some experiments in which the only probabilistic structure assumed is that which is imposed by the (known) randomization scheme itself. In other situations, when one prefers not to make any distributional assumptions about the ϵ's, one can, for testing certain hypotheses, make use of nonparametric "randomization" tests which are based only on the probability structure induced by the randomization.

Usually the statistical analysis of data is based on further assumptions which include specifying some underlying distribution for the ϵ's. Generally the assumption is that the $\{\epsilon_i\}$ behave as if they were *independent* realizations of a random variable with distribution function $F(\epsilon)$, where F is specified. The assumption of independence is crucial to most statistical analyses, more so, in fact, than the correct specification of F. In practice, one can expect the assumption of independence to be quite wrong. *If*, however, randomization is carried out and the y's are assigned randomly to the ϵ's, then (for reasons we shall not discuss here) the observed y's often behave approximately *as if* the ϵ's were drawn independently from some distribution and then tacked onto the η's to give the y's. Besides ensuring a "fair game," therefore, randomization gives our statistical assumptions a firmer footing than they would otherwise deserve.

So far, we have discussed randomization in connection with the order of performing experimental runs. More generally, randomization involves the assignment of *treatments* (combinations of the controlled variables) to experimental *units* (e.g., intervals in time).

For an example in which the experimental units are not just time intervals, let us consider a hypothetical experiment which has to do with the spinning of straw into gold. We shall suppose there are two processes, A and B, for doing this, and we wish to test experimentally to see which process is superior. We may select 20 bales of straw, say, and we decide to put 10 of these through process A and 10 through process B. Which 10 bales should we assign to process A? By the principle of randomization, we should choose at random one of the

$$\binom{20}{10}$$

possible subsets of bales out of 20, so that each subset has an equal chance of being selected. In particular, randomization will ensure that a particularly "well-nourished" bale (i.e., one which would provide, under either process, more gold than most of its fellow bales on the same process) has the same chance of being assigned to process A as to process B.

Similarly, a "dud" bale would have an equal chance of being assigned to each of the two processes, thus ensuring a fair game. Once the bales had been assigned to the two processes, we would apply randomization once more to determine the order of performing the runs.

In more complex experiments, the appropriate randomization procedure is not so obvious. In general, it is helpful to think of each experiment as being systematically designed up to a certain point, after which we throw up our hands and randomize. In the above example, for instance, we carefully specified that

1. there are to be 20 bales of straw in all,
2. 10 bales are to be assigned to each process. (2.2.2)

The restrictions (2.2.2) correspond to our systematic design. The design is not completely specified, however, until we determine

1. the assignment of bales to processes, and
2. the order of the experimental runs. (2.2.3)

The decisions (2.2.3) are made by randomization, which essentially considers all designs which satisfy (2.2.2) and gives each an equal chance of being selected.

One of the most common mistakes when randomizing has been to ignore part of the class of designs which satisfy the systematic requirements. That is, some designs which satisfy these requirements have no chance of being selected. Often the effect of this is that a choice is made from a class of designs which has more restrictions than were designed into the experiment. No harm is done until the data are analyzed, at which time the experimenter estimates his error without taking into account the added restrictions which he has unwittingly imposed through faulty randomization. An example of this type of mistake will be given later in connection with Latin square designs.

2.3 Blocking

Returning to our straw-into-gold example, we now consider a situation in which it is possible to separate the bales of straw into two parts and submit one part to process A and the other to process B. To keep the total number of tests the same, we shall now assume a total of ten bales altogether. Assuming that an individual bale is more homogeneous in "nourishment" than different bales, we can now obtain comparisons between processes without regard to the differences in nourishment between bales. Even if a bale were especially well nourished compared with the others, that fact would not affect the comparison of the two processes *within* that bale. By averaging the differences between the yields of process A and process B within all the bales, we obtain an estimate of the "true" difference which is much more precise than the "one bale, one process" procedure we discussed in Sect. 2.2, simply because the "between bale" variation is no longer associated with our estimate. The only variation due to nourishment which is now associated with our estimate is that which represents the inhomogeneity within bales.

This is an example of the technique known as *blocking*, which is simply the arrangement of the experimental units into separate blocks (e.g., bales), based on the expectation that the variation in response within blocks will be considerably lower than the variation between blocks. An experiment which is to be performed by several National Laboratories, for example, might be designed so that the experimental runs performed by a single laboratory are all considered part of the same block, while those performed by different laboratories are assigned to different blocks. In this way, overall differences in results from one laboratory to the next would not be associated with the estimates of the effects of the variables of interest, even though data from all the laboratories are used to obtain these estimates.

We should remark that most blocking arrangements are made assuming that, for a given treatment, the mean difference in response from one block to the next is simply a constant which depends on the pair of blocks involved, but which does not depend on the particular choice of treatment. In other words, the mean difference in response between treatment A, block 1, and treatment A, block 2, is exactly the same as the mean difference in response between treatment B, block 1, and treatment B, block 2. While this assumption is seldom true in practice, it is a good first step toward accounting for differences in response between blocks. Though it is implicit in most standard blocked designs, this assumption can be relaxed if more sophisticated blocking is desired.

One of the most useful applications of blocking has to do with the order in which runs are performed, a problem which has already been discussed in terms of complete randomization. If the random errors are expected to follow some sort of a time trend, it is often advisable to set up the experimental program in a sequence of subexperiments, where each subexperiment is treated as a single block. Within each subexperiment, or block, the runs would be performed in random order.

The incorporation of blocks into a design belongs to the "systematic" phase of designing the experiment. To analyze the data from a blocked design, we need to include the appropriate blocking variables in our model. (These blocking variables are generally "controlled variables which are not of interest," in terms of the classification of Sect. 2.1.) When we decide to block instead of randomizing completely, we are, in effect, considering a more elaborate model. Where are we to draw the line? We might be tempted to get carried away with our blocking and end up making every experimental run a single block. Such a design would be doomed, of course, because it complicates the model too much for the number of runs at hand. (In general, blocking a design into b blocks usually adds $b - 1$ parameters to the mathematical model.) There are no hard and fast rules which specify when to stop blocking and start randomizing, but one useful rule of thumb is that the number of parameters in the model, including the $b - 1$ blocking parameters, should be no greater than three-fourths of the total number of observations.

2.4 "Fixed Variables"

In most experiments, there is a set of controllable variables which could be varied during the course of the experiment but which are held constant, or "fixed," to avoid overcomplicating the design. These variables are often of secondary interest to the experimenter, who holds them constant so that the effects of the variables of primary interest will stand out more clearly.

While this is often a good idea, care must be taken in interpreting the results, since any conclusions drawn from the experiment are technically valid *only for the levels of the fixed variables* which pertain during the experiment. In order to make a more powerful statement, it is necessary to assume, either by conjecture or by referring to other data, that the effects of the primary variables on the response are the same no matter what the levels of the fixed variables. In the jargon of the statistician, this is equivalent to saying that there is no *interaction* between the fixed variables and the primary variables. When such interactions do exist, it is advisable to "free" the fixed variables which interact with the primary variables, by varying them in the experiment. Though this may appear to overcomplicate the experiment, there are designs, as we shall see, which are particularly good for estimating the main effects and interactions of a large number of variables, all of which are varied in a highly systematic fashion, during the course of the experiment.

2.5 Replication

Repetition, or replication, of an experiment or part of an experiment can be a very valuable feature of an experimental program. In designed experiments, replicating some of the experimental runs allows the experimenter to obtain an estimate of the random error. If, for example, we are observing yield as a function of temperature in a chemical reaction, and if we test the yield at n different temperatures, then all the observed variation in yield will come from two sources:

1. variation due to the difference in yield from one temperature to another, and
2. random variation which would be present even if the temperature were constant.

Without replicating some points, that is, making several tests at the same temperature for one or more temperature levels, we would not be able to tell how much of the total variation was due to source 1 and how much to source 2. The variation among multiple observations at a single point, however, is due strictly to source 2, and we can get an estimate of it by combining such information from all points with repeat observations.

The estimate of random error arising from repeat observations is valid no matter what model is used to fit the data. This estimate is extremely useful in checking the adequacy of any proposed model. If, for example, our estimate of random error, under the assumption that the model is correct, is much larger than the estimate of random error from repeat observations, then we have evidence that the model is inadequate.

Replication can also be used by the experimenter to check the predictive power of his conclusions from the initial experiment. Nothing adds more weight to a theory than evidence of its ability to predict the results of future experiments with reasonable accuracy.

2.6 Sequential Experimentation

Virtually all experimental programs are sequential in nature. An initial experiment is run, the data are analyzed, tentative conclusions are drawn which indicate the direction of further experimentation, and so forth. In general, the advantages of sequential experimentation make it undesirable to perform too elaborate an experiment initially, no matter how sophisticated or statistically appealing the design may appear.

Unfortunately, because of the mathematical difficulties involved in considering the design at each stage of a sequential experiment to be, in effect, a random variable depending on the results at the previous stage, much of the statistical literature has so far ignored the formal aspects of sequential design. In practice, however, it is often sufficient to "play it by ear," designing and analyzing each stage in the usual fashion (as if it were a single experiment), always utilizing the knowledge gained at previous stages. Often it is helpful to "pool" the data from a series of stages and analyze them all together as if the whole design had been laid out beforehand. When this is done, it is important to include whatever block differences there might be (from stage to stage) as terms in the model.

3. SOME STANDARD DESIGNS

3.1 Completely Randomized Design

We consider an experiment in which five treatments, A, B, C, D, and E, are to be compared. The systematic part of the design specifies that each treatment is to be tested three times. Subject to this restriction, a random order for performing the runs was selected, as follows:

$$\text{E C B B D E D A D E B C A A C} \tag{3.1.1}$$

The model corresponding to this design would probably be of the form:

$$y_i = \mu + t_{(i)} + \epsilon_i , \qquad i = 1, 2, \ldots, 15 , \tag{3.1.2}$$

where μ is an overall "base level" for the response, $t_{(i)}$ is the effect of whatever treatment appears in the ith run, and ϵ_i is the random error.

3.2 Randomized Block Design

Now suppose that, instead of the order given in (3.1.1), the result of our randomization happened to be the order AAABBBCCCDDDEEE. It would be tempting to assert that this ordering is "not random," discard it, and try again. However systematic the ordering might appear, the fact would remain that it was selected fairly and squarely, every other ordering having been given an equal chance. If we ask ourselves why this ordering is "not random," we shall probably find that the reason we are worried about the clustering of treatments is that, in the presence of a time trend, one treatment might get all the high random errors and

present a false impression of superiority. We can solve this problem by dividing the runs into three sequentially run blocks and specifying that each treatment is to appear in each block exactly once. The randomization procedure is now applied to give the order of runs *within* blocks. For example:

Block	Order
1	A D B C E
2	B A E C D
3	B A D E C

(3.2.1)

This type of design is called a *randomized block design*. It successfully prevents the severe clustering of treatments which might occur in a completely randomized design.

When the blocking parameters are incorporated into the model (3.1.2) we have an appropriate model for the design in (3.2.1):

$$y_{ij} = \mu + b_i + t_{(ij)} + \epsilon_{ij}, \qquad \begin{array}{l} i = 1, 2, 3, \\ j = 1, 2, \ldots, 5, \end{array} \tag{3.2.2}$$

where the subscript i refers to the block number and the subscript ij refers to the jth observation in the ith block. Again, μ is an overall "base level" of response, $t_{(ij)}$ is the effect of whatever treatment occurs at the ijth observation, and ϵ_{ij} is the random error at the ijth observation. The three blocking parameters b_1, b_2, and b_3 allow for a constant shift in response from block to block.

3.3 Latin Square Design

A design with even more systematic restrictions than the randomized block design was used in an experiment whose purpose was to compare the effects of the following five different types of background music on the efficiency of workers in a defense plant:

A: Country and Western

B: None

C: Viennese waltzes

D: Rock and roll

E: Sen. Dirksen

Because it is felt that the day of the week and even the week itself have an effect on worker efficiency, over and above that of the background music, the experiment was carefully planned to ensure that all five treatments (types of music) can be compared not only within each row (weeks) but also within each column (days). For example:

Week	Day				
	M	T	W	T	F
1	A	B	C	D	E
2	E	C	D	A	B
3	C	A	B	E	D
4	D	E	A	B	C
5	B	D	E	C	A

Inspection of the design shows that, indeed, each treatment appears once in each row and once in each column. This type of arrangement is called a *Latin square* design. In this particular example, it is an application of blocking with respect to two variables (weeks and days) which are not of direct interest but whose influence on the response we wish to account for in our model. As a result of using this type of design, the variation in worker efficiency due to differences between days and between weeks does not cloud up our estimates of the differences between types of background music.

To apply the principle of randomization in this case, we must select a single design from all those which satisfy the requirements of a Latin square, namely, that each letter should appear exactly once in each row and once in each column. It would not do to make up a convenient Latin square, for example,

A	B	C	D	E
E	A	B	C	D
D	E	A	B	C
C	D	E	A	B
B	C	D	E	A

and then "*randomize*" by assigning the letters randomly to the treatments, because this would restrict our selection to a particular *subset* of all 5×5 Latin squares. If the data from a design derived from such a "randomization" procedure are analyzed according to the usual Latin square analysis, the estimate of error would probably be all wrong.

The usual model corresponding to an $m \times m$ Latin square design is

$$y_{ij} = \mu + r_i + c_j + t_{(ij)} + \epsilon_{ij}, \qquad \begin{array}{l} i = 1, 2, \ldots, m, \\ j = 1, 2, \ldots, m, \end{array} \tag{3.3.1}$$

where μ is the overall "base level" of response, r_i is the effect of the ith level of the factor associated with rows, c_j is the jth level of the factor associated with columns, $t_{(ij)}$ is the effect of whatever treatment is associated with the ijth observation, and ϵ_{ij} is the corresponding random error.

Latin square designs are useful when blocking with respect to two variables, as in the above example, or when there are no interactions among the variables associated with rows, columns, and letters. When such interactions are present, however, the results of the analysis of a Latin square design can be quite misleading.

3.4 Graeco-Latin Squares

In many cases, the Latin square designs can be generalized to include yet another restriction. Whereas with the Latin square we dealt with three variables (represented by rows, columns, and letters) we shall now consider four variables (represented by rows, columns, Latin letters, and Greek letters). Again, our restrictions require that any level of any variable must be associated once and only once with any level of any other variable.

A 4×4 Graeco-Latin square is given below.

$$\begin{array}{cccc} A\alpha & B\beta & C\gamma & D\delta \\ B\gamma & A\delta & D\alpha & C\beta \\ C\delta & D\gamma & A\beta & B\alpha \\ D\beta & C\alpha & B\delta & A\gamma \end{array}$$

Note that the problem of constructing a 4×4 Graeco-Latin square is the same as that of the following parlor game:

Take the four highest cards of each suit in a regular deck of cards and arrange them in a 4×4 array in such a way that each row and each column contains exactly one card from each suit (clubs, diamonds, hearts, or spades) and one card from each denomination (jack, queen, king, ace).

The question of the existence of Graeco-Latin squares of specified dimension became a classical mathematical problem nearly 200 years before these squares were ever utilized in the design of experiments. The great Swiss mathematician Euler (1707–83), having shown that $n \times n$ Graeco-Latin squares exist when n is odd or a multiple of 4, conjectured that they did not exist for any other value of n. In 1901, the French mathematician Tarry showed Euler's conjecture to be correct for $n = 6$. It was not until 1958 that the problem was solved. Three mathematicians, R. C. Bose, S. S. Shrikande, and E. T. Parker, all specialists in the combinatorial aspects of experimental designs, finally proved Euler's conjecture to be wrong in every case except for $n = 6$. (An interesting account of this discovery is given in the November 1959 issue of *Scientific American.*)

3.5 Balanced Incomplete Block Design

Often the number of available experimental units is restricted in such a way that it is not possible to apply each treatment once within each block. A hypothetical example is the problem of having seven swimmers race in a three-lane swimming pool. An obvious approach is to arrange the race in several heats. It is not possible to compare all seven treatments (swimmers) within each block (heat). We can, however, set up seven heats in such a way that each swimmer swims in three heats and competes against every other swimmer exactly once during the course of the meet. For example:

	Heat						
	1	2	3	4	5	6	7
Swimmers	1	2	3	4	1	2	1
	2	3	4	5	5	6	3
	4	5	6	7	6	7	7

Such an arrangement is called a *balanced incomplete block design,* incomplete in the sense that the whole set of treatments does not appear within each block, and balanced in the sense that every pair of treatments appears together in the same number of blocks.

3.6 Youden Square Design

If, in the comparison of swimmers in Sect. 3.5, we feel that certain lanes are generally more favorable than others, we can arrange the heats in such a way that each swimmer, during the course of his three heats, swims once in each of the three lanes. This added restriction removes from the analysis of differences among swimmers that part of the variation which is due merely to differences among lanes. Such a design is called an incomplete Latin square, or a *Youden square,* and is shown below:

Lane	Heat						
	1	2	3	4	5	6	7
1	1	2	3	4	5	6	7
2	2	3	4	5	6	7	1
3	4	5	6	7	1	2	3

4. FACTORIAL AND FRACTIONAL FACTORIAL DESIGNS

4.1 Factorial Designs (General)

One of the simplest designs for investigating the effect of several variables on a response is the *factorial design,* in which every combination of the levels of the variables of interest is investigated. If, for example, there are m variables of interest (often called "factors"), and we consider p_j levels of the jth factor, $j = 1, 2, \ldots, m$, then the factorial design, which includes every possible combination of factor levels, consists of $p_1 \times p_2 \times \cdots \times p_m$ runs.

An advantage of factorial designs is that they are easy to construct and analyze. The regular pattern of the observations makes the data easy to "look at." If, for example, we find that some of the factors have a negligible effect on the response, then the data can often be viewed as a replicated factorial in the remaining variables.

The main disadvantage of factorial designs is that the number of runs required soon gets out of hand if the number of variables or the number of levels of the variables becomes large. In Sects. 4.3 to 4.5, we shall see how to reduce the number of runs by performing only a specially selected subset of the runs required by the full factorial.

There are several empirical models which can be fitted to the data from a full factorial design. One is the "multiway" classification model with one observation per cell, which is often discussed in connection with analysis of variance techniques. For example, consider the usual two-way classification with I rows and J columns, where I is the number of levels of factor A and J is the number of levels of factor B. This is, of course, just an $I \times J$ factorial, and the model is:

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}, \qquad \begin{array}{l} i = 1, 2, \ldots, I, \\ j = 1, 2, \ldots, J, \end{array} \tag{4.1.1}$$

where y_{ij} is the observation at the ith level of A and the jth level of B, μ is an overall base level of response, α_i is the main effect of the ith level of A, β_j is the main effect of the jth level of B, γ_{ij} is the interaction of the ith level of A with the jth level of B, and ϵ_{ij} is the random error. (If there were repeat observations, they would be used to estimate the random error. When there are no repeat observations, some of the interaction terms γ_{ij} must be assumed negligible if an estimate of error is desired.)

The form of the model (4.1.1) can easily be extended to apply to any factorial design.

A second type of model, which is really a reparameterization of model (4.1.1), is often used in connection with factorial designs. This reparameterization is performed for reasons of mathematical convenience – most of the individual terms in the model are not physically meaningful. In this lecture, we shall give an example of this model in the 3×3 case and let it go at that.

Suppose we are investigating two factors, A and B, each at three levels. For the purposes of this model, we label the levels 0, 1, and 2 for both variables. Letting y_{ij} denote the observation at the ith level of A and the jth level of B, we write the model

$$y_{ij} = \mu + (A)_i + (B)_j + (AB)_{i+j} + (AB^2)_{i+2j}, \tag{4.1.2}$$

where the subscripts, which are all integers (mod 3), indicate the "levels" of the "effects" (A), (B), (AB), and (AB^2). The physical significance of these parameters seems rather abstract. For example, $(AB^2)_k$ is defined as $(i) - (ii)$, where (i) is the mean of the treatment combinations for which the level of A plus twice the level of B is equal to k (mod 3) and (ii) is the mean of all treatment combinations.

It is fair to ask why such an apparently meaningless parameterization should be introduced at all. The answer comes when we are faced with the problem of selecting a subset of treatment combinations which is to be run instead of the full factorial. In this situation, it turns out that the parameterization (4.1.2), with its "modulo p" operations and all, is much more productive than (4.1.1). A more complete discussion of the "modular" model (4.1.2) is given in Kempthorne, 1952.

Another model which is used in connection with factorial designs is the polynomial regression model, which we shall discuss later with regard to response surface designs. This type of model

is more restrictive than the models previously discussed in that it forces the response to be represented by a polynomial in the controlled variables. It is often extremely useful, however, especially in giving an overall "picture" of the response.

There are many other kinds of models which have been used to fit data from factorial experiments. This is because factorial designs arose principally on the basis of their own merits, rather than from consideration of a particular model. In effect, models such as (4.1.1) and (4.1.2) were constructed to fit the design. More recent work in experimental design theory has tended to emphasize the opposite approach: the selection of a design to suit the current working model.

4.2 2^k Factorial Designs

A particularly important class of factorial designs, in its own right and as a building block for more elaborate designs, is the class of 2^k factorials. Although each of the k factors takes on only two levels in this type of experiment, it is often remarkable how much information can be obtained through the use of these designs.

In the 2^k factorials, the main effects and interactions have a clear physical interpretation. To illustrate, let us focus our attention on a single variable A. We see that every combination of levels of the remaining variables is associated once with the low level of A and once with the high level of A (where "low" and "high" are convenient ways of referring to the two levels of a variable). The difference between the "true" responses at these two points represents the effect of A at that particular combination of the other variables. There are $N/2$ such differences, where $N = 2^k$ = total number of observations. We now define the *main effect* of A to be the average of all these differences. The main effect of A is, in other words, indicative of the overall change in response (averaged over the whole design) when A is switched from its low level to its high level. The main effects of all the other variables are defined in a similar way.

Now suppose we consider the factorial design which corresponds only to those observations taken at the high level of factor A, and we define, for this design, the main effect of another variable B in the usual way. We then do the same thing for the factorial design corresponding to the low level of factor A. We now have (i) the main effect of B at the high level of A and (ii) the main effect of B at the low level of A. The interaction between A and B is defined to be one-half the difference between (i) and (ii), that is,

$$AB \text{ interaction} = \frac{1}{2}[(i) - (ii)] .$$

It is a measure of the change in the main effect of factor B when factor A is changed from its low level to its high level.

If we were to consider, instead, the change in the main effect of factor A when B is changed from its low level to its high level, we would obtain exactly the same formal expression. In

other words, the AB interaction is happily the same as the BA interaction. The remaining two-factor interactions can be defined in the same way.

Similarly, we can define three-factor interactions, four-factor interactions, and so forth. For example, the ABC interaction is equal to one-half the difference between the AB interaction at the high level of C and the AB interaction at the low level of C. Equivalently, the ABC interaction could be defined in terms of the change in the AC interaction between the low and high levels of B, or the change in the BC interaction between the low and high levels of A – all such definitions result in the same formal expression.

To see what form is taken by the formal expressions for the main effects and interactions, we shall consider the following 2^3 design:

A	B	C	"true" response
−	−	−	η_1
+	−	−	η_2
−	+	−	η_3
+	+	−	η_4
−	−	+	η_5
+	−	+	η_6
−	+	+	η_7
+	+	+	η_8

(4.2.1)

where we use a minus sign to represent the low level of a factor and a plus sign to represent the high level. We note that all possible combinations of levels of the three factors are included in the design. The main effect of A is, by our definition above,

$$\frac{1}{4}[(\eta_2 - \eta_1) + (\eta_4 - \eta_3) + (\eta_6 - \eta_5) + (\eta_8 - \eta_7)]$$

$$= \frac{1}{4}[-\eta_1 + \eta_2 - \eta_3 + \eta_4 - \eta_5 + \eta_6 - \eta_7 + \eta_8] \,. \tag{4.2.2}$$

The expression (4.2.2) can be obtained mechanically by multiplying each of the η's by the corresponding sign in the A column, taking the sum, and dividing by the number of plus signs (4) in the column. It turns out that the same sort of rule can be followed to obtain the expression for any interaction without having to go back to the basic definition. For example, to obtain the expression for the ABC interaction we would first form column ABC, the ith element of which is the product of the ith elements of columns A, B, and C. The ABC column is therefore:

ABC
−
+
+
−
+
−
−
+

and the ABC interaction is found by multipling the η's by the corresponding elements of this column and dividing by the number of pluses (4):

$$ABC \text{ interaction} = \frac{1}{4}(-\eta_1 + \eta_2 + \eta_3 - \eta_4 + \eta_5 - \eta_6 - \eta_7 + \eta_8) . \tag{4.2.3}$$

Expressions for all the main effects and interactions can be obtained in the same way.

Estimates of the effects are obtained merely by substituting the observed $\{y_i\}$ for the $\{\eta_i\}$ in the expressions for the effects. For example, the *estimated* ABC interaction is:

$$\frac{1}{4}(-y_1 + y_2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8) .$$

Although the distinction between the true effects (in terms of the η's) and the estimated effects (in terms of the y's) should be understood, we shall often refer to both merely as "effects."

Some appealing properties of the estimated effects are:

1. They are all uncorrelated (if the original observations are uncorrelated).
2. They all have variance $4\sigma^2/N$, where σ^2 is the variance of the original observations and N is the total number of observations. This means that for $N > 4$, the effects are much more precisely determined than the expectations of the individual observations.
3. They are usually approximately normally distributed, even though the distribution of the original observations may be quite nonnormal.

4.3 2^{k-1} Fractional Factorials

The principal disadvantage of the 2^k factorial designs is that, for k even moderately large, the number of runs required ($N = 2^k$) is excessive. In this case, we often perform only a subset of these runs, where the subset is chosen so that the important information (e.g., main effect estimates) is retained while less important information (e.g., high-order interaction estimates) is sacrificed.

Suppose, for example, we are considering four factors (which we shall label 1, 2, 3, and 4) but we want to perform only half the runs required by the full 2^4 factorial. Let us arbitrarily choose to perform only those eight runs which correspond to a plus sign in the 1234 column of

the 2^4 design. (We recall that the ith element of the 1234 column is obtained by multiplying the ith elements of the 1, 2, 3, and 4 columns of the 2^4 design, where [as in (4.2.1)] we represent the low level of a variable by a minus sign and the high level by a plus sign.) The corresponding combinations of levels are given by the following design matrix:

Factor				Observation
1	2	3	4	
−	−	−	−	y_1
+	−	−	+	y_2
−	+	−	+	y_3
+	+	−	−	y_4
−	−	+	+	y_5
+	−	+	−	y_6
−	+	+	−	y_7
+	+	+	+	y_8

(4.3.1)

This design is an example of a 2^{4-1} *fractional factorial* design, where the 4 in the superscript stands for the number of variables and the −1 indicates that this is a 2^{-1} (or one-half) fraction of the full factorial.

Note that if we construct the 1234 column in this design, it will turn out to be a column of pluses. This is to be expected, since we have deliberately chosen only those runs of the 2^4 factorial which have a + sign in the 1234 column. Symbolically, we write

$$I = 1234 \ , \tag{4.3.2}$$

where the letter I stands for a column of pluses. We call (4.3.2) the *defining relation* of the design. The defining relation, which arises from the way in which the fractional factorial is constructed, is extremely useful in determining the properties of the design. In this lecture, however, we shall not go into this important aspect of the theory.

Now suppose that we use column 1 of (4.3.1) to form the linear combination:

$$\ell_1 = \frac{1}{4}(-y_1 + y_2 - y_3 + y_4 - y_5 + y_6 - y_7 + y_8) \ . \tag{4.3.3}$$

We might expect ℓ_1, by analogy with the full factorial analysis, to be a valid estimate of the 1 effect. To our disappointment, however, we find that ℓ_1 estimates not the 1 effect but the *sum* of the 1 effect and the 234 interaction, where the effects are as originally defined in connection with the full 2^4 factorial. Similarly,

$$
\begin{array}{lll}
\ell_2 & \text{estimates} & 2 + 134 \,, \\
\ell_3 & \text{estimates} & 3 + 124 \,, \\
\ell_{12} & \text{estimates} & 12 + 34 \,, \\
\ell_{13} & \text{estimates} & 13 + 24 \,, \\
\ell_{23} & \text{estimates} & 23 + 14 \,, \\
\ell_{123} & \text{estimates} & 4 + 123 \,,
\end{array}
\tag{4.3.4}
$$

where the subscript on an ℓ indicates the column of pluses and minuses which is used to obtain that particular linear combination.

We say that effect 1 is *confounded* with (or *aliased* with) the 234 interaction, effect 2 is confounded with the 134 interaction, etc. Confounding is the price of doing only half the runs of the full factorial. If, however, we can assume three-factor interactions to be negligible compared with main effects and two-factor interactions, then ℓ_1, ℓ_2, ℓ_3, and ℓ_{123} estimate, essentially, the main effects 1, 2, 3, and 4, respectively.

In fact, the design of this example was chosen, through the defining relation (4.3.2), so that this sort of confounding would occur. In general, if we choose for the defining relation of a half fraction of a 2^k factorial

$$I = 12 \ldots k \,, \tag{4.3.5}$$

it will turn out that (1) main effects are confounded with $(k - 1)$-factor interactions and (2) two-factor interactions are confounded with $(k - 2)$-factor interactions, etc.

The same sort of confounding occurs if we choose

$$I = -12 \ldots k \,, \tag{4.3.6}$$

that is, if our half fraction consists of those runs which have a *minus* sign in the $12 \ldots k$ column of the full factorial. Had we done this in our 2^{4-1} example above, the confounding relationship (4.3.4) would have been:

$$
\begin{array}{lll}
\ell_1 & \text{estimates} & 1 - 234 \,, \\
\ell_2 & \text{estimates} & 2 - 134 \,, \\
 & \vdots & \\
\ell_{123} & \text{estimates} & 123 - 4 \,.
\end{array}
\tag{4.3.7}
$$

The design should, of course, be chosen in the light of whatever prior knowledge exists about the relative importance of the variables. The experimenter will attempt to choose a design which will confound those effects assumed to be important with those presumed to be unimportant.

In the absence of prior information, it is generally assumed that main effects are more important than two-factor interactions, which are more important than three-factor interactions, etc. This kind of rationale supports the choice of defining relations such as (4.3.5) and (4.3.6).

4.4 2^{k-p} Fractional Factorials

In the previous section, we constructed a 2^{k-1} fractional factorial by deciding to perform only those runs from the 2^k factorial which corresponded to the plus signs in a specially selected column. (As we remarked earlier, we could just as well have chosen the half fraction associated with minus signs in the selected column.) If we want to cut this fraction again in half, we can select another column, half of which is minus and half of which is plus, and do the same sort of thing.

If we do this p times, the resulting design is a 2^{k-p} *fractional factorial*, where the k in the superscript refers to the number of variables and the $-p$ refers to the fact that we have a 2^{-p} fraction of the full factorial. The number of runs in a 2^{k-p} design is, of course, 2^{k-p}.

In the original 2^k design there are, in general, $2^p - 1$ columns which, when a 2^{k-p} fraction is taken, appear in that fraction as columns of all pluses or all minuses. This means that the defining relation [see (4.3.2)] of a 2^{k-p} design contains 2^p terms, including I. The confounding relationships, which can be obtained directly from the defining relation (in a way which we shall not discuss here), are correspondingly more complicated than in the 2^{k-1} case.

In this lecture, we shall avoid a general discussion of the 2^{k-p} designs and shall give, instead, methods for constructing two very useful kinds of these designs.

Resolution III Designs

One important class of 2^{k-p} designs consists of the so-called "resolution III" designs (where the III refers to the number of variables in the "shortest" term of the defining relation, not counting I). Resolution III designs are characterized by the property that all main effects can be estimated clear of one another, but some of them are confounded with interactions involving two or more factors. If all interactions are assumed negligible, then clear estimates of all the main effects can be obtained.

Resolution III designs can be constructed simply as follows. Suppose we are investigating k variables and we wish to construct a resolution III design in 2^m runs, $m < k$. We first write down the m columns corresponding to a full 2^m factorial and label these columns 1, 2, ... , m. (In what follows, these will be called the "original m columns.") Each of the remaining columns $m+1$, $m+2$, ... , k is constructed by taking some product of the original m columns. There are $2^m - m - 1$ such products, and it does not matter which ones we select to complete our resolution III design.

To give a specific example, we now construct a resolution III design in $k = 5$ variables and $2^m = 8$ runs ($m = 3$):

1	2	3	4 = 12	5 = 13	Observations
−	−	−	+	+	y_1
+	−	−	−	−	y_2
−	+	−	−	+	y_3
+	+	−	+	−	y_4
−	−	+	+	−	y_5
+	−	+	−	+	y_6
−	+	+	−	−	y_7
+	+	+	+	+	y_8

(4.4.1)

The original m columns correspond to variables 1, 2, and 3 and form a 2^3 factorial design. The remaining variables (4 and 5) are associated with the product of columns 1 and 2 and the product of columns 1 and 3 respectively. Our design is now completely specified. If we can assume interactions negligible relative to main effects, the estimate of any main effect can be obtained simply by multiplying the signs in the appropriate column by the corresponding observations, adding, and dividing the result by the number of plus signs in the column.

Since there are $2^m - m - 1$ products of the original m columns which can be used to give the columns associated with new variables, the maximum number of new variables which can be accommodated is $2^m - m - 1$. The total number of variables in such a *saturated* design is therefore $2^m - 1$. That is, in the absence of interaction, we can investigate $2^m - 1$ main effects in 2^m runs.

These designs are used frequently in "screening" situations, where there are a large number of variables present and the experiment is to indicate which factors have the largest main effects.

Resolution IV Designs

Resolution IV designs are characterized by the property that main effects are not confounded with two-factor interactions but are confounded with interactions involving three or more factors. Two-factor interactions are confounded with each other and with higher-order interactions.

These designs are particularly useful when the experimenter is interested primarily in estimating main effects but cannot assume that two-factor interactions are negligible.

The simplest way to construct a resolution IV design in k variables and 2^m runs is first to write down a resolution III design in $k - 1$ variables and 2^{m-1} runs. Denoting this array of pluses and minuses by D, we write a new array

$$D^* = \begin{bmatrix} D \\ -D \end{bmatrix},$$

that is, the lower half of array D is just the "mirror image" of the upper half, in the sense that each sign in the upper half is changed to the *opposite* sign in the lower half. At this point, we have a design in 2^m runs but only $k - 1$ variables. We complete the design by adding the kth column, which consists of 2^{m-1} plus signs followed by 2^{m-1} minus signs.

As an example we shall construct a resolution IV design to accommodate four variables in eight runs. We first construct, as described above, a resolution III design in three variables and four runs:

1	2	3 = 12
−	−	+
+	−	−
−	+	−
+	+	+

We now tack the "mirror image" of this design onto itself to give:

1	2	3
−	−	+
+	−	−
−	+	−
+	+	+
+	+	−
−	+	+
+	−	+
−	−	−

The last column (for variable 4), which consists of four pluses followed by four minuses, is now constructed to give our resolution IV design:

1	2	3	4
−	−	+	+
+	−	−	+
−	+	−	+
+	+	+	+
+	+	−	−
−	+	+	−
+	−	+	−
−	−	−	−

Note that this design turns out to be the same as the design (4.3.1), with the runs rearranged.

4.5 Other Fractions of 2^k Factorials: The Plackett-Burman Designs

The fractional factorials described above are not the only designs which could be selected as subsets of the 2^k designs. Much of the recent literature on two-level designs has, in fact, been devoted to studying the properties of such "irregular" fractions.

One class of "irregular" designs which has been in use for some time is the class of Plackett-Burman designs. These designs are essentially "main effect plans" in that they, like the resolution III designs, provide clear estimates of all main effects as long as all interactions are negligible. The number of runs in any Plackett-Burman design is always a multiple of 4. This is less restrictive than the fractional factorials, which require the number of runs to be a power of 2. Methods for constructing the Plackett-Burman designs can be obtained through the references at the end of these notes.

4.6 Fractions of General Factorial Designs

When selecting fractions of general factorial designs, the same basic principles apply as when selecting fractions of two-level factorials. The goal, as usual, is to derive a fraction with the least undesirable confounding properties.

The mathematics is somewhat more difficult, however, and we shall not go into it here. Methods for choosing fractions of factorials of form p^k, where p is a prime number, are quite well established and involve the use of mathematical operations "modulo p." When the number of levels (p) of each of the k factors is not a prime but is a power of a prime, the problem of choosing a fraction can be solved through the introduction of "pseudofactors." Again, we shall not discuss this approach here. It has been only within the past year or two that an approach has been developed to handle the most general problem, that of selecting fractions of a $p_1 \times p_2 \times \ldots \times p_k$ factorial, where the p's are not necessarily equal.

4.7 Blocking Factorial and Fractional Factorial Designs

We remark in passing that there exist methods for blocking factorial and fractional factorial designs, and these are very closely related to the methods for selecting fractions which we discussed above. Each block is, after all, a fraction in its own right. Historically, in fact, fractional factorials first arose as blocks of the full factorials, long before they were applied on their own.

5. THE MODEL-ORIENTED APPROACH TO EXPERIMENTAL DESIGN

5.1 Notation

In this section, it will be convenient to define a set of "coded" variables $\{x_i\}$ which are related to the controlled variables $\{\xi_i\}$ in an experiment by the following type of linear transformation:

$$x_i = \frac{\xi_i - \xi_{i0}}{S_i}, \qquad i = 1, 2, \ldots, k, \tag{5.1.1}$$

where ξ_{i0} and S_i are arbitrary constants, chosen for convenience.

For example, consider two controlled variables, temperature (T) and concentration (C), of a chemical in an experiment involving a chemical reaction. The design

T	C
200	1.4
400	1.4
200	2.0
400	2.0

(5.1.2)

can be put into "coded" form by writing

$$x_1 = \frac{T - 300}{100}, \qquad x_2 = \frac{C - 1.7}{0.3}$$

to give

x_1	x_2
–1	–1
1	–1
–1	1
1	1

(5.1.3)

which is now in the form of a 2^2 factorial design.

When we discuss a particular type of design, we shall generally write it in terms of the coded variables. It should be understood that the coded design is actually a *representative* of a large class of designs, each member of which can be transformed into the coded design by means of the appropriate coding (5.1.1).

Without loss of generality, we shall also write our *models* in terms of the coded variables. If our original model is a polynomial of degree d in the ξ's, as it often will be, we note that the corresponding model in terms of the x's will also be a polynomial of degree d.

5.2 First-Order Designs

In recent years, much of the work in design theory has departed from the traditional approach of constructing a model to suit a given design and has turned instead to the problem of choosing a design to suit a conjectured model.

The simplest sort of empirical model for the response is one which is linear in the controlled variables. If our k controlled variables are $\xi_1, \xi_2, \ldots, \xi_k$, we write

$$y_u = \beta_0 + \beta_1 x_{1u} + \beta_2 x_{2u} + \dots + \beta_k x_{ku} + \epsilon_u , \qquad u = 1, 2, \dots, n , \tag{5.2.1}$$

where the x's are the coded values of the ξ's. This model is called a first-order linear model, where "first order" refers to the x's and "linear" refers to the β's. When it is understood that the model is linear in the β's, it is called simply a "first-order model."

What sort of designs are suitable for the type of model (5.2.1)? This depends on the criteria for a "good" design, but in this case the generally accepted criterion requires that the design should be chosen so that the variances of the estimated coefficients ($\hat{\beta}$) are minimized. It can be shown that if the "spread" of the design points is fixed, where the "spread" of the design points corresponding to the ith controlled variable is defined as

$$c_i = \sum_{u=1}^{n} (x_{iu} - \bar{x}_i)^2 ,$$

then the n-run design which achieves the minimum variances for all the $\hat{\beta}$'s simultaneously must be "orthogonal," that is,

$$\sum_{u=1}^{n} (x_{iu} - \bar{x}_i)(x_{ju} - \bar{x}_j) = 0 \qquad \text{for all } i \neq j , \quad i, j = 1, 2, \dots, k .$$

Designs possessing this desirable property are, for example, the factorials, fractional factorials, and Plackett-Burman designs discussed above.

If we increase the "spread" of the variables, we find that the variances of the estimated coefficients decrease. The "improvement" achieved in this way is illusory, however, since the model (5.2.1), which may be a good *local* approximation to the response in a small region of the **x** space (space of the coded variables), usually becomes less and less adequate as the "spread" of the x's increases. It is important to remember that the design criteria we shall discuss are often developed under the assumption that the model is good. We must always be careful not to let these criteria lead us to a design for which this assumption no longer holds.

The most economical class of orthogonal first-order designs is the class of *simplex designs*, which can be used to investigate $n - 1$ variables in n runs. We have already encountered some simplex designs, namely, the saturated two-level fractional factorials of resolution III and the Plackett-Burman designs. Geometrically, in the **x** space, the general simplex design consists of n points in $n - 1$ dimensions, arranged so that all the distances between points are equal. For example, in two dimensions (i.e., for two variables x_1 and x_2) the points of a simplex design form an equilateral triangle; in three dimensions, a tetrahedron. Although the simplex designs are efficient in that the number of runs is equal to the number of parameters to be estimated, it should be remembered that, unless they are augmented in some way, they cannot provide a check on the adequacy of the model. For this reason, two-level factorials and unsaturated fractional factorials are often preferred as first-order designs.

5.3 Second-Order Designs

When a first-order model proves inadequate, it is often helpful to consider a second-order model, that is, a model which represents the response as a quadratic function in the x's. In two variables x_1 and x_2, for example, a second-order model would take the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \,. \tag{5.3.1}$$

We mentioned previously that such models are sometimes used to fit data from the class of 3^k factorial designs. Another class of designs, one which was developed specifically for second-order models, is the class of *central composite designs*. A central composite design in k variables is made up of three basic parts:

1. a "cube," which may be a 2^k factorial or 2^{k-p} fractional factorial,
2. $2k$ "star points," which lie on the k axes (two on each axis, both at the same distance from the center but in different directions),
3. "center points," which are repeat points located at the center ($x_1 = 0, x_2 = 0, \ldots, x_k = 0$) of the design.

An example of a central composite design in two variables is as follows:

	x_1	x_2
"cube"	−1	−1
	1	−1
	−1	1
	1	1
"star"	−1.414	0
	1.414	0
	0	−1.414
	0	1.414
"center"	0	0
	0	0
	0	0
	0	0

(5.3.2)

The distance (1.414) of the star points from the center of the design (5.3.2) was chosen so that the variance of the estimated response at any point in the $\mathbf{x}$ space depends only on the distance of that point from the center. [Note: the estimated response at any point $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ is obtained by substituting the estimated coefficients ($\hat{\beta}$) for the β's in the model and dropping the random error term.] This criterion, which requires equality of the variances of the estimated responses on any sphere centered at the origin, is called the criterion of *rotatability*. We shall not go into its justification here, except to say, rather weakly, that the corresponding rotatable designs essentially give equal information about the response in all directions from the origin.

The need to introduce a new criterion, such as rotatability, in connection with second-order designs, when the minimization of the variances of the coefficients served us so well in the first-order case, arises because the latter criterion is no longer applicable. In general, it is not possible to select a second-order design which minimizes the variances of all the $\hat{\beta}$'s *simultaneously*. In an attempt to meet this difficulty a "generalized variance," which is the determinant of the variance-covariance matrix of the $\hat{\beta}$'s, has been suggested as a single overall measure of the variances of the coefficients. Minimization of this generalized variance is equivalent to minimization of the volume of the confidence ellipsoid for the β's, a fact which supports its use as a design criterion.

Another approach in the selection of design criteria has been to consider not the variances of the coefficients but the variance of the estimated response function $\hat{y}(\mathbf{x})$, where $\mathbf{x}$ is a point in the space of the coded variables. Even then, no unique criterion has been accepted. Some which have been suggested are:

1. Rotatability: $V[\hat{y}(\mathbf{x})]$ is a function only of $r = (\mathbf{x}'\mathbf{x})^{1/2}$.
2. Minimize $\max_{\mathbf{x} \in R} V[\hat{y}(\mathbf{x})]$, where R is some "region of interest" in the $\mathbf{x}$ space.
3. Minimize $\int_R w(\mathbf{x})\, V[\hat{y}(\mathbf{x})]\, d\mathbf{x}$, where $w(\mathbf{x})$ is some weight function which indicates the importance of estimating the response well at the point $\mathbf{x}$.

Other criteria have been introduced which consider, in addition to the *variances*, the *bias* which might arise if the model is inadequate.

The existence of so many competing criteria, no one of which is clearly superior, points up the futility of trying to select an "optimum" design for practical use. What is generally required in practice is not a design which will be optimum with respect to a finely defined set of criteria but one which will be "good" with respect to many criteria.

G. E. P. Box and J. S. Hunter have listed some key properties of "good" designs for polynomial models, which should serve as useful guidelines:

1. The design should allow the approximating polynomial of degree d (tentatively assumed to be representationally adequate) to be estimated with satisfactory accuracy within the region of interest.
2. It should allow a check to be made on the representational accuracy of the assumed polynomial.
3. It should not contain an excessively large number of experimental points.
4. It should lend itself to blocking.
5. It should form a nucleus from which a satisfactory design of higher order can be built in case the assumed degree of polynomial proves inadequate.

5.4 Higher-Order Designs and Designs for Nonlinear Models

Just as a quadratic model can be useful when a first-order model is inadequate, so can cubic and even higher-order models be applied if necessary. The number of parameters in these

higher-order models becomes rather large, however, and it is difficult to associate a clear physical meaning with many of them. Often a more "parsimonious" parameterization can be achieved by considering nonlinear models, that is, models which are no longer linear in the β's. A simple example of such a model is a sum of exponentials:

$$y_u = \beta_{11} e^{-\beta_{12} x_{1u}} + \beta_{21} e^{-\beta_{22} x_{2u}} + \epsilon_u , \qquad u = 1, 2, \ldots, n .$$

Until a few years ago, there was very little in the statistical literature concerning designs for such models. Some recent "first steps" are indicated in the references at the end of these notes.

6. RESPONSE SURFACE METHODOLOGY

6.1 Viewing the Response as a Surface

The characteristic concept of response surface methodology is that of considering the "true" response, which is a function of the controlled variables, as a *surface* over the space of those variables.

This idea is most clearly understood when we consider just two (coded) controlled variables x_1 and x_2. The true response at the points of the x_1-x_2 plane is viewed as a surface hovering over that plane. If this response surface is reasonably smooth, we can expect it to be adequately represented by a linear function within a small neighborhood of any given point (x_{10}, x_{20}). Over a slightly larger region a quadratic approximation may be necessary, a larger region may require a cubic, and so forth. This rationale is often employed to justify the use of the polynomial models described above.

6.2 An Optimum-Seeking Method: Steepest Ascent

We shall now briefly illustrate an application of experimental design, in conjunction with response surface concepts, to the problem of finding the point of maximum (or, equivalently, a minimum) response.

Suppose that there are two (coded) controlled variables x_1 and x_2 and that we wish to find the levels of x_1 and x_2 at which the maximum "true" response can be attained. We might first "guess" a point (x_{10}, x_{20}), which we feel is near the region of maximum response, and center a first-order design about that point. This design could well include a few repeat points at the center, not only to provide an estimate of the random error but also as a check on possible curvature in the surface.

When the first-order model has been fitted:

$$\hat{y} = \beta_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 , \tag{6.2.1}$$

we move our next experiments away from the region of the first experiment, along the *direction of steepest ascent,* that is, the direction in which the fitted surface rises most sharply. This direction can be calculated easily from (6.2.1), and we have that any point (x_1, x_2) on the line of steepest ascent satisfies:

$$(x_1 - x_{10}, x_2 - x_{20}) = \lambda(\hat{\beta}_1, \hat{\beta}_2) , \tag{6.2.2}$$

where λ is just an arbitrary positive constant which determines how far the point (x_1, x_2) lies from the center (x_{10}, x_{20}) of the initial experiment. The usual procedure is then to perform several experimental runs at points along the line of steepest ascent (6.2.2), continuing to move along that line until the observed response no longer increases. By this time, we may be quite far away from the original conditions (x_{10}, x_{20}). We now center a new first-order design about the point we have reached, calculate a new direction of steepest ascent leading away from that point, and proceed as before.

Eventually, we shall come to a region for which the linear model is no longer adequate. This may mean that we are near the top of the "hill" we have been climbing. At this point, we can easily augment a first-order design, perhaps with some "star" points, to give a second-order design. When the second-order model is fitted, it should indicate the location of the point of maximum response. A few confirming runs should then determine whether or not the actual maximum has been attained.

In practice, there are other considerations, such as the problem of maintaining other responses within certain specification limits while a single response is being "optimized." Sometimes, in fact, it is never possible to "optimize" – one just tries to do better today than one did yesterday. The principles involved in the steepest ascent approach are valuable, however, and have often been applied with considerable success.

REFERENCES

General

R. A. Fisher, *The Design of Experiments,* Hafner, New York, 1949.

W. G. Cochran and G. M. Cox, *Experimental Designs,* Wiley, New York, 1957.

O. L. Davies (ed.), *Design and Analysis of Industrial Experiments,* Oliver and Boyd, Edinburgh, 1954.

O. Kempthorne, *Design and Analysis of Experiments,* Wiley, New York, 1952.

Fractional Factorial Designs

S. Addelman, "Techniques for Constructing Fractional Replicate Plans," *J. Am. Statist. Assoc.* **58**, 45–71 (1963).

G. E. P. Box and J. S. Hunter, "The 2^{k-p} Fractional Factorial Designs," *Technometrics* **3**, 311–51, 449–58 (1961).

First-Order Designs

G. E. P. Box, "Multi-Factor Designs of First Order," *Biometrika* **39**, 49–57 (1952).
R. L. Plackett and J. P. Burman, "The Design of Multifactorial Experiments," *Biometrika* **33**, 305–25 (1946).

Second-Order and Higher-Order Designs

G. E. P. Box and J. S. Hunter, "Multi-Factor Experimental Designs for Exploring Response Surfaces," *Ann. Math. Statist.* **28**, 195–241 (1957).
G. E. P. Box and N. R. Draper, "A Basis for the Selection of a Response Surface Design," *J. Am. Statist. Assoc.* **54**, 622–54 (1959).

Designs for Non-Linear Models

G. E. P. Box and H. L. Lucas, "Design of Experiments in Non-Linear Situations," *Biometrika* **46**, 77–90 (1959).

Response Surface Methodology

W. J. Hill and W. G. Hunter, "A Review of Response Surface Methodology: A Literature Survey," *Technometrics* **8**, 571–90 (1966).
G. E. P. Box and K. B. Wilson, "On the Experimental Attainment of Optimum Conditions," *J. Roy. Statist. Soc.*, Ser. B **13**, 1–45 (1951).

LECTURE X: ELEMENTS OF STOCHASTIC PROCESSES

V. R. Rao Uppuluri

1. DEFINITIONS AND PRELIMINARIES

Stochastic process: An indexed family of random variables $X(t)$ is called a stochastic process. The parameter t belongs to a set T referred to as the index set.

If $T = \{0, 1, 2, \ldots\}$, then we will say that $X(t)$ is a *discrete-time stochastic* process. When T is discrete we shall write X_n instead of $X(n)$. If $T = (-\infty, \infty)$, then $X(t)$ is called a *continuous-time stochastic process.*

We note that the index set T can at times be an arbitrary set. For example, $X(t)$ could be the height of an ocean wave at a location given by the latitude and longitude and denoted by t.

The space in which the possible values of $X(t)$ lie is called the state space and is denoted by S. If $S = \{0, 1, 2, \ldots\}$, we refer to the stochastic process as a discrete-state process. If $S = (-\infty, \infty)$, then we will call $X(t)$ a real-valued stochastic process. If S is Euclidean k space, then $X(t)$ is said to be a k-vector process.

Generally, the relationships among the random variables $X(t)$, $t \in T$, are specified by giving the joint distribution function of *every finite* family $X(t_1), X(t_2), \ldots, X(t_n)$ of variables of the process.

For the purpose of this lecture, a stochastic process may be considered as well defined once its state space, index parameter, and family of joint distributions are prescribed.

Process with independent increments: Let $T = [0, \infty)$, and let $t_1 < t_2 < \ldots < t_n$ be an arbitrary choice of time points. If the random variables $X(t_2) - X(t_1), X(t_3) - X(t_2), \ldots, X(t_n) - X(t_{n-1})$ are independent, then we say that $X(t)$ is a process with independent increments.

Markov process: Let $T = [0, \infty)$, and let $t_1 < t_2 < \ldots < t_n < t$ be an arbitrary choice of time points. A process is said to be Markovian if

$$P[a < X(t) \leq b \mid X(t_1) = x_1, X(t_2) = x_2, \ldots, X(t_n) = x_n] = P[a < X(t) \leq b \mid X(t_n) = x_n] .$$

In other words, a Markov process is a process with the property that, given the value of $X(t)$, the values of $X(s)$, $s > t$, do not depend on the values of $X(u)$, $u < t$; that is, the probability of any particular future behavior of the process, when the present state is known exactly, is not altered by additional knowledge concerning its past.

The function

$$P(x, s; t, A) = P[X(t) \in A \mid X(s) = x] ,$$

$t > s$ and A an interval of the real line, is called the *transition probability function* and is basic to study the structure of Markov processes.

Stationary processes: A stochastic process $X(t)$ is said to be *strictly stationary* if the joint distributions of the families of random variables $[X(t_1 + h), X(t_2 + h), \ldots, X(t_n + h)]$ and $[X(t_1), \ldots, X(t_n)]$ are the same for all h and arbitrary selections $t_1, t_2, \ldots, t_n$ of T. In particular, the distribution of $X(t)$ is the same for all $t \in T$. This condition asserts that in essence the process is in probabilistic equilibrium and the particular times at which we observe the process are of no relevance.

A stochastic process $X(t)$ is said to be *wide sense stationary* or *covariance stationary* if it possesses finite second moments and if $\mathrm{Cov}[X(t), X(t + h)] = E[X(t)\, X(t + h)] - E[X(t)]\, E[X(t + h)]$ depends only on h for all $t \in T$.

Stationary processes are found to be appropriate models for describing many phenomena that occur in communication theory, astronomy, biology, and economics.

2. EXAMPLES

2.1 Brownian Motion Process (or Wiener Process)

The Brownian motion process is a special type of stochastic process which is the most renowned and historically the first that was thoroughly investigated. As a physical phenomenon the Brownian motion was discovered by the English botanist Brown in 1827. In 1905, a mathematical description of this phenomenon was first derived from the laws of physics by Einstein. The physical theory was further perfected by Smoluchowski, Fokker, Planck, Burger, Furth, Ornstein, Uhlenbeck, Chandrasekhar, Kramers, and others. The first concise mathematical formulation of the theory was given by Wiener in his 1918 dissertation and later papers. In terms of the general framework of stochastic processes, the Brownian motion process is an example of a continuous-time, continuous-state-space Markov process.

In 1827, Brown observed that small particles immersed in a liquid exhibit ceaseless irregular motions. In 1905, Einstein explained this motion by postulating that the particles under observation are subject to perpetual collision with the molecules of the surrounding medium. Let $X(t)$ be the position of the particle at time t, and let $X(0) = x_0$. Let $p(x, t \mid x_0)$ be the conditional probability density of $X(t + t_0)$, given that $X(t_0) = x_0$, so that

$$p(x, t \mid x_0) \geq 0, \quad \int_{-\infty}^{\infty} p(x, t \mid x_0)\, dx = 1 .$$

We postulate that the probability law governing the transitions is stationary in time; that is, $p(x, t \mid x_0)$ does not depend on t_0. Further, we assume that

$$\lim_{t \to 0} p(x, t \mid x_0) = 0 \quad \text{for } x \neq x_0 ,$$

which implies that $X(t + t_0)$ is likely to be near $X(t_0)$ for small t. From physical principles Einstein showed that $p(x, t \mid x_0)$ must satisfy the partial differential equation

$$\frac{\partial p}{\partial t} = D \frac{\partial^2 p}{\partial x^2}, \tag{2.1}$$

which is called the diffusion equation, D being the diffusion coefficient. The evaluation of D is based on the formula $D = 2RT/Nf$, where R is the gas constant, T is the temperature, N is Avogadro's number, and f is the coefficient of friction. By choosing the proper scale we may take $D = \frac{1}{2}$. Then we can show that

$$p(x, t \mid x_0) = (1/\sqrt{2\pi t}) \exp [-(x - x_0)^2/2t] \tag{2.2}$$

is the unique solution of (2.1) with the above boundary conditions. Thus $p(x, t \mid x_0)$ gives the probability density function of $X(t) - X(0)$. The complete Brownian motion process is given by the following definition.

Brownian motion process is a stochastic process $X(t)$, $t \geqq 0$, with the following properties:

1. $X(t)$ is a process with independent increments.
2. Every increment $X(t + s) - X(s)$ is normally distributed with mean 0 and variance ct; $c > 0$ is a fixed constant.

The physical origins of the Brownian motion process suggest that the possible realizations $X(t)$ (sample paths) whose movements result from continuous collisions in the surrounding medium are continuous functions. The proof of this statement requires delicate analysis. The sample paths $X(t)$, although continuous, are very kinky, and their derivatives exist nowhere. This fact is rather deep. Using the property of continuity, we can compute some interesting probabilities associated with the Brownian motion. For instance, it can be shown that

$$P[\max_{0 \leqq u \leqq T} X(u) \geqq a] = \frac{2}{\sqrt{2\pi T}} \int_a^\infty e^{-x^2/2T}\, dx \tag{2.3}$$

under the condition $X(0) = 0$.

If T_a denotes the random time (variable) at which $X(t)$ first attains the value a, where $X(0) = 0$, it can be shown that

$$P(T_a \leqq t) = P[\max_{0 \leqq u \leqq t} X(u) \geqq a \mid X(0) = 0] , \tag{2.4}$$

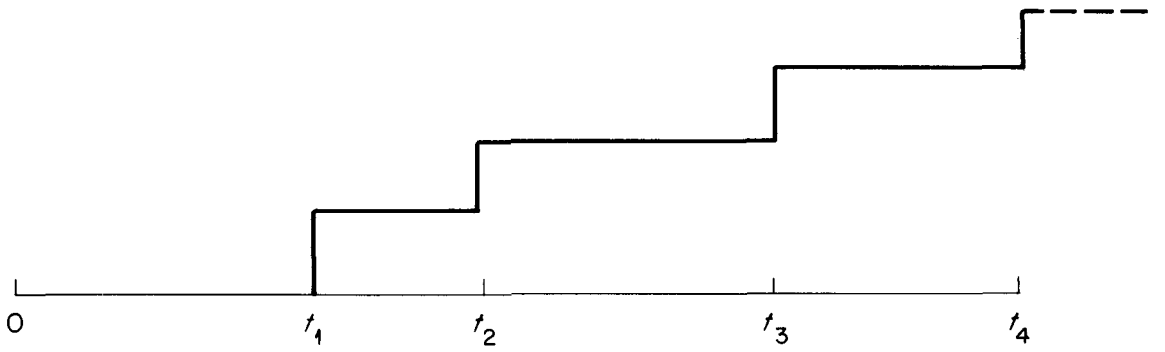
which gives us the probability density function of T_a.

The probability that $X(t)$ has at least one zero in the interval (t_0, t_1), given $X(0) = 0$, can be shown to be

$$\frac{2}{\pi} \arccos \sqrt{t_0/t_1} . \tag{2.5}$$

2.2 Poisson Process

Another important example of a continuous-time $\left(T = [0, \infty)\right)$ stochastic process is the Poisson process. The sample function $X(t)$ counts the number of times a specified event occurs during the period $[0, t)$. Thus each sample path $X(t)$ is represented by a nondecreasing step function.



A sample function of a Poisson process, with X(0) = 0.

Concrete examples of Poisson processes are the following:

1. the number of x rays emitted by a substance undergoing radioactive decay,
2. the number of telephone calls originating in a given locality,
3. the occurrence of accidents at a certain intersection,
4. the occurrence of errors in a page of typing,
5. the breakdowns of a machine.

These examples may be viewed as Poisson processes based on the concept of the law of rare events. We have a situation of several Bernoulli trials with a small probability of success at each trial, where the expected number of successes is constant. Under these conditions it is a familiar theorem that the actual number of events occurring follows a Poisson law. In the case of radioactive decay the Poisson approximation is excellent if the period of observation is very short compared with the half-life of the radioactive substance.

We will now give a set of postulates which lead to the fact that $X(t)$ follows a Poisson distribution, where $X(t)$ denotes the number of events during $[0, t)$.

Postulates

1. The numbers of events $[X(t_2) - X(t_1)]$ and $[X(t_3) - X(t_2)]$ are independent, where $t_1 < t_2 < t_3$.
2. $P[X(t + h) - X(t) \leqq x]$ depends only on h and x, but not on t or on $X(t)$.
3. The probability of at least one event happening in a time period of duration h is

$$p(h) = ah + o(h) \ ,$$

where $o(h)/h$ tends to 0 as $h \to 0$.

4. The probability of two or more events happening in time h is $o(h)$. (This leads to the impossibility of the simultaneous occurrence of two or more events.)

From these postulates, it can be shown that

$$P[X(t) = m] = \frac{(at)^m \, e^{-at}}{m!}, \quad m = 0, 1, 2, \ldots .$$

Often the Poisson process arises in a form where the time parameter is replaced by a suitable spatial parameter, as illustrated by the following formal example. Consider an array of points distributed in a space E. Let N_R denote the number of points contained in the region R of E. We postulate that N_R is a random variable. The collection $\{N_R\}$ of random variables, where R varies over all subsets of E, is said to be a homogeneous Poisson process if the following assumptions are fulfilled:

1. The numbers of points in nonoverlapping regions are independent random variables.
2. For any region R of finite volume, N_R is Poisson distributed with mean $\lambda\, V(R)$, where $V(R)$ is the volume of R. The parameter λ is fixed and measures in a sense the intensity component of the distribution, which is independent of the size or shape.

Spatial Poisson processes arise in considering the distribution of stars or galaxies in space, the spatial distribution of plants and animals, the distribution of bacteria on a slide, etc.

We shall now give some more properties of the Poisson process. Let us recall that from the definition of a continuous-time, discrete-state Markov process, with stationary transition probabilities,

$$p_{ij}(t) = P[X(t+u) = j \mid X(u) = i]$$

will be independent of u, for all $i, j = 0, 1, 2, \ldots$.

The Poisson process is a Markov process on the nonnegative integers which has the following properties:

1. $P[X(t+h) - X(t) = 1 \mid X(t) = x] = ah + o(h), \quad x = 0, 1, 2, \ldots,$
2. $P[X(t+h) - X(t) = 0 \mid X(t) = x] = 1 - ah + o(h),$
3. $X(0) = 0.$

At times the Poisson process is referred to as the completely random process, as it distributes points "at random" over the interval $[0, \infty)$ in much the same way that the uniform distribution distributes points over a finite interval. In particular, the probability of an observation falling in a subinterval is a function of its length only, and the numbers of events occurring in two disjoint time intervals are independent random variables.

Another illustrative example of the Poisson process is that of fishing. Let $X(t)$ denote the number of fish caught in the time interval $[0, t)$. Suppose that the number of fish available is very large, that the enthusiast stands no better chance of catching fish than the rest of us, and that as many fish are likely to nibble at one instant of time as at another. Under these "ideal" conditions, the process $\{X(t), t \geqq 0\}$ may be considered as a Poisson process. This example serves to point up the Markov property (the chance of catching a fish does not depend upon the number caught) and the "no premium for waiting" property, which is the most distinctive property possessed by the Poisson process. It means that the fisherman who has just arrived at the pier has as good a chance of catching a fish in the next instant of time as he who has been waiting for a bite for a few hours without success.

Another example is afforded by problems arising in the theory of counters. If $X(t)$ is the number of radioactive disintegrations detected by a Geiger counter in the interval $[0, t)$, the process is Poisson as long as the half-life of the substance is large relative to t. This provision ensures that the chance for a disintegration per unit of time may be considered as constant over time.

Let T_k denote the time (random) between the kth and $(k+1)$st event. The random variables T_k are called the "waiting times" between successive events, and

$$S_k = T_0 + T_1 + \ldots + T_{k-1}$$

denotes the time at which the kth event occurs.

In the case of a Poisson process, we have the following:

Theorem: The waiting times T_k are independent and identically distributed random variables, following an exponential distribution with parameter a.

3. MARKOV CHAINS

3.1 Definitions and Preliminaries

A *discrete-time Markov chain* $\{X_n\}$ is a Markov stochastic process whose state space is a countable or finite set and for which the index set T is $(0, 1, 2, \ldots)$. It is convenient to label the state space of the process by $\{0, 1, 2, \ldots\}$, and it is customary to speak of X_n being in state i if $X_n = i$.

The probability of X_{n+1} being in state j, given that X_n is in state i (called a one-step transition probability), is denoted by

$$p_{ij}^{n,n+1} = P[X_{n+1} = j \mid X_n = i] .$$

When the one-step transition probabilities are independent of the time variable (i.e., independent of the value n), we say that the Markov chain has stationary transition probabilities. In this case $p_{ij}^{n,n+1} = p_{ij}$, and p_{ij} is the probability that the state value undergoes a transition from i to j in one trial. It is customary to arrange the elements p_{ij} as a matrix and to refer to the matrix

$$P = (p_{ij},\ 0 \leqq i,\ j \leqq \infty)$$

as the transition probability matrix of the Markov chain; we note that all the elements of this matrix, p_{ij}, are nonnegative and that the sum of the elements in any ith row,

$$\sum_{j=0}^{\infty} p_{ij} ,$$

is equal to unity.

The Markov chain is completely determined when once the transition probability matrix P and the value of X_0 are specified. We note that

$$P(X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) = p_{i_o} \cdot p_{i_o, i_1} \cdot p_{i_1, i_2} \cdot \ldots \cdot p_{i_{n-1}, i_n} . \quad (3.1.1)$$

3.2 One-Dimensional Random Walks

A one-dimensional random walk is a Markov chain whose state space is a finite or infinite subset $(a, a+1, \ldots, b)$ of the integers, in which the particle (state of the system), if it is in state i, can in a single transition either stay in i or move to one of the adjacent states $i-1$, $i+1$. If the state space is taken as the nonnegative integers, the transition matrix of a random walk has the form

$$P = \begin{pmatrix} r_0 & p_0 & 0 & 0 & \cdots \\ q_1 & r_1 & p_1 & 0 & \cdots \\ 0 & q_2 & r_2 & p_2 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \end{pmatrix}$$

where

$$q_i + r_i + p_i = 1,\ i = 1, 2, \ldots,\ r_0 + p_0 = 1 .$$

The fortune of a player engaged in a series of contests is often depicted by a random walk process. Specifically, suppose an individual (player A) with fortune k plays a game against an infinitely rich adversary and has probability p_k of winning one unit and probability $q_k = 1 - p_k$ $(k \geqq 1)$ of losing one unit in each contest, and $r_0 = 1$. The process $\{X_n\}$, where X_n represents his fortune after n contests, is a random walk. Note that once state 0 is reached (i.e., the player A is wiped out), the process remains in that state. This process is also commonly known as the "gambler's ruin."

If $p_0 = 0$ and $r_0 = 1$, then 0 acts as an *absorbing barrier*. Once the particle reaches zero it remains there forever. If $p_0 > 0$ and $r_0 > 0$, then 0 is a *partially reflecting barrier*.

A classical mathematical model of diffusion through a membrane is the famous Ehrenfest model, namely, a random walk on a finite set of states where the boundary states are reflecting. The random walk is restricted to the states $i = -a, -a+1, \ldots, -1, 0, 1, \ldots, a$ with transition probability matrix

$$p_{ij} = \begin{cases} \dfrac{a-i}{2a} & \text{if } j = i+1 \\ \dfrac{a+i}{2a} & \text{if } j = i-1 \\ 0 & \text{otherwise} \end{cases}$$

The physical interpretation of this model is as follows: Imagine two containers containing a total of $2a$ balls. Suppose the first container, labeled A, holds k balls and the second container B holds $2a - k$ balls. A ball is selected at random (all selections are equally likely) from among the totality of the $2a$ balls and moved to the other container. Each selection generates a transition of the process. Let the state of the system be determined by the number of balls in A.

3.3 A Discrete Queueing Markov Chain

Customers arrive for service and take their place in a waiting line. During each period of time a single customer is served, provided that at least one customer is present. If no customer awaits service, then during this period no service is performed. During a service period new customers may arrive. We suppose that the actual number of arrivals in the nth period is a random variable ξ_n whose distribution function is independent of the period and is given by

$$P(k \text{ customers arrive in a service period}) = P(\xi_n = k) = a_k, \quad k = 0, 1, 2, \ldots, \quad a_k \geqq 0, \ \sum a_k = 1 .$$

We also assume the random variables ξ_n are independent. The state of the system at the start of each period is defined to be the number of customers waiting in line for service. If the present state is i, then after a lapse of one period the state is

$$j = \begin{cases} i - 1 + \xi & \text{if } i \geqq 1 \\ \xi & \text{if } i = 0 , \end{cases}$$

where ξ is the number of new customers that arrived in this period while a single customer was serviced. In terms of the random variables of the process, we can express this as

$$X_{n+1} = \max(X_n - 1, 0) + \xi_n .$$

The transition probability matrix is given by

$$\begin{bmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & \cdots \\ a_0 & a_1 & a_2 & a_3 & a_4 & \cdots \\ 0 & a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & 0 & a_0 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix} .$$

It is intuitively clear that if the expected number of new customers,

$$\sum_{k=0}^{\infty} k a_k ,$$

that arrive during a service period exceeds 1, then certainly with the passage of the time the length of the waiting line increases without limit. On the other hand, if $\sum k a_k < 1$, then we shall see that the length of the waiting line approaches an equilibrium. If $\sum k a_k = 1$, a situation of gross instability develops.

3.4 Success Runs

Consider a Markov chain on the nonnegative integers with transition probability matrix of the form

$$P = \begin{bmatrix} p_0 & q_0 & 0 & 0 & \cdots \\ p_1 & 0 & q_1 & 0 & \cdots \\ p_2 & 0 & 0 & q_2 & \cdots \\ p_3 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix},$$

where $q_i > 0$, $p_i > 0$, and $q_i + p_i = 1$, $i = 0, 1, 2, \ldots$. The zero state plays a distinguished role in that it can be reached in one transition from any other state, while state $i + 1$ can be reached only from state i.

A special case of this transition matrix arises when one is dealing with success runs resulting from repeated trials, each of which admits two possible outcomes, success (S) or failure (F). More explicitly, consider a sequence of trials with two possible outcomes, S or F. Moreover, suppose that in each trial, the probability of S is α and the probability of F is $\beta = 1 - \alpha$. We say a success run of length r happened at trial n if the outcomes in the preceding $r + 1$ trials, including the present trial as the last, were, respectively, F, S, S, ..., S. Let us now label the present state of the process by the length of the success run currently under way. In particular, if the last trial resulted in a failure, then the state is zero. Similarly, when the preceding $r + 1$ trials in order had the outcomes F, S, ..., S, the state variable would carry the label r. The process is clearly Markovian, and the transition matrix has the above form with

$$p_n = \beta, \quad n = 0, 1, 2 \ldots .$$

3.5 Branching Processes

Suppose an organism at the end of its lifetime produces a random number ξ of offspring with probability distribution

$$P(\xi = k) = a_k, \quad k = 0, 1, 2, \ldots,$$

where $a_k \geqq 0$ and $\Sigma a_k = 1$. We assume that all offspring act independently of each other and at the end of their lifetime (for simplicity, the life-spans of all organisms are assumed to be the same) individually have progeny in accordance with the same probability distribution, thus propagating their species. The process $\{X_n\}$, where X_n is the population size at the nth generation, is a Markov chain. The transition matrix is given by

$$p_{ij} = P(X_{n+1} = j \mid X_n = i) = P(\xi_1 + \ldots + \xi_i = j) ,$$

where the ξ's are independent identically distributed random variables.

3.6 A Genetic Model

The following idealized genetics model was introduced by S. Wright to investigate the fluctuation of gene frequency under the influence of mutation and selection. We begin by describing a so-called simple haploid model of random reproduction, disregarding mutation pressures and selective forces. We assume that we are dealing with a fixed population of $2N$ genes composed of type a and type A individuals. The makeup of the next generation is determined by $2N$ independent binomial trials as follows: If the parent population consists of j genes of type a and $2N - j$ genes of type A, then each trial results in a or A with probabilities $p_j = j/2N$ and $q_j = 1 - (j/2N)$ respectively. Repeated selections are done with replacement. By this procedure we generate a Markov chain $\{X_n\}$, where X_n is the number of a genes in the nth generation among a constant population size of $2N$ elements. The state space contains the $2N + 1$ values $(0, 1, 2, \ldots, 2N)$. The transition probability matrix is computed according to the binomial distribution as

$$P(X_{n+1} = k \mid X_n = j) = p_{jk} = \binom{2N}{k} p_j^k q_j^{2N-k}, \quad k, j = 0, 1, \ldots, 2N .$$

4. PROPERTIES OF MARKOV CHAINS

4.1 n-Step Probabilities

A Markov chain is completely defined by its one-step transition probability matrix and the specification of a probability distribution on the state of the process at time 0. Let p_{ij}^n denote the probability that the process goes from state i to state j in n transitions, that is,

$$p_{ij}^n = P(X_{n+m} = j \mid X_m = i) .$$

It can be shown that

$$p_{ij}^n = \sum_{k=0}^{\infty} p_{ik}^r p_{kj}^s, \quad \text{where } r + s = n .$$

If the probability of the process initially being in state j is p_j, that is, $P(X_0 = j) = p_j$, then the probability of the process being in state k at time n is

$$p_k^n = \sum_{j=0}^{\infty} p_j p_{jk}^n = P(X_n = k) .$$

In order to study the limiting behavior of p_{ij}^n as $n \to \infty$, we need to introduce some principles of classifying the states of a Markov chain.

4.2 Classification of States

A state j is said to be *accessible* from state i if, for some integer $n \geqq 0$, $p_{ij}^n > 0$. Two states i and j, each accessible to the other, are said to *communicate*. A Markov chain is said to be *irreducible* if all states communicate with each other.

We define the *period of state i*, written $d(i)$, to be the greatest common divisor of all integers $n \geqq 1$ for which $p_{ii}^n > 0$. (If $p_{ii}^n = 0$ for all $n \geqq 1$ define $d(i) = 0$.) A Markov chain in which each state has period unity is called *aperiodic*.

Let f_{jk}^n stand for the probability that, in a process starting from state j, the first entry to state k occurs at the nth step. Let us put

$$f_{jk}^{(0)} = 0, \quad f_{jk} = \sum_{n=1}^{\infty} f_{jk}^n, \quad \mu_j = \sum_{n=1}^{\infty} n f_{jj}^n .$$

Clearly f_{jk} is the probability that, starting from state j, the system will ever pass through state k.

The state j is said to be *persistent* if $f_{jj} = 1$ and *transient* if $f_{jj} < 1$. A persistent state j is called a *null state* if its mean recurrence time $\mu_j = \infty$.

An aperiodic persistent state j with $\mu_j < \infty$ is called *ergodic*.

4.3 Limit Theorems

Theorem:

1. The state j is transient if and only if

$$\sum_{n=0}^{\infty} p_{jj}^n < \infty .$$

In this case

$$\sum_{n=0}^{\infty} p_{ij}^n < \infty$$

for all i.

2. The (persistent) state j is null if and only if

$$\sum_{n=0}^{\infty} p_{jj}^n = \infty$$

but $p_{jj}^n \to 0$ as $n \to \infty$. In this case $p_{ij}^n \to 0$ for all i.

3. An aperiodic state j (persistent) is ergodic if and only if $\mu_j < \infty$. In this case as $n \to \infty$, $p_{ij}^n \to f_{ij}/\mu_j$.

A Markov *chain* is said to be *ergodic* if all its states are aperiodic and persistent with finite mean recurrence times. We have the following important theorem.

Theorem: In an irreducible ergodic chain the limits

$$\lim_{n\to\infty} p_{jk}^n = u_k$$

exist and are independent of the initial state j. Furthermore,

$$u_k > 0, \quad \sum u_k = 1, \quad u_j = \sum_i u_i p_{ij} ,$$

for all j.

Conversely, suppose that the chain is irreducible and aperiodic and that there exist numbers $u_k \geqq 0$ satisfying

$$\sum u_k = 1, \quad u_j = \sum_i u_i p_{ij} .$$

Then the chain is ergodic, the u_k are given by

$$u_k = \lim_{n\to\infty} p_{jk}^n ,$$

and $u_k = 1/\mu_k$, where μ_k is the mean recurrence time of state k.

ACKNOWLEDGMENT

The main source of this lecture has been the book *A First Course in Stochastic Processes*, by S. Karlin, Academic Press, 1966. The following additional references may also be of great value.

REFERENCES

Bartlett, M. S., *An Introduction to Stochastic Processes with Special Reference to Methods and Applications*, Cambridge Univ. Press, London and New York, 1955.

Bartlett, M. S., *Stochastic Population Models in Ecology and Epidemiology*, Wiley, New York, 1960.

Bharucha-Reid, A. T., *Elements of the Theory of Markov Processes and Their Applications*, McGraw-Hill, New York, 1960.

Bailey, N. T. J., *The Elements of Stochastic Processes with Application to the Natural Sciences*, Wiley, New York, 1964.

Chung, K. L., *Markov Chains with Stationary Transition Probabilities*, Springer-Verlag, 1960.

Doob, J. L., *Stochastic Processes*, Wiley, New York, 1953.

Dynkin, E. B., *Theory of Markov Processes*, Academic, New York, 1965.

Feller, W., *An Introduction to Probability Theory and Its Applications*, vol. I, 3d ed., Wiley, New York, 1966.

Feller, W., *An Introduction to Probability Theory and Its Applications*, vol. II, Wiley, New York, 1968.

Gnedenko, B. V., *Theory of Probability*, Chelsea, New York, 1962.

Harris, T. E., *The Theory of Branching Processes*, Springer-Verlag, Berlin, 1963.

Ito, K., and H. P. McKean, *Diffusion Processes and Their Sample Paths*, Springer-Verlag, Berlin, 1965.

Kac, M., *Probability and Related Topics in Physical Sciences*, Wiley, New York, 1959.

Kemeny, J. G., and J. L. Snell, *Finite Markov Chains*, Van Nostrand, Princeton, N.J., 1960.

Kemeny, J. G., J. L. Snell, and A. W. Knapp, *Denumerable Markov Chains*, Van Nostrand, Princeton, N.J., 1966.

Lévy, P., *Processus Stochastiques et Mouvement Brownien*, Gauther-Villars, Paris, 1948.

Parzen, E., *Stochastic Processes*, Holden-Day, San Francisco, 1962.

Prabhu, N. U., *Stochastic Processes*, Macmillan, New York, 1965.

Rosenblatt, M., *Random Processes*, Oxford Univ. Press, London and New York, 1962.

Skorokhod, A. V., *Studies in the Theory of Random Processes,* Addison-Wesley, Reading, 1965.

Spitzer, F., *Principles of Random Walk,* Van Nostrand, Princeton, N.J., 1964.

Takacs, L., *Stochastic Processes,* Methuen's Monographs, 1960.

Yaglom, A. M., *Stationary Random Functions,* Prentice-Hall, Englewood Cliffs, N.J., 1962.

ORNL-4347
UC-32 – Mathematics and Computers

INTERNAL DISTRIBUTION

1. Biology Library
2–4. Central Research Library
5–6. ORNL – Y-12 Technical Library
Document Reference Section
7–26. Laboratory Records Department
27. Laboratory Records, ORNL R.C.
28. T. D. Anderson
29. T. W. Armstrong
30. M. R. Arnett
31. Russell Baldock
32. S. J. Ball
33. C. D. Baumann
34–43. J. J. Beauchamp
44. R. G. Berggren
45. H. V. Bertini
46. J. A. Biggerstaff
47. F. T. Binford
48. E. G. Bohlmann
49–50. K. O. Bowman
51. R. L. Bradshaw
52. H. R. Brashear
53. J. C. Bresee
54. W. H. Bridges
55. A. A. Brooks
56. G. M. Brown
57. M. D. Brown
58. W. E. Brundage
59. J. B. Bullock
60. C. T. Butler
61. R. S. Carlsmith
62. G. T. Chapman
63. R. H. Chapman
64. R. O. Chester
65. W. H. Christie
66. C. J. Claffey
67. H. O. Cohn
68. D. A. Costanzo
69. G. E. Creek
70. G. A. Cristy
71. Arline Culkowski
72. J. W. Curlin
73. M. D. Danford
74. F. J. Davis
75. W. Davis, Jr.
76. J. G. Delene
77. L. C. Emerson
78. J. I. Federer
79. M. Fontana
80. R. E. Funderlic
81. C. H. Gabbard
82. W. R. Gambill
83. D. A. Gardiner
84. D. Gosslee
85. R. L. Hahn
86. M. L. Halbert
87. J. Halperin
88. C. E. Hammons
89. F. F. Haywood
90. T. L. Hebble
91. R. W. Hendricks
92. P. G. Herndon
93. N. W. Hill
94. N. E. Hinkle
95. D. G. Hoel
96. H. F. Holmes
97. A. S. Householder
98. H. H. Hubbell
99. Norma C. Hull
100. H. B. Hupf
101. F. C. Hutton
102. C. K. Johnson
103. G. F. R. Johnson
104. P. C. Johnson
105. A. R. Jones
106. W. H. Jordan
107. H. W. Joy
108–117. M. A. Kastenbaum
118. S. V. Kaye
119. R. J. Kedl
120. B. H. Ketelle
121. E. M. Kidd
122. J. R. Knight
123. D. J. Knowles
124. W. W. Knowles
125. K. A. Kraus
126. T. S. Kress
127. R. A. Kuebbing
128. B. Kuperstock

129. C. E. Larson
130. R. J. Lauer
131. M. E. LaVerne
132. Elmon Leach
133. Claudia S. Lever
134. W. E. Lever
135. M. H. Lietzke
136. D. B. Lloyd
137. R. A. Lorenz
138. M. I. Lundin
139. R. N. Lyon
140. H. G. MacPherson
141. H. A. Mahlman
142. F. C. Maienschein
143. A. P. Malinauskas
144. J. A. Martin
145. W. J. Martin
146. H. J. Metz
147. F. L. Miller, Jr.
148. R. V. Miskell
149. T. J. Mitchell
150. J. P. Moore
151. S. E. Moore
152. M. T. Morgan
153. Erlie McDaniel
154. F. K. McGowan
155. V. A. McKay
156. J. R. McWherter
157. A. H. Narten
158. C. W. Nestor, Jr.
159–163. J. S. Olson
164. R. B. Parker
165. L. F. Parsly
166. H. O. Phillips
167. L. R. Phillips
168. P. H. Pitkanen
169. F. Plasil
170. J. K. Poggenburg
171. F. A. Posey
172. G. L. Ragan
173. J. W. Reynolds
174. Paul Rubel
175. R. M. Rush
176. A. C. Rutenburg
177. L. M. Scott
178. P. E. Scott
179. S. T. Sekula
180. C. E. Sessions
181. M. J. Skinner
182. C. J. Sparks
183. W. C. Stoddart
184. W. J. Stelzman
185. J. O. Stiegler
186. J. A. Stockdale
187. R. S. Stone
188. E. A. Straker
189. J. G. Sullivan
190. D. A. Sundberg
191. J. R. Tallackson
192. J. R. Tarrant
193. W. A. Thomas
194. W. E. Unger
195. V. R. Rao Uppuluri
196. R. S. Valachovic
197. R. A. Vandermeer
198. J. L. Wantland
199. G. M. Watson
200. A. M. Weinberg
201. A. J. Weinberger
202. M. E. Whatley
203. Robin Williams
204. C. S. Williams
205. F. R. Winslow
206. E. C. Witt
207. H. A. Wright
208. G. T. Yahr
209. H. L. Yakel
210. Henry Zeldes
211. N. F. Ziegler
212. Walter Zobel

EXTERNAL DISTRIBUTION

213. J. R. Totter, U.S. Atomic Energy Commission, Washington, D.C.
214. J. S. Kirby-Smith, U.S. Atomic Energy Commission, Washington, D.C.
215. C. W. Edington, U.S. Atomic Energy Commission, Washington, D.C.
216. R. W. Touchberry, U.S. Atomic Energy Commission, Washington, D.C.
217. C. V. L. Smith, U.S. Atomic Energy Commission, Washington, D.C.
218. Harry Polachek, U.S. Atomic Energy Commission, Washington, D.C.
219. G. B. Darling, Director, Atomic Bomb Casualty Commission, Hiroshima, Japan
220. G. W. Beebe, Follow-Up Agency, National Academy of Sciences, National Research Council, Washington, D.C.
221. J. J. Harley, Director, Health and Safety Laboratory, U.S. Atomic Energy Commission, New York, N.Y.
222. J. L. Jaech, Battelle-Northwest, Richland, Washington
223. J. M. Cameron, National Bureau of Standards, Washington, D.C.
224. R. D. Sanders, Statistics Department, University of Tennessee, Knoxville, Tennessee
225. J. A. Swartout, Union Carbide Corporation, New York, N.Y.
226. Documents Room, Computing Center, University of Notre Dame, Notre Dame, Indiana
227. Laboratory and University Division, AEC, ORO

228–479. Given distribution as shown in TID-4500 under Mathematics and Computers category (25 copies – CFSTI)