UNCLASSIFIED

AD NUMBER

AD865424

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; Dec 1969. Other requests shall be referred to Army Materiel Command, Attn: AMCRD-TV, Washington, DC 20315.

AUTHORITY

US Army Materiel Command ltr dtd 14 Jan 1972

THIS PAGE IS UNCLASSIFIED

STATEMENT #2 UNCLASSIFIED
This document is subject to special export controls and each transmittal to foreign governments or foreign nationals may be made only with prior approval of: Army Materiel Command, Attn AMCRD-TV, Washington, D.C. 20315

ENGINEERING DESIGN HANDBOOK

EXPERIMENTAL STATISTICS

SECTION 4

SPECIAL TOPICS

HEADQUARTERS, U.S. ARMY MATERIEL COMMAND ECEMBER 1969

HEADQUARTERS
UNITED STATES ARMY MATERIEL COMMAND
WASHINGTON, D.C. 20315

AMC PAMPHLET
No. 706-113*

16 December 1969

ENGINEERING DESIGN H[illegible]OOK

EXPERIMENTAL STATISTICS (SEC 4)

Paragraph *Page*

CHAPTER 15

SOME SHORTCUT TESTS FOR SMALL SAMPLES FROM NORMAL POPULATIONS

*This pamphlet supersedes AMCP 706-113, 10 March 1966.

Paragraph *Page*

CHAPTER 16

SOME TESTS WHICH ARE INDEPENDENT OF THE FORM OF THE DISTRIBUTION

TABLE OF CONTENTS

Paragraph *Page*

CHAPTER 17

THE TREATMENT OF OUTLIERS

CHAPTER 18

THE PLACE OF CONTROL CHARTS IN EXPERIMENTAL WORK

Paragraph *Page*

CHAPTER 19

STATISTICAL TECHNIQUES FOR ANALYZING EXTREME-VALUE DATA

CHAPTER 20

THE USE OF TRANSFORMATIONS

TABLE OF CONTENTS

Paragraph *Page*

CHAPTER 23

EXPRESSION OF THE UNCERTAINTIES OF FINAL RESULTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

FOREWORD

INTRODUCTION

This is one of a group of handbooks covering the engineering information and quantitative data needed in the design, development, construction, and test of military equipment which (as a group) constitute the Army Materiel Command Engineering Design Handbook.

PURPOSE OF HANDBOOK

The Handbook on Experimental Statistics has been prepared as an aid to scientists and engineers engaged in Army research and development programs, and especially as a guide and ready reference for military and civilian personnel who have responsibility for the planning and interpretation of experiments and tests relating to the performance of Army equipment in the design and developmental stages of production.

SCOPE AND USE OF HANDBOOK

This Handbook is a collection of statistical procedures and tables. It is presented in five sections, viz:

AMCP 706-110, Section 1, Basic Concepts and Analysis of Measurement Data (Chapters 1-6)

AMCP 706-111, Section 2, Analysis of Enumerative and Classificatory Data (Chapters 7-10)

AMCP 706-112, Section 3, Planning and Analysis of Comparative Experiments (Chapters 11-14)

AMCP 706-113, Section 4, Special Topics (Chapters 15-23)

AMCP 706-114, Section 5, Tables

Section 1 provides an elementary introduction to basic statistical concepts and furnishes full details on standard statistical techniques for the analysis and interpretation of measurement data. Section 2 provides detailed procedures for the analysis and interpretation of enumerative and classificatory data. Section 3 has to do with the planning and analysis of comparative experiments. Section 4 is devoted to consideration and exemplification of a number of important but as yet non-standard statistical techniques, and to discussion of various other special topics. An index for the material in all four sections is placed at the end of Section 4. Section 5 contains all the mathematical tables needed for application of the procedures given in Sections 1 through 4.

An understanding of a few basic statistical concepts, as given in Chapter 1, is necesssary; otherwise each of the first four sections is largely independent of the others. Each procedure, test, and technique described is illustrated by means of a worked example. A list of authoritative references is included, where appropriate, at the end of each chapter. Step-by-step instructions are given for attaining a stated goal, and the conditions under which a particular procedure is strictly valid are stated explicitly. An attempt is made to indicate the extent to which results obtained by a given procedure are valid to a good approximation when these conditions are not fully met. Alternative procedures are given for handling cases where the more standard procedures cannot be trusted to yield reliable results.

The Handbook is intended for the user with an engineering background who, although he has an occasional need for statistical techniques, does not have the time or inclination to become an expert on statistical theory and methodology.

The Handbook has been written with three types of users in mind. The first is the person who has had a course or two in statistics, and who may even have had some practical experience in applying statistical methods in the past, but who does not have statistical ideas and techniques at his fingertips. For him, the Handbook will provide a ready reference source of once familiar ideas and techniques. The second is the person who feels, or has been advised, that some particular problem can be solved by means of fairly simple statistical techniques, and is in need of a book that will enable him to obtain the solution to his problem with a minimum of outside assistance. The Handbook should enable such a person to become familiar with the statistical ideas, and reasonably adept at the techniques, that are most fruitful in his particular line of research and development work. Finally, there is the individual who, as the head of, or as a member of a service group, has responsibility for analyzing and interpreting experimental and test data brought in by scientists and engineers engaged in Army research and development work. This individual needs a ready source of model work sheets and worked examples corresponding to the more common applications of statistics, to free him from the need of translating textbook discussions into step-by-step procedures that can be followed by individuals having little or no previous experience with statistical methods.

It is with this last need in mind that some of the procedures included in the Handbook have been explained and illustrated in detail twice: once for the case where the important question is whether the performance of a new material, product, or process exceeds an established standard; and again for the case where the important question is whether its performance is not up to the specified standards. Small but serious errors are often made in changing "greater than" procedures into "less than" procedures.

AUTHORSHIP AND ACKNOWLEDGMENTS

The Handbook on Experimental Statistics was prepared in the Statistical Engineering Laboratory, National Bureau of Standards, under a contract with the Department of Army. The project was under the general guidance of Churchill Eisenhart, Chief, Statistical Engineering Laboratory.

Most of the present text is by Mary G. Natrella, who had overall responsibility for the completion of the final version of the Handbook. The original plans for coverage, a first draft of the text, and some original tables were prepared by Paul N. Somerville. Chapter 6 is by Joseph M. Cameron; most of Chapter 1 and all of Chapters 20 and 23 are by Churchill Eisenhart; and Chapter 10 is based on a nearly-final draft by Mary L. Epling.

Other members of the staff of the Statistical Engineering Laboratory have aided in various ways through the years, and the assistance of all who helped is gratefully acknowledged. Particular mention should be made of Norman C. Severo, for assistance with Section 2, and of Shirley Young Lehman for help in the collection and computation of examples.

Editorial assistance and art preparation were provided by John I. Thompson & Company, Washington, D. C. Final preparation and arrangement for publication of the Handbook were performed by the Engineering Handbook Office, Duke University.

Appreciation is expressed for the generous cooperation of publishers and authors in granting permission for the use of their source material. References for tables and other material, taken wholly or in part, from published works, are given on the respective first pages.

Elements of the U. S. Army Materiel Command having need for handbooks may submit requisitions or official requests directly to the Publications and Reproduction Agency, Letterkenny Army Depot, Chambersburg, Pennsylvania 17201. Contractors should submit such requisitions or requests to their contracting officers.

Comments and suggestions on this handbook are welcome and should be addressed to Army Research Office-Durham, Box CM, Duke Station, Durham, North Carolina 27706.

PREFACE

This listing is a guide to the Section and Chapter subject coverage in all Sections of the Handbook on Experimental Statistics.

Chapter No. *Title*

AMCP 706-110 (SECTION 1) — BASIC STATISTICAL CONCEPTS AND STANDARD TECHNIQUES FOR ANALYSIS AND INTERPRETATION OF MEASUREMENT DATA

AMCP 706-111 (SECTION 2) — ANALYSIS OF ENUMERATIVE AND CLASSIFICATORY DATA

AMCP 706-112 (SECTION 3) — THE PLANNING AND ANALYSIS OF COMPARATIVE EXPERIMENTS

AMCP 706-113 (SECTION 4) — SPECIAL TOPICS

AMCP 706-114 (SECTION 5) — TABLES

DISCUSSION OF TECHNIQUES IN CHAPTERS 15 THROUGH 23

In this Section, a number of important but as yet non-standard techniques are presented for answering questions similar to those considered in AMCP 706-110, Section 1. In addition, various special topics, such as transformation of data to simplify the statistical analysis, treatment of outlying observations, expression of uncertainties of final results, use of control charts in experimental work, etc., are discussed in sufficient detail to serve as an introduction for the reader who wishes to pursue these topics further in the published literature.

All A-Tables referenced in these Chapters are contained in AMCP 706-114, Section 5.

CHAPTER 15

SOME SHORTCUT TESTS FOR SMALL SAMPLES FROM NORMAL POPULATIONS

15-1 GENERAL

Shortcut tests are characterized by their nplicity. The calculations are simple, and often may be done on a slide rule. Further, they are easily learned. An additional advantage in their use is that their simplicity implies fewer errors, and this may be important where time spent in checking is costly.

The main disadvantage of the shortcut tests as compared to the tests given in AMCP 706-110, Chapters 3 and 4, is that with the same values of α and n, the shortcut test will, in general, have a larger β, — i.e., it will result in a higher proportion of *errors of the second kind*. For the tests given in this chapter, this increase in error will usually be rather small if the sample sizes involved are each of the order of 10 or less.

Unlike the *nonparametric* tests of Chapter 16, these tests require the assumption of *normality* of the underlying populations. Small departures from normality, however, will usually have a negligible effect on the test — i.e., the values of α and β, in general, will differ from their intended values by only a slight amount.

No descriptions of the operating characteristics of the tests or of methods of determining sample size are given in this chapter.

15-2 COMPARING THE AVERAGE OF A NEW PRODUCT WITH THAT OF A STANDARD

15-2.1 DOES THE AVERAGE OF THE NEW PRODUCT DIFFER FROM THE STANDARD?

Data Sample 15-2.1 — Depth of Penetration

Ten rounds of a new type of shell are fired into a target, and the depth of penetration is measured for each round. The depths of penetration are:

10.0, 9.8, 10.2, 10.5, 11.4, 10.8, 9.8, 12.2, 11.6, 9.9 cms.

The average penetration depth, m_0, of the standard comparable shell is 10.0 cm.

The question to be answered is: Does the new type differ from the standard type with respect to average penetration depth (either a decrease, or an increase, being of interest)?

Procedure	Example
(1) Choose α, the significance level of the test.	(1) Let $\alpha = .01$
(2) Look up $\varphi_{1-\alpha/2}$ in Table A-12 for the appropriate n.	(2) $n = 10$ $\varphi_{.995} = 0.333$
(3) Compute $\bar{X}$, the mean of the n observations.	(3) $\bar{X} = 10.62$
(4) Compute w, the difference between the largest and smallest of the n observations.	(4) $w = 2.4$
(5) Compute $\varphi = (\bar{X} - m_0)/w$	(5) $\varphi = \frac{10.62 - 10.00}{2.4}$ $= 0.258$
(6) If $\lvert\varphi\rvert > \varphi_{1-\alpha/2}$, conclude that the average performance of the new product differs from that of the standard; otherwise, there is no reason to believe that they differ.	(6) Since 0.258 is not larger than 0.333, there is no reason to believe that the new type shell differs from the standard.

15-2.2 DOES [illegible] AVERAGE OF THE NEW PRODUCT EXCEED THE STANDARD?

In terms of [illegible] a Sample 15-2.1, let us suppose that — in advance of looking at the data — the important question is: Does the average of the new type exceed that of the standard?

Procedure	Example
(1) Choose α, the significance level of the test.	(1) Let $\alpha = .01$
(2) Look up $\varphi_{1-\alpha}$ in Table A-12, for the appropriate n.	(2) $n = 10$ $\varphi_{.99} = 0.288$
(3) Compute $\bar{X}$, the mean of the n observations.	(3) $\bar{X} = 10.62$
(4) Compute w, the difference between the largest and smallest of the n observations.	(4) $w = 2.4$
(5) Compute $\varphi = (\bar{X} - m_0)/w$	(5) $\varphi = \frac{10.62 - 10.00}{2.4}$ $= 0.258$
(6) If $\varphi > \varphi_{1-\alpha}$, conclude that the average of the new product exceeds that of the standard; otherwise, there is no reason to believe that the average of the new product exceeds the standard.	(6) Since 0.258 is not larger than 0.288, there is no reason to believe that the average of the new type exceeds that of the standard.

15-2.3 IS THE AVERAGE OF THE NEW PRODUCT LESS THAN THE STANDARD?

In terms of Data Sample 15-2.1, let us suppose that [illegible] advance of looking at the data — the important question is: Is the average of the new type *less than* that of the standard?

Procedure	Example
(1) Choose α, the significance level of the test.	(1) Let $\alpha = .01$
(2) Look up $\varphi_{1-\alpha}$ in Table A-12, for the appropriate n.	(2) $n = 10$ $\varphi_{.99} = 0.288$
(3) Compute $\bar{X}$, the mean of the n observations.	(3) $\bar{X} =$ [illegible]
(4) Compute w, the difference between the largest and smallest of the n observations.	(4) $w = 2.4$
(5) Compute $\varphi = (m_0 - \bar{X})/w$	(5) $\varphi = \frac{10.00 - 10.62}{2.4}$ $= -0.258$
(6) If $\varphi > \varphi_{1-\alpha}$, conclude that the average of the new product is less than that of the standard; otherwise, there is no reason to believe that the average of the new product is less than that of the standard.	(6) Since -0.258 is not larger than 0.288, there is no reason to believe that the average of the new type is less than that of the standard.

15-3 COMPARING THE AVERAGES OF TWO PRODUCTS

15-3.1 DO THE PRODUCTS A AND B DIFFER IN AVERAGE PERFORMANCE?

Data Sample 15-3.1 — Capacity of Batteries

Form: A set of n measurements is available from each of two materials or products. The procedure* given requires that both sets contain the same number of measurements (i.e., $n_A = n_B = n$).

Example: There are available two independent sets of measurements of battery capacity.

Set A	Set B
138	140
143	141
136	139
141	143
140	138
142	140
142	142
146	139
137	141
135	138

Procedure	Example
(1) Choose α, the significance level of the test.	(1) Let $\alpha = .01$
(2) Look up $\varphi'_{1-\alpha/2}$ in Table A-13, for the appropriate n.	(2) $n = 10$ $\varphi'_{.995} = 0.419$
(3) Compute $\bar{X}_A$, $\bar{X}_B$, the means of the two samples.	(3) $\bar{X}_A = 140.0$ $\bar{X}_B = 140.1$
(4) Compute w_A, w_B, the ranges (or difference between the largest and smallest values) for each sample.	(4) $w_A = 146 - 135 = 11$ $w_B = 143 - 138 = 5$
(5) Compute $\varphi' = \dfrac{\bar{X}_A - \bar{X}_B}{\frac{1}{2}(w_A + w_B)}$	(5) $\varphi' = \dfrac{140.0 - 140.1}{8} = -0.0125$
(6) If $\|\varphi'\| > \varphi'_{1-\alpha/2}$, conclude that the averages of the two products differ; otherwise, there is no reason to believe that the averages of A and B differ.	(6) Since 0.0125 is not larger than 0.419, there is no reason to believe that the average of A differs from the average of B.

* This procedure is not appropriate when the observations are "paired", i.e., when each measurement from A is associated with a corresponding measurement from B (see Paragraph 3-3.1.4). In the paired observation case, the question may be answered by the following procedure: compute $\bar{X}_d$ as shown in Paragraph 3-3.1.4 and follow the procedure of Paragraph 15-2.1, using $\bar{X} = \bar{X}_d$ and $m_0 = 0$.

15-3.2 DOES THE AVERAGE OF PRODUCT A EXCEED THE AVERAGE OF PRODUCT B?

In terms of Data Sample 15-3.1, let us suppose that — in advance of looking at the data — the important question is: Does the average of A exceed the average of B?

Again, as in Paragraph 15-3.1, the procedure is appropriate when two independent sets of measurements are available, each containing the same number of observations ($n_A = n_B = n$), but is not appropriate when the observations are paired (see Paragraph 3-3.1.4). In the paired observation case, the question may be answered by the following procedure: compute $\bar{X}_d$ as shown in Paragraph 3-3.2.4, and follow the procedure of Paragraph 15-2.2, using $\bar{X} = \bar{X}_d$ and $m_0 = 0$.

Procedure	Example
(1) Choose α, the significance level of the test.	(1) Let $\alpha = .05$
(2) Look up $\varphi'_{1-\alpha}$ in Table A-13, for the appropriate n.	(2) $n = 10$ $\varphi'_{.95} = .250$
(3) Compute $\bar{X}_A$, $\bar{X}_B$, the means of the two samples.	(3) $\bar{X}_A = 140.0$ $\bar{X}_B = 140.1$
(4) Compute w_A, w_B, the ranges (or difference between the largest and smallest values) for each sample.	(4) $w_A = 11$ $w_B = 5$
(5) Compute $$\varphi' = \frac{\bar{X}_A - \bar{X}_B}{\frac{1}{2}(w_A + w_B)}$$	(5) $\varphi' = \frac{140.0 - 140.1}{8}$ $= -0.0125$
(6) If $\varphi' > \varphi'_{1-\alpha}$, conclude that the average of A exceeds that of B; otherwise, there is no reason to believe that the average of A exceeds that of B.	(6) Since -0.0125 is not larger than 0.250, there is no reason to believe that the average of A exceeds the average of B.

15-4 COMPARING THE AVERAGES OF SEVERAL PRODUCTS DO THE AVERAGES OF t PRODUCTS DIFFER?

Data Sample 15-4 — Breaking-Strength of Cement Briquettes

The following data relate to breaking-strength of cement briquettes (in pounds per square inch).

	Group				
	1	2	3	4	5
	518	508	554	555	536
	560	574	598	567	492
	538	528	579	550	528
	510	534	538	535	572
	544	538	544	540	506
ΣX_i	2670	2682	2813	2747	2634
n_i	5	5	5	5	5
$\bar{X}_i$	534.0	536.4	562.6	549.4	526.8

The question to be answered is: Does the average breaking-strength differ for the different groups?

Procedure

(1) Choose α, the significance level of the test.

(2) Look up L_α in Table A-15, corresponding to t and n.
$n = n_1 = n_2 = \ldots = n_t$, the number of observations on each product.

(3) Compute $w_1, w_2, \ldots, w_t$, the ranges of the n observations from each product.

(4) Compute $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_t$, the means of the observations from each product.

(5) Compute $w' = w_1 + w_2 + \ldots + w_t$.
Compute w'', the difference between the largest and the smallest of the means $\bar{X}_i$.

Example

(1) Let $\alpha = .01$

(2) $t = 5$
$n = 5$
$L_\alpha = 1.02$

(3) $w_1 = 50$
$w_2 = 66$
$w_3 = 60$
$w_4 = 32$
$w_5 = 80$

(4) $\bar{X}_1 = 534.0$
$\bar{X}_2 = 536.4$
$\bar{X}_3 = 562.6$
$\bar{X}_4 = 549.4$
$\bar{X}_5 = 526.8$

(5) $w' = 288$
$w'' = 562.6 - 526.8$
$= 35.8$

Procedure (Cont)	Example (Cont)
(6) Compute $L = nw''/w'$	(6) $L = 179/288$ $= 0.62$
(7) If $L > L_\alpha$, conclude that the averages of the t products differ; otherwise, there is no reason to believe that the averages differ.	(7) Since L is less than L_α, there is no reason to believe that the group averages differ.

15-5 COMPARING TWO PRODUCTS WITH RESPECT TO VARIABILITY OF PERFORMANCE

15-5.1 DOES THE VARIABILITY OF PRODUCT A DIFFER FROM THAT OF PRODUCT B?

The data of Data Sample 15-3.1 are used to illustrate the procedure.
The question to be answered is: Does the variability of A differ from the variability of B?

Procedure	Example
(1) Choose α, the significance level of the test.	(1) Let $\alpha = .01$
(2) Look up $F'_{\alpha/2}(n_A, n_B)$ and $F'_{1-\alpha/2}(n_A, n_B)$ in Table A-11*.	(2) $n_A = 10$ $n_B = 10$ $F'_{.005}(10, 10) = .37$ $F'_{.995}(10, 10) = 2.7$
(3) Compute w_A, w_B, the ranges (or difference between the largest and smallest observations) for A and B, respectively.	(3) $w_A = 11$ $w_B = 5$
(4) Compute $F' = w_A/w_B$	(4) $F' = 11/5$ $= 2.2$
(5) If $F' < F'_{\alpha/2}(n_A, n_B)$ or $F' > F'_{1-\alpha/2}(n_A, n_B)$, conclude that the variability in performance differs; otherwise, there is no reason to believe that the variability differs.	(5) Since F' is not less than .37 and is not greater than 2.7, there is no reason to believe that the variability differs.

* When using Table A-11, sample sizes need not be equal, but cannot be larger than 10.

15-5.2 DOES THE VARIABILITY OF PRODUCT A EXCEED THAT OF PRODUCT B?

In terms of Data Sample 15-3.1, the question to be answered is: Does the variability of A exceed the variability of B?

Procedure	Example
(1) Choose α, the significance level of the test.	(1) Let $\alpha = .01$
(2) Look up $F'_{1-\alpha}(n_A, n_B)$ in Table A-11*.	(2) $n_A = 10$ $n_B = 10$ $F'_{.99}(10, 10) = 2.4$
(3) Compute w_A, w_B, the ranges (or difference between the largest and smallest observations) for A and B, respectively.	(3) $w_A = 11$ $w_B = 5$
(4) Compute $F' = w_A/w_B$	(4) $F' = 11/5$ $= 2.2$
(5) If $F' > F'_{1-\alpha}(n_A, n_B)$, conclude that the variability in performance of A exceeds the variability in performance of B; otherwise, there is no reason to believe that the variability in performance of A exceeds that of B.	(5) Since F' is not larger than $F'_{.99}$, there is no reason to believe that the variability of set A exceeds that of set B.

* When using Table A-11, sample sizes need not be equal, but cannot be larger than 10.

CHAPTER 16

SOME TESTS WHICH ARE INDEPENDENT OF THE FORM OF THE DISTRIBUTION

16-1 GENERAL

This chapter outlines a number of test procedures in which very little is assumed about the nature of the population distributions. In particular, the population distributions are not assumed to be "normal". These tests are often called "nonparametric" tests. The assumptions made here are that the individual observations are independent* and that all observations on a given material (product, or process) have the same underlying distribution. The procedures are strictly correct only if the underlying distribution is continuous, and suitable warnings in this regard are given in each test procedure.

In this chapter, the same wording is used for the problems as was used in AMCP 706-110, Chapter 3 (e.g., "Does the average differ from a standard?"), because the general import of the questions is the same. The specific tests employed, however, are fundamentally different.

If the underlying populations are indeed normal, these tests are poorer than the ones given in Chapter 3, in the sense that β, the probability of the second kind of error, is always larger for given α and n. For some other distributions, however, the nonparametric tests actually may have a smaller error of the second kind. The increase in the second kind of error, when nonparametric tests are applied to normal data, is surprisingly small and is an indication that these tests should receive more use.

Operating characteristic curves and methods of obtaining sample sizes are not given for these tests. Roughly speaking, most of the tests of this chapter require a sample size about 1.1 times that required by the tests given in Chapter 3 (see Paragraphs 3-2 and 3-3 for appropriate normal sample size formulas). For the sign test (Paragraphs 16-2.1, 16-3.1, 16-4.1, 16-5.1, and 16-6.1), a factor of 1.2 is more appropriate.

For the problem of comparing with a standard (Paragraphs 16-2, 16-3, and 16-4), two methods of solution are given and the choice may be made by the user. The sign test (Paragraphs 16-2.1, 16-3.1, and 16-4.1) is a very simple test which is useful under very general conditions. The Wilcoxon signed-ranks test (Paragraphs 16-2.2, 16-3.2, and 16-4.2) requires the assumption that the underlying distribution is symmetrical. When the assumption of symmetry can be made, the signed-ranks test is a more powerful test than the sign test, and is not very burdensome for fairly small samples.

For the problem of comparing two products (Paragraphs 16-5 and 16-6), two methods of solution are also given, but each applies to a specific situation with regard to the source of the data.

The procedures of this chapter assume that the pertinent question has been chosen before taking the observations.

* Except for certain techniques which are given for "paired observations"; in that case, the *pairs* are assumed to be *independent.*

16-2 DOES THE AVERAGE OF A NEW PRODUCT DIFFER FROM A STANDARD?

Data Sample 16-2 — Reverse-Bias Collector Current of Ten Transistors

The data are measurements of I_{CBO} for ten transistors of the same type, where I_{CBO} is the reverse-bias collector current recorded in microamperes.

The standard value m_0 is 0.28μa.

Transistor	I_{CBO}
1	0.28
2	.18
3	.24
4	.30
5	.40
6	.36
7	.15
8	.42
9	.23
10	.48

16-2.1 DOES THE AVERAGE OF A NEW PRODUCT DIFFER FROM A STANDARD? THE SIGN TEST

Procedure

(1) Choose α, the significance level of the test. Table A-33 provides for values of $\alpha = .25$, .10, .05, and .01 for this two-sided test.

(2) Discard observations which happen to be equal to m_0, and let n be the number of observations actually used. (If more than 20% of the observations need to be discarded, this procedure should not be used).

(3) For each observation X_i, record the sign of the difference $X_i - m_0$.
Count the number of occurrences of the less frequent sign. Call this number r.

(4) Look up $r(\alpha, n)$, in Table A-33.

(5) If r is less than, or is equal to, $r(\alpha, n)$, conclude that the average of the new product differs from the standard; otherwise, there is no reason to believe that the averages differ.

Example

(1) Let $\alpha = .05$

(2) In Data Sample 16-2, $m_0 = .28$. Discard the first observation.

$$n = 9$$

(3) The less frequent sign is $-$.

Since there are 4 minus signs,

$$r = 4$$

(4) $r(.05, 9) = 1$

(5) Since r is not less than $r(.05, 9)$, there is no reason to believe that the average current differs from $m_0 = .28\mu$a.

16-2.2 DOES THE AVERAGE OF A NEW PRODUCT DIFFER FROM A STANDARD? THE WILCOXON SIGNED-RANKS TEST

Procedure	Example
(1) Choose α, the significance level of the test. Table A-34 provides for values of $\alpha = .05$, .02, and .01 for this two-sided test. Discard any observations which happen to be equal to m_0, and let n be the number of observations actually used.	(1) Let $\alpha = .05$ In Data Sample 16-2, $m_0 = .28$. Discard the first observation. $n = 9$
(2) Look up $T_\alpha(n)$, in Table A-34.	(2) $T_{.05}(9) = 6$
(3) For each observation X_i, compute $X'_i = X_i - m_0$	(3) (4)
(4) Disregarding signs, rank the X'_i according to their numerical value, i.e., assign the rank of 1 to the X'_i which is numerically smallest, the rank of 2 to the X'_i which is next smallest, etc. In case of ties, assign the average of the ranks which would have been assigned had the X'_i's differed only slightly. (If more than 20% of the observations are involved in ties, this procedure should not be used.) To the assigned ranks 1, 2, 3, etc., prefix a + or a − sign, according to whether the corresponding X'_i is positive or negative.	see table below
(5) Sum the ranks prefixed by a + sign, and the ranks prefixed by a − sign. Let T be the smaller (disregarding sign) of the two sums.	(5) Sum + = 28 Sum − = 17 $T = 17$
(6) If $T \le T_\alpha(n)$, conclude that the average performance of the new type differs from that of the standard; otherwise, there is no reason to believe that the averages differ.	(6) Since T is not less than $T_{.05}(9)$, there is no reason to believe that the average current differs from $m_0 = .28\mu a$.

$X_i - m_0$	Signed rank
−.10	−5
−.04	−2
+.02	+1
+.12	+6
+.08	+4
−.13	−7
+.14	+8
−.05	−3
+.20	+9

16-3 DOES THE AVERAGE OF A NEW PRODUCT EXCEED THAT OF A STANDARD?

Data Sample 16-3 — Reverse-Bias Collector Current of Twenty Transistors

The data are a set of measurements I_{CBO} for 20 transistors, where I_{CBO} is the reverse-bias collector current recorded in microamperes.

The standard value m_0 is 0.28μa.

Transistor	I_{CBO}
1	0.20μa
2	.16
3	.20
4	.48
5	.92
6	.33
7	.20
8	.53
9	.42
10	.50
11	.19
12	.22
13	.18
14	.17
15	1.20
16	.14
17	.09
18	.13
19	.26
20	.66

16-3.1 DOES THE AVERAGE OF A NEW PRODUCT EXCEED THAT OF A STANDARD? THE SIGN TEST

Procedure

(1) Choose α, the significance level of the test. Table A-33 provides for values of α = .125, .05, .025, and .005 for this one-sided test.

(2) Discard observations which happen to be equal to m_0, and let n be the number of observations actually used. (If more than 20% of the observations need to be discarded, this procedure should not be used.)

(3) For each observation X_i, record the sign of the difference $X_i - m_0$.
Count the number of minus signs.
Call this number r.

(4) Look up $r(\alpha, n)$, in Table A-33.

Example

(1) Let $\alpha = .025$

(2) In Data Sample 16-3, $m_0 = .28$. Since no observations are equal to m_0,

$$n = 20$$

(3)

$$r = 12$$

(4) $r(.025, 20) = 5$

Procedure (Cont)

(5) If r is less than, or is equal to, $r(\alpha, n)$, conclude that the average of the new product exceeds the standard; otherwise, there is no reason to believe that the average of the new product exceeds that of the standard.

Example (Cont)

(5) Since r is not less than $r(.025, 20)$, there is no reason to believe that the average current exceeds $m_0 = .28\mu a$.

16-3.2 DOES THE AVERAGE OF A NEW PRODUCT EXCEED THAT OF A STANDARD? THE WILCOXON SIGNED-RANKS TEST

Procedure

(1) Choose α, the significance level of the test. Table A-34 provides for values of $\alpha = .025$, .01, and .005 for this one-sided test. Discard any observations which happen to be equal to m_0, and let n be the number of observations actually used.

(2) Look up $T_\alpha(n)$, in Table A-34.

(3) For each observation X_i, compute

$$X'_i = X_i - m_0.$$

(4) Disregarding signs, rank the X'_i according to their numerical value, i.e., assign the rank of 1 to the X'_i which is numerically smallest, the rank of 2 to the X'_i which is next smallest, etc. In case of ties, assign the average of the ranks which would have been assigned had the X'_i's differed only slightly. (If more than 20% of the observations are involved in ties, this procedure should not be used.)
To the assigned ranks 1, 2, 3, etc., prefix a + or a − sign according to whether the X'_i is positive or negative.

(5) Let T be the absolute value of the sum of the ranks preceded by a negative sign.

(6) If $T \le T_\alpha(n)$, conclude that the average performance of the new product exceeds that of the standard; otherwise, there is no reason to believe that the average of the new product exceeds that of the standard.

Example

(1) Let $\alpha = .025$
In Data Sample 16-3, $m_0 = .28\mu a$. Since no observations are equal to m_0,

$$n = 20$$

(2) $T_{.025}(20) = 52$

(3) (4)

$X_i - m_0$	Signed Rank
−0.08	− 5
−0.12	−10
−0.08	− 5
0.20	+15
0.64	+19
0.05	+ 2
−0.08	− 5
0.25	+17
0.14	+11.5
0.22	+16
−0.09	− 7
−0.06	− 3
−0.10	− 8
−0.11	− 9
0.92	+20
−0.14	−11.5
−0.19	−14
−0.15	−13
−0.02	− 1
0.38	+18

(5) $T = 91.5$

(6) Since T is not smaller than $T_{.025}(20)$, there is no reason to believe that the average current exceeds $m_0 = .28\mu a$.

16-4 IS THE AVERAGE OF A NEW PRODUCT LESS THAN THAT OF A STANDARD?

Data Sample 16-4 — Tensile Strength of Aluminum Alloy

The data are measurements of ultimate tensile strength (psi) for twenty test specimens of aluminum alloy. The standard value for tensile strength is m_0 = 27,000 psi.

Specimen	Ultimate Tensile Strength (psi)
1	24,200
2	25,900
3	26,000
4	26,000
5	26,300
6	26,450
7	27,250
8	27,450
9	27,550
10	28,550
11	29,150
12	29,900
13	30,000
14	30,400
15	30,450
16	30,450
17	31,450
18	31,600
19	32,400
20	33,750

16-4.1 IS THE AVERAGE OF A NEW PRODUCT LESS THAN THAT OF A STANDARD? THE SIGN TEST

Procedure	Example
(1) Choose α, the significance level of the test. Table A-33 provides for values of α = .125, .05, .025, and .005 for this one-sided test.	(1) Let $\alpha = .025$
(2) Discard observations which happen to be equal to m_0, and let n be the number of observations actually used. (If more than 20% of the observations need to be discarded, this procedure should not be used.)	(2) In Data Sample 16-4, m_0 = 27,000. Since no observations are equal to m_0, $n = 20$
(3) For each observation X_i, record the sign of the difference $X_i - m_0$. Count the number of plus signs. Call this number r.	(3) There are 14 plus signs. $r = 14$
(4) Look up $r(\alpha, n)$, in Table A-33.	(4) $r(.025, 20) = 5$

Procedure (Cont)

(5) If r is less than, or is equal to, $r(\alpha, n)$, conclude that the average of the new product is less than the standard; otherwise, there is no reason to believe that the average of the new product is less than the standard.

Example (Cont)

(5) Since r is not less than $r(.025, 20)$, there is no reason to believe that the average tensile strength is less than $m_0 = 27{,}000$ psi.

16-4.2 IS THE AVERAGE OF A NEW PRODUCT LESS THAN THAT OF A STANDARD? THE WILCOXON SIGNED-RANKS TEST

Procedure

(1) Choose α, the significance level of the test. Table A-34 provides for values of $\alpha = .025$, .01, and .005 for this one-sided test. Discard any observations which happen to be equal to m_0, and let n be the number of observations actually used.

(2) Look up $T_\alpha(n)$, in Table A-34.

(3) For each observation X_i, compute

$$X'_i = X_i - m_0.$$

(4) Disregarding signs, rank the X'_i according to their numerical value, i.e., assign the rank of 1 to the X'_i which is numerically smallest, the rank of 2 to the X'_i which is next smallest, etc. In case of ties, assign the average of the ranks which would have been assigned had the X'_i's differed only slightly. (If more than 20% of the observations are involved in ties, this procedure should not be used.)
To the assigned ranks 1, 2, 3, etc., prefix a + or a − sign according to whether the corresponding X'_i is positive or negative.

(5) Let T be the sum of the ranks preceded by a + sign.

(6) If $T \le T_\alpha(n)$, conclude that the average of the new product is less than that of the standard; otherwise, there is no reason to believe that the average of the new product is less than that of the standard.

Example

(1) Let $\alpha = .025$
In Data Sample 16-4,

$$m_0 = 27{,}000.$$

Since no observations are equal to m_0,

$$n = 20$$

(2) $T_{.025}(20) = 52$

(3) (4)

$X_i - m_0$	Signed Rank
−2800	−11
−1100	− 8
−1000	− 6.5
−1000	− 6.5
− 700	− 5
− 550	− 3.5
250	+ 1
450	+ 2
550	+ 3.5
1550	+ 9
2150	+10
2900	+12
3000	+13
3400	+14
3450	+15.5
3450	+15.5
4450	+17
4600	+18
5400	+19
6750	+20

(5) $T = 169.5$

(6) Since T is not less than $T_{.025}(20)$, there is no reason to believe that the average tensile strength is less than $m_0 = 27{,}000$ psi.

16-5 DO PRODUCTS A AND B DIFFER IN AVERAGE PERFORMANCE?

Two procedures are given to answer this question. Each of the procedures is applicable to a different situation, depending upon how the data have been taken.

Situation 1 (for which the sign test of Paragraph 16-5.1 is applicable) is the case where observations on the two things being compared have been obtained in pairs. Each of the two observations on a pair has been obtained under similar conditions, but the different pairs need not have been obtained under similar conditions. Specifically, the sign test procedure tests whether the *median* difference between A and B can be considered equal to zero.

Situation 2 (for which we use the Wilcoxon-Mann-Whitney test of Paragraph 16-5.2) is the case where two independent samples have been drawn — one from population A and one from population B. This test answers the following kind of questions — if the two distributions are of the same form, are they displaced with respect to each other? Or, if the distributions are quite different in form, do the observations on A systematically tend to exceed the observations on B?

16-5.1 DO PRODUCTS A AND B DIFFER IN AVERAGE PERFORMANCE? THE SIGN TEST FOR PAIRED OBSERVATIONS

Data Sample 16-5.1 — Reverse-Bias Collector Currents of Two Types of Transistors

Ten pairs of measurements of I_{CBO} on two types of transistors are available, as follows:

Type A	Type B
.19	.21
.22	.27
.18	.15
.17	.18
1.20	.40
.14	.08
.09	.14
.13	.28
.26	.30
.66	.68

Procedure	Example
(1) Choose α, the significance level of the test. Table A-33 provides for values of $\alpha = .25$, .10, .05, and .01 for this two-sided test.	(1) Let $\alpha = .10$
(2) For each pair, record the sign of the difference $X_A - X_B$. Discard any difference which happens to equal zero. Let n be the number of differences remaining. (If more than 20% of the observations need to be discarded, this procedure should not be used.)	(2) In Data Sample 16-5.1, $n = 10$
(3) Count the number of occurrences of the less frequent sign. Call this r.	(3) There are 3 plus signs. $r = 3$
(4) Look up $r(\alpha, n)$, in Table A-33.	(4) $r(.10, 10) = 1$
(5) If r is less than, or is equal to, $r(\alpha, n)$, conclude that the averages differ; otherwise, there is no reason to believe that the averages differ.	(5) Since r is not less than $r(.10, 10)$, there is no reason to believe that the two types differ in average current.

Note: The Wilcoxon Signed-Ranks Test also may be used to compare the averages of two products in the paired-sample situation; follow the procedure of Paragraph 16-2.2, substituting $X'_i = X_A - X_B$ for $X'_i = X_i - m_0$ in step (3) of that procedure.

16-5.2 DO PRODUCTS A AND B DIFFER IN AVERAGE PERFORMANCE? THE WILCOXON-MANN-WHITNEY TEST FOR TWO INDEPENDENT SAMPLES

Data Sample 16-5.2 — Forward Current Transfer Ratio of Two Types of Transistors

The data are measurements of h_{fe} for two independent groups of transistors, where h_{fe} is the small-signal short-circuit forward current transfer ratio.

Group A	Group B
50.5 (9)*	57.0 (17)
37.5 (1)	52.0 (11)
49.8 (7)	51.0 (10)
56.0 (15.5)	44.2 (3)
42.0 (2)	55.0 (14)
56.0 (15.5)	62.0 (19)
50.0 (8)	59.0 (18)
54.0 (13)	45.2 (5)
48.0 (6)	53.5 (12)
	44.4 (4)

* The numbers shown in parentheses are the ranks, from lowest to highest, for all observations combined, as required in Step (2) of the following Procedure and Example.

Procedure	Example
(1) Choose α, the significance level of the test. Table A-35 provides for values of $\alpha = .01$, .05, .10, and .20 for this two-sided test when $n_A, n_B \le 20$.	(1) Let $\alpha = .10$
(2) Combine the observations from the two samples, and rank them in order of increasing size from smallest to largest. Assign the rank of 1 to the lowest, a rank of 2 to the next lowest, etc. (Use algebraic size, i.e., the lowest rank is assigned to the largest negative number, if there are negative numbers). In case of ties, assign to each the average of the ranks which would have been assigned had the tied observations differed only slightly. (If more than 20% of the observations are involved in ties, this procedure should not be used.)	(2) In Data Sample 16-5.2, the ranks of the nineteen individual observations, from lowest to highest, are shown in parentheses beside the respective observations. Note that the two tied observations (56.0) are each given the rank 15.5 (instead of ranks 15 and 16), and that the next larger observation is given the rank 17.
(3) Let: n_1 = smaller sample n_2 = larger sample $n = n_1 + n_2$	(3) $n_1 = 9$ $n_2 = 10$ $n = 19$
(4) Compute R, the sum of the ranks for the smaller sample. (If the two samples are equal in size, use the sum of the ranks for either sample.) Compute $R' = n_1(n+1) - R$	(4) $R = 77$ $R' = 9\ (20) - 77 = 103$
(5) Look up $R_\alpha(n_1, n_2)$, in Table A-35.	(5) $R_{.10}(9, 10) = 69$
(6) If either R or R' is smaller than, or is equal to, $R_\alpha(n_1, n_2)$, conclude that the averages of the two products differ; otherwise, there is no reason to believe that the averages of the two products differ.	(6) Since neither R nor R' is smaller than $R_{.10}(9, 10)$, there is no reason to believe that the averages of the two groups differ.

16-6 DOES THE AVERAGE OF PRODUCT A EXCEED THAT OF PRODUCT B?

Two procedures are given to answer this question. In order to choose the procedure that is appropriate to a particular situation, read the discussion in Paragraph 16-5.

16-6.1 DOES THE AVERAGE OF PRODUCT A EXCEED THAT OF PRODUCT B? THE SIGN TEST FOR PAIRED OBSERVATIONS

In terms of Data Sample 16-5.1, assume that we had asked in advance (not after looking at the data) whether the average I_{CBO} was larger for Type A than for Type B.

Procedure	Example
(1) Choose α, the significance level of the test. Table A-33 provides for values of $\alpha = .125$, .05, .025, and .005 for this one-sided test.	(1) Let $\alpha = .025$
(2) For each pair, record the sign of the difference $X_A - X_B$. Discard any difference which happens to equal zero. Let n be the number of differences remaining. (If more than 20% of the observations need to be discarded, this procedure should not be used.)	(2) In Data Sample 16-5.1, $n = 10$
(3) Count the number of minus signs. Call this number r.	(3) There are 7 minus signs. $r = 7$
(4) Look up $r(\alpha, n)$, in Table A-33.	(4) $r(.025, 10) = 1$
(5) If r is less than, or is equal to, $r(\alpha, n)$, conclude that the average of product A exceeds the average of product B; otherwise, there is no reason to believe that the average of product A exceeds that of product B.	(5) Since r is not less than $r(.025, 10)$, there is no reason to believe that the average of Type A exceeds the average of Type B.

Note: The Wilcoxon Signed-Ranks Test also may be used to compare the averages of two products in the paired-sample situations; follow the procedure of Paragraph 16-3.2, substituting $X_i' = X_A - X_B$ for $X_i' = X_i - m_0$ in Step (3) of that Procedure.

16-6.2 DOES THE AVERAGE OF PRODUCT A EXCEED THAT OF PRODUCT B? THE WILCOXON-MANN-WHITNEY TEST FOR TWO INDEPENDENT SAMPLES

Data Sample 16-6.2 — Output Admittance of Two Types of Transistors

The data are observations of h_{ob} for two types of transistors, where h_{ob} = small-signal open-circuit output admittance.

Type A	Type B
.291 (5)*	.246 (1)
.390 (10)	.252 (2)
.305 (7)	.300 (6)
.331 (9)	.289 (4)
.316 (8)	.258 (3)

* The numbers shown in parentheses are the ranks, from lowest to highest, for all observations combined, as required in Step (2) of the following Procedure and Example.

Does the average h_{ob} for Type A exceed that for Type B?

Procedure	Example
(1) Choose α, the significance level of the test. Table A-35 provides for values of $\alpha = .005, .025, .05$, and .10 for this one-sided test, when $n_A, n_B \leq 20$.	(1) Let $\alpha = .05$
(2) Combine the observations from the two populations, and rank them in order of increasing size from smallest to largest. Assign the rank of 1 to the lowest, a rank of 2 to the next lowest, etc. (Use algebraic size, i.e., the lowest rank is assigned to the largest negative number if there are negative numbers). In case of ties, assign to each the average of the ranks which would have been assigned had the tied observations differed only slightly. (If more than 20% of the observations are involved in ties, this procedure should not be used.)	(2) In Data Sample 16-6.2, the ranks of the ten individual observations, from lowest to highest, are shown beside the respective observations.
(3) Let: n_1 = smaller sample n_2 = larger sample $n = n_1 + n_2$	(3) $n_1 = 5$ $n_2 = 5$ $n = 10$
(4) Look up $R_\alpha(n_1, n_2)$, in Table A-35.	(4) $R_{.05}(5, 5) = 19$
(5a) If the two samples are equal in size, or if n_B is the smaller, compute R_B the sum of the ranks for sample B. If R_B is less than, or is equal to, $R_\alpha(n_1, n_2)$, conclude that the average for product A exceeds that for product B; otherwise, there is no reason to believe that the average for product A exceeds that for product B.	(5a) $R_B = 16$ Since R_B is less than $R_{.05}(5, 5)$, conclude that the average for Type A exceeds that for Type B.
(5b) If n_A is smaller than n_B, compute R_A the sum of the ranks for sample A, and compute $R'_A = n_A(n + 1) - R_A$. If R'_A is less than, or is equal to, $R_\alpha(n_1, n_2)$, conclude that the average for product A exceeds that for product B; otherwise, there is no reason to believe that the two products differ.	

16-7 COMPARING THE AVERAGES OF SEVERAL PRODUCTS DO THE AVERAGES OF t PRODUCTS DIFFER?

Data Sample 16-7 — Life Tests of Three Types of Stopwatches

Samples from each of three types of stopwatches were tested. The following data are thousands of cycles (on-off-restart) survived until some part of the mechanism failed.

Type 1	Type 2	Type 3
1.7 (1)*	13.6 (6)	13.4 (5)
1.9 (2)	19.8 (8)	20.9 (9)
6.1 (3)	25.2 (12)	25.1 (10.5)
12.5 (4)	46.2 (16.5)	29.7 (13)
16.5 (7)	46.2 (16.5)	46.9 (18)
25.1 (10.5)	61.1 (19)	
30.5 (14)		
42.1 (15)		
82.5 (20)		

* The numbers shown in parentheses are the ranks, from lowest to highest, for all observations combined, as required in Step (3) of the following Procedure and Example.

TABLE 16-1. WORK TABLE FOR DATA SAMPLE 16-7

	Ranks Type 1	Ranks Type 2	Ranks Type 3
	1	6	5
	2	8	9
	3	12	10.5
	4	16.5	13
	7	16.5	18
	10.5	19	
	14		
	15		
	20		
R_i	$R_1 = 76.5$	$R_2 = 78.0$	$R_3 = 55.5$
n_i	9	6	5
R_i^2/n_i	650.25	1014.00	616.05

Does the average length of "life" differ for the three types?

Procedure	Example
(1) Choose α, the significance level of the test.	(1) Let $\alpha = .10$
(2) Look up $\chi^2_{1-\alpha}$ for $t - 1$ degrees of freedom, in Table A-3, where t is the number of products to be compared.	(2) $t = 3$ $\chi^2_{.90}$ for 2 d.f. $= 4.61$
(3) We have $n_1, n_2, \ldots, n_t$ observations on each of the products $1, 2, \ldots, t$. $N = n_1 + n_2 + \ldots + n_t$. Assign ranks to each observation according to its size in relation to all N observations. That is, assign rank 1 to the smallest, 2 to the next larger, etc., and N to the largest. In case of ties, assign to each of the tied observations the average of the ranks which would have been assigned had the observations differed slightly. (If more than 20% of the observations are involved in ties, this procedure should not be used.)	(3) In Data Sample 16-7, $N = 9 + 6 + 5 = 20$. The assigned ranks are shown in Data Sample 16-7 and in Table 16-1.
(4) Compute R_i, the sum of the ranks of the observations on the ith product, for each of the products.	(4) $R_1 = 76.5$ $R_2 = 78.0$ $R_3 = 55.5$
(5) Compute $H = \frac{12}{N(N+1)} \sum_{i=1}^{t} \frac{R_i^2}{n_i} - 3(N+1)$	(5) $H = \frac{12}{420}(2280.30) - 63$ $= 2.15$
(6) If $H > \chi^2_{1-\alpha}$, conclude that the averages of the t products differ; otherwise, there is no reason to believe that the averages differ.	(6) Since H is not larger than $\chi^2_{.90}$, there is no reason to believe that the averages for the three types differ.

Note: When using this Procedure, each of the n_i should be at least 5. If any n_i are less than 5, the level of significance α may be considerably different from the intended value.

CHAPTER 17

THE TREATMENT OF OUTLIERS

17-1 THE PROBLEM OF REJECTING OBSERVATIONS

Every experimenter, at some time, has obtained a set of observations, purportedly taken under the same conditions, in which one observation was widely different, or an outlier from the rest.

The problem that confronts the experimenter is whether he should keep the suspect observation in computation, or whether he should discard it as being a faulty measurement. The word reject will mean *reject in computation*, since every observation should be recorded. A careful experimenter will want to make a record of his "rejected" observations and, where possible, detect and carefully analyze their cause(s).

It should be emphasized that we are not discussing the case where we *know* that the observation differs because of an assignable cause, i.e., a dirty test-tube, or a change in operating conditions. We are dealing with the situation where, as far as we are able to ascertain, all the observations are on approximately the same footing. One observation is suspect however, in that it seems to be set apart from the others. We wonder whether it is not so far from the others that we can reject it as being caused by some assignable but thus far unascertained cause.

When a measurement is far-removed from the great majority of a set of measurements of a quantity, and thus possibly reflects a gross error, the question of whether that measurement should have a full vote, a diminished vote, or no vote in the final average — and in the determination of precision — is a very difficult question to answer completely in general terms. If on investigation, a trustworthy explanation of the discrepancy is found, common sense dictates that the value concerned should be excluded from the final average and from the estimate of precision, since these presumably are intended to apply to the unadulterated system. If, on the other hand, no explanation for the apparent anomalousness is found, then common sense would seem to indicate that it should be included in computing the final average and the estimate of precision. Experienced investigators differ in this matter. Some, e.g., J. W. Bessel, would always include it. Others would be inclined to exclude it, on the grounds that it is better to exclude a possibly "good" measurement than to include a possibly "bad" one. The argument for exclusion is that *when a "good" measurement is excluded* we simply lose some of the relevant information, with consequent decrease in precision and the introduction of some bias (both being theoretically computable); whereas, *when a truly anomalous measurement is included* it vitiates our results, biasing both the final average and the estimate of precision by unknown, and generally unknowable, amounts.

There have been many criteria proposed for guiding the rejection of observations. For an excellent summary and critical review of the classical rejection procedures, and some more modern ones, see P. R. Rider[1]. One of the more famous classical rejection rules is "Chauvenet's criterion," which is not recommended. This criterion is based on the normal distribution and advises rejection of an extreme observation if the probability of occurrence of such deviation from the mean of the n measurements is less than $\frac{1}{2}n$. Obviously, for small n, such a criterion rejects too easily.

A review of the history of rejection criteria, and the fact that new criteria are still being proposed, leads us to realize that no completely satisfactory rule can be devised for any and all situations. We cannot devise a criterion that will not reject a predictable amount from endless arrays of perfectly good data; the amount of data rejected of course depends on the rule used. This is the price we pay for using any rule for rejection of data. No available criteria are superior to the judgment of an experienced investigator who is thoroughly familiar with his measurement process. For an excellent discussion of this point, see E. B. Wilson, Jr.[2]. Statistical rules are given primarily for the benefit of inexperienced investigators, those working with a new process, or those who simply want justification for what they would have done anyway.

Whatever rule is used, it must bear some resemblance to the experimenter's feelings about the nature and possible frequency of errors. For an extreme example — if the experimenter feels that about one outlier in twenty reflects an actual blunder, and he uses a rejection rule that throws out the two extremes in every sample, then his reported data obviously will be "clean" with respect to extreme blunders — but the effects of "little" blunders may still be present. The *one and only* sure way to avoid publishing any "bad" results is to throw away all results.

With the foregoing reservations, Paragraphs 17-2 and 17-3 give some suggested procedures for judging outliers. In general, the rules to be applied to a single experiment (see Paragraph 17-3) reject only what would be rejected by an experienced investigator anyway.

17-2 REJECTION OF OBSERVATIONS IN ROUTINE EXPERIMENTAL WORK

The best tools for detection of *errors* (e.g., systematic errors, gross errors) in routine work are the control charts for the mean and range. These charts are described in Chapter 18, which also contains a table of factors to facilitate their application, Table 18-2.

17-3 REJECTION OF OBSERVATIONS IN A SINGLE EXPERIMENT

We assume that our experimental observations (except for the truly discordant ones) come from a single normal population with mean m and standard deviation σ. In a particular experiment, we have obtained n observations and have arranged them in order from lowest to highest ($X_1 \leq X_2 \leq \ldots \leq X_n$). We consider procedures applicable to two situations: when observations which are either too large or too small would be considered faulty and rejectable, see Paragraph 17-3.1; when we consider rejectable those observations that are extreme in one direction only (e.g., when we want to reject observations that are too large but never those that are too small, or vice versa), see Paragraph 17-3.2. The proper choice between the situations must be made on *a priori* grounds, and not on the basis of the data to be analyzed.

For each situation, procedures are given for four possible cases with regard to our knowledge of m and σ.

17-3.1 WHEN EXTREME OBSERVATIONS IN EITHER DIRECTION ARE CONSIDERED REJECTABLE

17-3.1.1 Population Mean and Standard Deviation Unknown — Sample in Hand is the Only Source of Information.

[The Dixon Criterion]

Procedure

(1) Choose α, the probability or risk we are willing to take of rejecting an observation that really belongs in the group.

(2) If:

$3 \leq n \leq 7$	Compute r_{10}
$8 \leq n \leq 10$	Compute r_{11}
$11 \leq n \leq 13$	Compute r_{21}
$14 \leq n \leq 25$	Compute r_{22},

where r_{ij} is computed as follows:

r_{ij}	If X_n is Suspect	If X_1 is Suspect
r_{10}	$(X_n - X_{n-1})/(X_n - X_1)$	$(X_2 - X_1)/(X_n - X_1)$
r_{11}	$(X_n - X_{n-1})/(X_n - X_2)$	$(X_2 - X_1)/(X_{n-1} - X_1)$
r_{21}	$(X_n - X_{n-2})/(X_n - X_2)$	$(X_3 - X_1)/(X_{n-1} - X_1)$
r_{22}	$(X_n - X_{n-2})/(X_n - X_3)$	$(X_3 - X_1)/(X_{n-2} - X_1)$

(3) Look up $r_{1-\alpha/2}$ for the r_{ij} from Step (2), in Table A-14.

(4) If $r_{ij} > r_{1-\alpha/2}$, reject the suspect observation; otherwise, retain it.

17-3.1.2 Population Mean and Standard Deviation Unknown — Independent External Estimate of Standard Deviation is Available.

[The Studentized Range]

Procedure

(1) Choose α, the probability or risk we are willing to take of rejecting an observation that really belongs in the group.

(2) Look up $q_{1-\alpha}(n, \nu)$ in Table A-10. n is the number of observations in the sample, and ν is the number of degrees of freedom for s, the independent external estimate of the standard deviation obtained from concurrent or past data — *not* from the sample in hand.

(3) Compute $w = q_{1-\alpha}s$.

(4) If $X_n - X_1 > w$, reject the observation that is suspect; otherwise, retain it.

17-3.1.3 Population Mean Unknown — Value for Standard Deviation Assumed.

Procedure

(1) Choose α, the probability or risk we are willing to take of rejecting an observation that really belongs in the group.

(2) Look up $q_{1-\alpha}(n, \infty)$ in Table A-10.

(3) Compute $w = q_{1-\alpha}\sigma$.

(4) If $X_n - X_1 > w$, reject the observation that is suspect; otherwise, retain it.

17-3.1.4 Population Mean and Standard Deviation Known.

Procedure	Example
(1) Choose α, the probability or risk we are willing to take of rejecting an observation when all n really belong in the same group.	(1) Let $\alpha = .10$, for example.
(2) Compute $\alpha' = 1 - (1 - \alpha)^{1/n}$ (We can compute this value using logarithms, or by reference to a table of fractional powers.)	(2) If $n = 20$, for example, $\alpha' = 1 - (1 - .10)^{1/20} = 1 - (.90)^{1/20} = 1 - .9947 = .0053$
(3) Look up $z_{1-\alpha'/2}$ in Table A-2. (Interpolation in Table A-2 may be required. The recommended method is graphical interpolation, using probability paper.)	(3) $1 - \alpha'/2 = 1 - (.0053/2) = .9974$; $z_{.9974} = 2.80$
(4) Compute: $a = m - \sigma z_{1-\alpha'/2}$; $b = m + \sigma z_{1-\alpha'/2}$	(4) $a = m - 2.80\,\sigma$; $b = m + 2.80\,\sigma$
(5) Reject any observation that does not lie in the interval from a to b.	(5) Reject any observation that does not lie in the interval from $m - 2.80\,\sigma$ to $m + 2.80\,\sigma$.

17-3.2 WHEN EXTREME OBSERVATIONS IN ONLY ONE DIRECTION ARE CONSIDERED REJECTABLE

17-3.2.1 Population Mean and Standard Deviation Unknown — Sample in Hand is the Only Source of Information.

[The Dixon Criterion]

Procedure

(1) Choose α, the probability or risk we are willing to take of rejecting an observation that really belongs in the group.

(2) If:

$3 \le n \le 7$	Compute r_{10}
$8 \le n \le 10$	Compute r_{11}
$11 \le n \le 13$	Compute r_{21}
$14 \le n \le 25$	Compute r_{22},

where r_{ij} is computed as follows:

r_{ij}	If Only Large Values are Suspect	If Only Small Values are Suspect
r_{10}	$(X_n - X_{n-1})/(X_n - X_1)$	$(X_2 - X_1)/(X_n - X_1)$
r_{11}	$(X_n - X_{n-1})/(X_n - X_2)$	$(X_2 - X_1)/(X_{n-1} - X_1)$
r_{21}	$(X_n - X_{n-2})/(X_n - X_2)$	$(X_3 - X_1)/(X_{n-1} - X_1)$
r_{22}	$(X_n - X_{n-2})/(X_n - X_3)$	$(X_3 - X_1)/(X_{n-2} - X_1)$

(3) Look up $r_{1-\alpha}$ for the r_{ij} from Step (2), in Table A-14.

(4) If $r_{ij} > r_{1-\alpha}$, reject the suspect observation; otherwise, retain it.

17-3.2.2 Population Mean and Standard Deviation Unknown — Independent External Estimate of Standard Deviation is Available.

[Extreme Studentized Deviate From Sample Mean; The Nair Criterion]

Procedure

(1) Choose α, the probability or risk we are willing to take of rejecting an observation that really belongs in the group.

(2) Look up $t_\alpha(n, \nu)$ in Table A-16. n is the number of observations in the sample, and ν is the number of degrees of freedom for s_ν the independent external estimate of the standard deviation obtained from concurrent or past data — *not* from the sample in hand.

(3) If only observations that are too large are considered rejectable, compute

$$t_n = (X_n - \bar{X})/s_\nu .$$

Or, if only observations that are too small are considered rejectable, compute

$$t_1 = (\bar{X} - X_1)/s_\nu .$$

(4) If t_n (or t_1, as appropriate) is larger than $t_\alpha(n, \nu)$, reject the observation that is suspect; otherwise, retain it.

17-3.2.3 Population Mean Unknown — Value for Standard Deviation Assumed.

[Extreme Standardized Deviate From Sample Mean]

Procedure

(1) Choose α, the probability or risk we are willing to take of rejecting an observation that really belongs in the group.

(2) Look up $t_\alpha(n, \infty)$ in Table A-16.

(3) If observations that are too large are considered rejectable, compute

$$t_n = (X_n - \bar{X})/\sigma .$$

Or, if observations that are too small are considered rejectable, compute

$$t_1 = (\bar{X} - X_1)/\sigma .$$

(4) If t_n (or t_1, as appropriate) is larger than $t_\alpha(n, \infty)$, reject the observation that is suspect; otherwise, retain it.

17-3.2.4 Population Mean and Standard Deviation Known.

Procedure	Example
(1) Choose α, the probability or risk we are willing to take of rejecting an observation when all n really belong in the same group.	(1) Let $\alpha = .10$, for example.
(2) Compute $\alpha'/2 = 1 - (1 - \alpha)^{1/n}$. (We can compute this value using logarithms, or by reference to a table of fractional powers.)	(2) If $n = 20$, for example, $\alpha'/2 = 1 - (1 - .10)^{1/20} = 1 - (.90)^{1/20} = 1 - .9947 = .0053$
(3) Look up $z_{1-\alpha'/2}$ in Table A-2. (Interpolation in Table A-2 may be required. The recommended method is graphical interpolation using probability paper.)	(3) $1 - \alpha'/2 = 1 - .0053 = .9947$; $z_{.9947} = 2.55$
(4) Compute: $a = m - \sigma z_{1-\alpha'/2}$, $b = m + \sigma z_{1-\alpha'/2}$	(4) $a = m - 2.55\,\sigma$, $b = m + 2.55\,\sigma$
(5) Reject any observation that does not lie in the interval from a to b.	(5) Reject any observation that does not lie in the interval from $m - 2.55\,\sigma$ to $m + 2.55\,\sigma$.

REFERENCES

1. P. R. Rider, "Criteria for Rejection of Observations," *Washington University Studies — New Series, Science and Technology*, No. 8, October 1933.
2. E. B. Wilson, Jr., *An Introduction to Scientific Research*, pp. 256-258, McGraw-Hill Book Co., Inc., New York, N. Y., 1952.

ADDITIONAL READING

W. J. Dixon, "Processing Data for Outliers," *Biometrics*, Vol. 9, No. 1, pp. 74-89, March 1953.

F. E. Grubbs, "Sample Criteria for Testing Outlying Observations," *Annals of Mathematical Statistics*, Vol. 21, No. 1, pp. 27-58, March 1950.

A. Hald, *Statistical Theory with Engineering Applications*, pp. 333-336, John Wiley & Sons, Inc., New York, N. Y., 1952.

F. Proschan, "Testing Suspected Observations," *Industrial Quality Control*, Vol. XIII, No. 7, pp. 14-19, January 1957.

CHAPTER 18

THE PLACE OF CONTROL CHARTS IN EXPERIMENTAL WORK

18-1 PRIMARY OBJECTIVE OF CONTROL CHARTS

Control charts have very important functions in experimental work, although their use in laboratory situations has been discussed only briefly by most textbooks. Control charts can be used as a form of statistical test in which the primary objective is to test whether or not the process is in *statistical control*. The process is in *statistical control* when repeated samples from the process behave as random samples from a stable probability distribution; thus, the underlying conditions of a process *in control* are such that it is possible to make predictions in the probability sense.

The control limits are usually computed by using formulas which utilize the information from the samples themselves. The computed limits are placed as lines on the specific chart, and the decision is made that the process was *in control* if all points fall within the control limits. If all points are not within the limits, then the decision is made that the process is *not in control*.

The basic assumption underlying most statistical techniques is that the data are a random sample from a stable probability distribution, which is another way of saying that the process is in *statistical control*. It is the validity of this basic assumption which the control chart is designed to test. The control chart is used to demonstrate the existence of statistical control, and to monitor a controlled process. As a monitor, a given control chart indicates a particular type of departure from control.

18-2 INFORMATION PROVIDED BY CONTROL CHARTS

Control charts provide a running graphical record of small subgroups of data taken from a repetitive process. Control charts may be kept on any of various characteristics of each small subgroup — e.g., on the average, standard deviation, range, or proportion defective. The chart for each particular characteristic is designed to detect certain specified departures in the process from the assumed conditions. The process may be a measurement process as well as a production process. The order of groups is usually with respect to time, but not necessarily so. The grouping is such that the members of the same group are more likely to be alike than are members of different groups.

Primarily, control charts can be used to demonstrate whether or not the process is in statistical control. When the charts show lack of control, they indicate where or when the trouble occurred. Often they indicate the nature of the trouble, e.g., trends or runs, sudden shifts in the mean, increased variability, etc.

In addition to serving as a method of testing for control, control charts also provide additional and useful information in the form of estimates of the characteristics of a controlled process. This information is altogether too-frequently overlooked. For example, one very important piece of information which can be obtained from a control chart for the range or standard deviation is an estimate of the variability σ of a routine measurement or production process. It should be remembered that many of the techniques of Section 1, Chapter 3, are given in parallel for known σ and unknown σ. Most experimental scientists have very good knowledge of the variability of their measurements, but hesitate to assume *known* σ without additional justification. Control charts can be used to provide the justification.

Finally, as was pointed out in Chapter 17, Paragraph 17-2, a control chart is the most satisfactory criterion for rejection of observations in a *routine* laboratory operation. An excellent discussion of the use of control charts to detect particular kinds of trouble is given by Olmstead[1]. The three most important types of control charts in this connection are the charts for the average $\bar{X}$, range R, and standard deviation σ. The order of usefulness of each type of chart in particular situations is shown in Table 18-1, where a "1" means most useful, "2" is the next best, and dots denote "not appropriate".

As can be seen from Table 18-1, the $\bar{X}$ and R charts are the most useful of the three types. The R chart is preferred to the σ chart because of its simplicity and versatility; and, unless there are compelling reasons to use the σ chart, the R chart is the method of choice.

TABLE 18-1. TESTS FOR LOCATING AND IDENTIFYING SPECIFIC TYPES OF ASSIGNABLE CAUSES

Type of Assignable Cause	Control Charts*		
	$\bar{X}$	R	σ
Gross Error (Blunder)	1	2	. .
Shift in Average	1	. .	. .
Shift in Variability	. .	1	2
Slow Fluctuation (Trend)	1	. .	. .
Fast Fluctuation (Cycle)	. .	1	2
Combination:			
(a) Production	1	2	. .
(b) Research	. .	. .	. .
Covariation	1	. .	. .

* The numeral 1 denotes the most useful type of chart; 2 denotes the next best; and, . . denote charts which are not appropriate for the particular cause.

Adapted with permission from *Industrial Quality Control*, Vol. IX, No. 3, (November, 1952) and No. 4, (January, 1953) from article entitled "How to Detect the Type of an Assignable Cause" by P. S. Olmstead.

18-3 APPLICATIONS OF CONTROL CHARTS

Table 18-2 is a summary table of factors for control charts for $\bar{X}$, R, and σ, when equal size samples are involved. Note carefully the footnote to Table 18-2, beginning "When using $s = \sqrt{\dfrac{\Sigma(X_i - \bar{X})^2}{n-1}}$ . . . ", because s is so defined in *this* Handbook. The last column of Table 18-2 gives values of $\sqrt{\dfrac{n-1}{n}}$ for convenience in using the Table factors with values of s.

The most explicit details of application to a variety of possible situations, e.g., to samples of unequal size, are given in the ASTM Manual[2]; in using that Manual, however, the reader again must be wary of the difference between the definition of σ given therein, and the definition of s given in this Handbook.

Actual examples of laboratory applications in the chemical field can be found in a series of comprehensive bibliographies published in *Analytical Chemistry*[3,4,5,6]. These four articles are excellent reviews that successively bring up-to-date the recent developments in statistical theory and statistical applications that are of interest in chemistry. Further, these bibliographies are divided by subject matter, and thus provide means for locating articles on control charts in the laboratory. They are not limited to control chart applications, however.

Industrial Quality Control[7], the monthly journal of the American Society for Quality Control, is the most comprehensive publication in this field.

For a special technique with ordnance examples, see Grubbs[8].

TABLE 18-2. FACTORS FOR COMPUTING 3-SIGMA CONTROL LIMITS

Number of Observations in Sample, n	Chart for Averages			Chart for Standard Deviations						Chart for Ranges							$\sqrt{\frac{n-1}{n}}$
	Factors for Control Limits			Factors for Central Line		Factors for Control Limits				Factors for Central Line		Factors for Control Limits					
	A	A_1	A_2	c_2	$1/c_2$	B_1	B_2	B_3	B_4	d_2	$1/d_2$	d_3	D_1	D_2	D_3	D_4	
2	2.121	3.760	1.880	0.5642	1.7725	0	1.843	0	3.267	1.128	0.8865	0.853	0	3.686	0	3.267	.70711
3	1.732	2.394	1.023	0.7236	1.3820	0	1.858	0	2.568	1.693	0.5907	0.888	0	4.358	0	2.575	.81650
4	1.500	1.880	0.729	0.7979	1.2533	0	1.808	0	2.266	2.059	0.4857	0.880	0	4.698	0	2.282	.86603
5	1.342	1.596	0.577	0.8407	1.1894	0	1.756	0	2.089	2.326	0.4299	0.864	0	4.918	0	2.115	.89443
6	1.225	1.410	0.483	0.8686	1.1512	0.026	1.711	0.030	1.970	2.534	0.3946	0.848	0	5.078	0	2.004	.91287
7	1.134	1.277	0.419	0.8882	1.1259	0.105	1.672	0.118	1.882	2.704	0.3698	0.833	0.205	5.203	0.076	1.924	.92582
8	1.061	1.175	0.373	0.9027	1.1078	0.167	1.638	0.185	1.815	2.847	0.3512	0.820	0.387	5.307	0.136	1.864	.93541
9	1.000	1.094	0.337	0.9139	1.0942	0.219	1.609	0.239	1.761	2.970	0.3367	0.808	0.546	5.394	0.184	1.816	.94281
10	0.949	1.028	0.308	0.9227	1.0837	0.262	1.584	0.284	1.716	3.078	0.3249	0.797	0.687	5.469	0.223	1.777	.94868
11	0.905	0.973	0.285	0.9300	1.0753	0.299	1.561	0.321	1.679	3.173	0.3152	0.787	0.812	5.534	0.256	1.744	.95346
12	0.866	0.925	0.266	0.9359	1.0684	0.331	1.541	0.354	1.646	3.258	0.3069	0.778	0.924	5.592	0.284	1.716	.95743
13	0.832	0.884	0.249	0.9410	1.0627	0.359	1.523	0.382	1.618	3.336	0.2998	0.770	1.026	5.646	0.308	1.692	.96077
14	0.802	0.848	0.235	0.9453	[illegible]	[illegible]	1.507	0.406	1.594	3.407	0.2935	0.762	1.121	5.693	0.329	1.671	.96362
15	0.775	0.816	0.223	0.9490	[illegible]	[illegible]	1.492	0.428	1.572	3.472	0.2880	0.755	1.207	5.737	0.348	1.652	.96609
16	0.750	0.788	0.212	0.9523	[illegible]	[illegible]	1.478	0.448	1.552	3.532	0.2831	0.749	1.285	5.779	0.364	1.636	.96825
17	0.728	0.762	0.203	0.9551	1.0470	0.445	1.465	0.466	1.534	3.588	0.2787	0.743	1.359	5.817	0.379	1.621	.97014
18	0.707	0.738	0.194	0.9576	1.0442	0.461	1.454	0.482	1.518	3.640	0.2747	0.738	1.426	5.854	0.392	1.608	.97183
19	0.688	0.717	0.187	0.9599	1.0418	0.477	1.443	0.497	1.503	3.689	0.2711	0.733	1.490	5.888	0.404	1.596	.97333
20	0.671	0.697	0.180	0.9619	1.0396	0.491	1.433	0.510	1.490	3.735	0.2677	0.729	1.548	5.922	0.414	1.586	.97468
21	0.655	0.679	0.173	0.9638	1.0376	0.504	1.424	0.523	1.477	3.778	0.2647	0.724	1.606	5.950	0.425	1.575	.97590
22	0.640	0.662	0.167	0.9655	1.0358	0.516	1.415	0.534	1.466	3.819	0.2618	0.720	1.659	5.979	0.434	1.566	.97701
23	0.626	0.647	0.162	0.9670	1.0342	0.527	1.407	0.545	1.455	3.858	0.2592	0.716	1.710	6.006	0.443	1.557	.97802
24	0.612	0.632	0.157	0.9684	1.0327	0.538	1.399	0.555	1.445	3.895	0.2567	0.712	1.759	6.031	0.452	1.548	.97895
25	0.600	0.619	0.153	0.9696	1.0313	0.548	1.392	0.565	1.435	3.931	0.2544	0.709	1.804	6.058	0.459	1.541	.97980

FORMULAS*

Purpose of Chart	Chart for	Central Line	3-Sigma Control Limits
For analyzing past data for control ($\bar{X}$, $\bar{\sigma}$, $\bar{R}$ are average values for the data being analyzed)	Averages	$\bar{X}$	$\bar{X} \pm A_1\bar{\sigma}$, or $\bar{X} \pm A_2\bar{R}$
	Standard deviations	$\bar{\sigma}$	$B_3\bar{\sigma}$ and $B_4\bar{\sigma}$
	Ranges	$\bar{R}$	$D_3\bar{R}$ and $D_4\bar{R}$
For controlling performance to standard values ($\bar{X}'$, σ', R_n' are selected standard values; $R_n' = d_2\sigma'$ for samples of size n)	Averages	$\bar{X}'$	$\bar{X}' \pm A\sigma'$, or $\bar{X}' \pm A_2R_n'$
	Standard deviations	$c_2\sigma'$	$B_1\sigma'$ and $B_2\sigma'$
	Ranges	$d_2\sigma'$, or R_n'	$D_1\sigma'$ and $D_2\sigma'$, or D_3R_n' and D_4R_n'

* When using $s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}$ for the standard deviation of a sample instead of $\sigma = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n}}$, one must make the following changes in the formulas for the central line and for the 3-sigma limits:

(1) Replace A_1 by $\sqrt{\frac{n-1}{n}}\,A_1$; replace $\bar{\sigma}$ by $\bar{s}$; make no change in B_3 *and* B_4;

(2) Replace c_2, B_1, B_2 by $\sqrt{\frac{n}{n-1}}\,c_2$, $\sqrt{\frac{n}{n-1}}\,B_1$ and $\sqrt{\frac{n}{n-1}}\,B_2$, respectively.

REFERENCES

1. P. S. Olmstead, "How to Detect the Type of an Assignable Cause," Parts I and II, *Industrial Quality Control*, Vol. IX, Nos. 3, pp. 32-38 (November 1952) and 4, pp. 22-32 (January 1953).
2. *ASTM Manual on Quality Control of Materials*, American Society for Testing Materials, 1916 Race St., Philadelphia 3, Pa., 1951.
3. G. Wernimont, "Statistics Applied to Analysis", *Analytical Chemistry*, Vol. 21, p. 115, 1949.
4. R. J. Hader and W. J. Youden, "Experimental Statistics", *Analytical Chemistry*, Vol. 24, p. 120, 1952.
5. J. Mandel and F. J. Linnig, "Statistical Methods in Chemistry", *Analytical Chemistry*, Vol. 28, p. 770, 1956.
6. Ibid., Vol. 30, p. 739, 1958.
7. *Industrial Quality Control*, Monthly Journal, The American Society for Quality Control, 161 W. Wisconsin Ave., Milwaukee 3, Wis.
8. F. E. Grubbs, "The Difference Control Chart with Example of Its Use", *Industrial Quality Control*, Vol. III, No. 1, pp. 22-25, July 1946.

SOME RECOMMENDED GENERAL TEXTS

American Standard Z 1.3-1958, *Control Chart Method of Controlling Quality During Production*, available from American Standards Association, Inc., 10 East Fortieth St., New York 16, N. Y.

D. J. Cowden, *Statistical Methods in Quality Control*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1957.

A. J. Duncan, *Quality Control and Industrial Statistics*, Richard D. Irwin, Inc., Chicago, Ill., 1952.

E. L. Grant, *Statistical Quality Control*, (2d edition), McGraw-Hill Book Co., New York, N. Y., 1952.

W. A. Shewhart, *Economic Control of Quality of Manufactured Product*, Van Nostrand Inc., New York, N. Y., 1931. (The first and still a very significant book in the field).

CHAPTER 19

STATISTICAL TECHNIQUES FOR ANALYZING EXTREME-VALUE DATA*

19-1 EXTREME-VALUE DISTRIBUTIONS

Classical applications of statistical methods, which frequently concern average values and other quantities following the symmetrical normal distribution, are inadequate when the quantity of interest is the largest or the smallest in a set of magnitudes. Applications of the techniques described in this Chapter already have been made in a number of fields. Meteorological phenomena that involve extreme pressures, temperatures, rainfalls, wind velocities, etc., have been treated by extreme-value techniques. The techniques are also applicable in the study of floods and droughts.

Other examples of extreme-value problems occur in the fracturing of metals, textiles, and other materials under applied force, and in fatigue phenomena. In these instances, the observed strength of a specimen often differs from the calculated strength, and depends, among other things, upon the length and volume. An explanation is to be found in the existence of weakening flaws assumed to be distributed at random in the body and assumed not to influence one another in any way. The observed strength is determined by the strength of the weakest region — just as no chain is stronger than its weakest link. Thus, it is apparent that whenever extreme observations are encountered it will pay to consider the use of extreme-value techniques.

19-2 USE OF EXTREME-VALUE TECHNIQUES

19-2.1 LARGEST VALUES

A simplified account is given here. Primary sources for the detailed theory and methods are References 1, 2, 3, which also contain extensive bibliographies. References 4 through 10, also given at the end of this Chapter, provide additional information and examples of applications.

Figure 19-1 illustrates the frequency form of a typical curve for the distribution of largest observations.

The curve in Figure 19-1 is tl derivative of the function

$$\Phi(y) = \exp[-\exp(\quad)].$$

Unlike the normal distribution, this curve is skewed, with its maximum to the left of the mean and the longer of its tails extending to the right. The outstanding feature of such a distribution is that very large values are much more likely to occur than are very small values. This agrees with common experience. Very low maximum values are most unusual, while very high ones do occur occasionally. Theoretical considerations lead to a curve of this nature, called the *distribution of largest values* or the *extreme-value distribution.*

In using the extreme-value method, all the observed maxima, such as the largest wind velocity observed in each year during a fifty-

* Adapted with permission from *The American Statistician*, Vol. 8, No. 5, December 1954, from article entitled "Some Applications of Extreme-Value Methods" by E. J. Gumbel and J. Lieblein; and, from *National Bureau of Standards Technical News Bulletin 38*, No. 2, pp. 29-31, February 1954, from article entitled "Extreme-Value Methods for Engineering Problems".

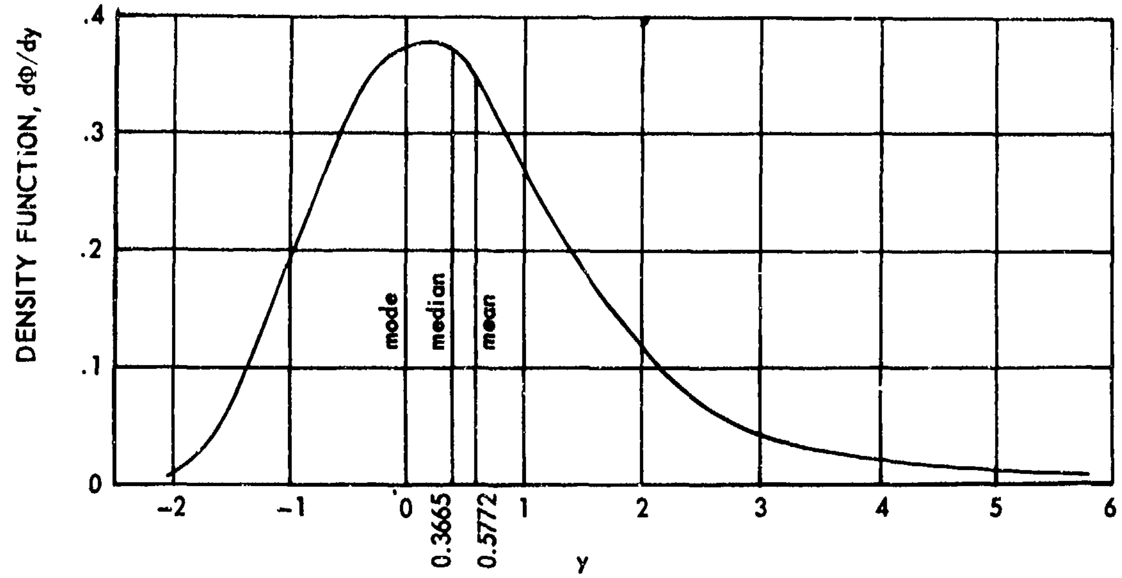


Figure 19-1. Theoretical distribution of largest values.

Adapted with permission from *The American Statistician*, Vol. 8, No. 5, December 1954, from article entitled "Some Applications of Extreme-Value Methods" by E. J. Gumbel and J. Lieblein.

year period, are first ranked in order of size from the smallest to the largest,

$$X_1 \leq X_2 \leq \ldots \leq X_i \leq \ldots \leq X_n .$$

A plotting position (X_i, P_i) is obtained for each observation by associating with X_i the probability coordinate $P_i = i/(n + 1)$, where i is the rank of the observation, counting from the smallest. The data are plotted on a special graph paper, called extreme-value probability paper*, designed so that the "ideal" extreme-value distribution will plot exactly as a straight line. Consequently, the closeness of the plotted points to a straight line is an indication of how well the data fit the theory.

* Extreme-value probability paper may be obtained from three sources: (a) U. S. Department of Commerce, Weather Bureau; (b) Environmental Protection Section, Research and Development Branch, Military Planning Division, Office of the Quartermaster General; (c) Technical and Engineering Aids for Management, 104 Belrose Ave., Lowell, Mass.

Extreme-value probability paper has a uniform scale along one axis, usually the vertical, which is used for the observed values as shown in Figure 19-2. The horizontal axis then serves as the probability scale, and is marked according to the doubly-exponential formula. Thus, in Figure 19-2, the space between 0.01 and 0.5 is much less than the space between 0.5 and 0.99. The limiting values zero and one are never reached, as is true of any probability paper designed for an unlimited variate.

An extreme-value plot (Figure 19-2) of the maximum atmospheric pressures in Bergen, Norway, for the period between 1857 and 1926, showed by inspection that the observed data satisfactorily fitted the theory. Fitting the line by eye may be sufficient. Details of fitting a computed line are given in Gumbel.[1] From the fitted straight line, it is possible to predict, for example, that a pressure of 793 mm corresponds to a probability of 0.994; that is, pres-

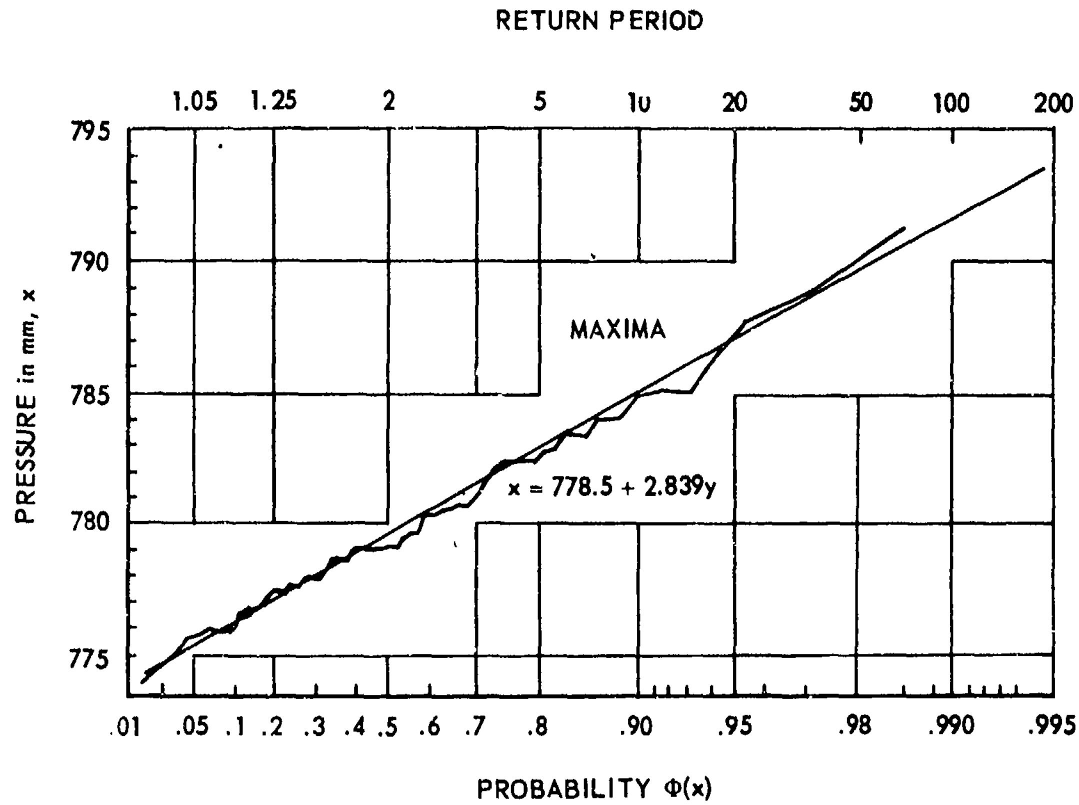


Figure 19-2. Annual maxima of atmospheric pressure, Bergen, Norway, 1857-1926.

Adapted with permission from *The American Statistician*, Vol. 8, No. 5, December 1954, from article entitled "... Applications of Extreme-Value Methods" by E. J. Gumbel and J. Lieblein.

sures of this magnitude have less than one chance in 100 of being exceeded in any particular year.

In studies of the normal acceleration increments experienced by an airplane flying through gusty air, see Gumbel and Carlson,[4] page 394, an instrument was employed that indicated only the maximum shocks. Thus, only one maximum value was obtained from a single flight. A plot representing 26 flights of the same aircraft indicated that the probability that the largest recorded gust will not be exceeded in any other flight was 0.96; i.e., a chance of four in 100 of encountering a gust more severe than any recorded. A more recent study, Lieblein,[5] presents refinements especially adapted to very small samples of extreme data, and also to larger samples where it is necessary to obtain the greatest amount of information from a limited set of costly data.

19-2.2 SMALLEST VALUES

Extreme-value theory can also be used to study the smallest observations, since the corresponding limiting distribution is simply related to the distribution of largest values. The steps in applying the "smallest-value" theory are very similar to those for the largest-value case. For example, engineers have long been interested in the problem of predicting the tensile strength of a bar or specimen of homogeneous material. One approach is to regard the specimen as being composed of a large number of pieces of very short length. The tensile strength of the entire specimen is limited by the strength of the weakest of these small pieces. Thus, the tensile strength at which the entire specimen will fail is a smallest-value phenomenon. The smallest-value approach can be used even though the number and individual strengths of the "small pieces" are unknown.

This method has been applied with considerable success by Kase[6] in studying the tensile testing of rubber. Using 200 specimens obtained so as to assure as much homogeneity as possible, he found that the observed distribution of their tensile strengths could be fitted remarkably well by the extreme-value distribution for smallest values. The fitted curve given by this data indicates that one-half of a test group of specimens may be expected to break under a tensile stress of 105 kg./cm.^{2} or more, while only one in 1,000 will survive a stress exceeding 126 kg./cm.^{2}.

Other examples of applications are given by Epstein and Brooks[7] and by Freudenthal and Gumbel[8][9].

19-2.3 MISSING OBSERVATIONS

It has been found that fatigue life of specimens under fixed stress can be treated in the same manner as tensile strength — by using the theory of smallest values. An extensive application of this method is given in Lieblein and Zelen[10].

In such cases, tests may be stopped before all specimens have failed. This results in a sample from which some observations are missing — a "censored" sample. Methods for handling such data are included in Lieblein and Zelen[10].

REFERENCES

1. E. J. Gumbel, *Statistical Theory of Extreme Values and Some Practical Applications*, National Bureau of Standards Applied Mathematics Series No. 33, U. S. Government Printing Office, Washington, D. C., 1954.

2. National Bureau of Standards, *Probability Tables for the Analysis of Extreme-Value Data*, Applied Mathematics Series No. 22, U. S. Government Printing Office, Washington, D. C., 1953.

3. E. J. Gumbel, *Statistics of Extremes*, Columbia University Press, New York, N. Y., 1958.

4. E. J. Gumbel and P. G. Carlson, "Extreme Values in Aeronautics", *Journal of the Aeronautical Sciences*, Vol. 21, No. 6, pp. 389-398, June 1954.

5. J. Lieblein, "A New Method for Analyzing Extreme-Value Data", *NACA Technical Note 3053*, National Advisory Committee for Aeronautics, January 1954.

6. S. Kase, "A Theoretical Analysis of the Distribution of Tensile Strength of Vulcanized Rubber," *Journal of Polymer Science*, Vol. 11, No. 5, pp. 425-431, November 1953.

7. B. Epstein and H. Brooks, "The Theory of Extreme Values and Its Implications in the Study of the Dielectric Strength of Paper Capacitors", *Journal of Applied Physics*, Vol. 19, pp. 544-550, 1948.

8. A. M. Freudenthal and E. J. Gumbel, "On the Statistical Interpretation of Fatigue Tests", *Proceedings of the Royal Society A*, Vol. 216, pp. 309-332, 1953.

9. A. M. Freudenthal and E. J. Gumbel, "Minimum Life in Fatigue", *Journal of the American Statistical Association*, Vol. 49, pp. 575-597, September 1954.

10. J. Lieblein and M. Zelen, "Statistical Investigation of the Fatigue Life of Deep-Groove Ball Bearings", *Journal of Research of the National Bureau of Standards*, Vol. 57, No. 5, pp. 273-316, November 1956.

CHAPTER 20

THE USE OF TRANSFORMATIONS

20-1 GENERAL REMARKS ON THE NEED FOR TRANSFORMATIONS

The scale on which a property is usually measured (that is, the units in which it is ordinarily expressed) may not be the most suitable for statistical analysis and interpretation. Statistical techniques are always based on assumptions. The validity of results obtained through their use in practice always depends, sometimes critically, on the assumed conditions being met, at least to a sufficient degree of approximation. Essentially all of the standard techniques for the statistical analysis and interpretation of measurement data (e.g., those given in AMCP 706-110,Section 1, Chapters 1 through 6) are based upon assumed *normality* of the underlying distribution involved; and many (e.g., the majority of those considered in Chapters 5 and 6) also require (at least approximate) *equality of variances* from group to group. Furthermore, the analysis-of-variance tests considered in AMCP 706-112,Section 3, depend not only on normality and equality of variances among subgroups, but also on *additivity* of the "effects" that characterize real differences of interest among the materials, processes, or products under consideration; see Eisenhart.[1]

Real-life data do not always conform to the conditions required for the strict, or even approximate, validity of otherwise appropriate techniques of statistical analysis. When this is the case, a transformation (change of scale) applied to the raw data may put the data in such form that the appropriate conventional analysis can be performed validly. Bartlett[2] provides a good general survey of the practical aspects of transformations, together with a fairly complete bibliography of the subject to 1947.

20-2 NORMALITY AND NORMALIZING TRANSFORMATIONS

20-2.1 IMPORTANCE OF NORMALITY

The dependence of many standard statistical techniques on normality of the underlying distribution is twofold. First, standard statistical techniques are in the main based on the sample mean $\bar{X}$, and the sample estimate s of the population standard deviation. A normal distribution is completely determined by its mean m and its standard deviation σ; and in sampling from a normal distribution, $\bar{X}$ and s together summarize all of the information available in the sample about the parent distribution. This 100% efficiency of $\bar{X}$ and s in samples from a normal distribution does not carry over to non-normal distributions. Consequently, if the population distribution of a characteristic of interest is markedly non-normal, confidence intervals for the population mean m and standard deviation σ based on $\bar{X}$ and s will tend to be wider, and tests of hypotheses regarding m or σ will have less power, than those based on the particular functions of the sample values that are the efficient estimators of the location and dispersion parameters of the non-normal distribution concerned. In other words, use of $\bar{X}$ and s as sample measures of the location and dispersion characteristics of a population distribution may result in an intrinsic loss of efficiency in the case of markedly non-normal distributions, even if the correct sampling distributions of χ^2, t, F, etc., appropriate to the non-normal distribution concerned are employed.

Second, the customary tables of percentage points of χ^2, t, F, and of factors for confidence intervals, tolerance limits, and so forth, are based on the assumption of sampling from a normal distribution. These percentage points, tolerance-limit factors, and so forth, are not strictly valid when sampling from non-normal distributions. The distribution of s^2, which is identically that of $\chi^2\sigma^2/\nu$ for ν degrees of freedom in the case of sampling from a normal distribution, is especially sensitive to departures from normality. Consequently, the actual significance levels, confidence coefficients, etc., associated with the procedures of Chapter 4 may differ somewhat from their nominal values when sampling from only moderately non-normal material is involved. Fortunately, the percentage points of t- and F-tests of hypotheses about means are not so sensitive to departures from normality, so that the standard tests of hypotheses about, and confidence intervals for, population means will be valid to a good approximation for moderately non-normal populations — but there may be some loss of efficiency, as noted above.

20-2.2 NORMALIZATION BY AVERAGING

Many physical measurement processes produce approximately normally-distributed data; some do not. Even when measurement errors are approximately normally distributed, sampling of a material, product, or process may be involved, and the distribution of the characteristic of interest in the sampled population may be definitely non-normal — or, at least, it may be considered risky to assume normality. In such cases, especially when the basic measurements are plentiful or easy to obtain in large numbers, an effective *normalization* almost always can be achieved — except for extremely non-normal distributions — if the questions of interest with respect to the population concerned can be rephrased in terms of the parameters of the corresponding sampling distribution of the arithmetic means of random samples of size four or more. This normalizational trick is of extremely wide applicability; but results, of course, in a substantial reduction in the number of observations available for statistical analysis. Consequently, it should not be applied when the basic measurements themselves are few in number and costly to obtain. In such cases, if assumption of normality of the population distribution of the basic observations is considered risky, or definitely is known to be false, then we may take recourse in available distribution-free techniques; see Chapter 16.

20-2.3 NORMALIZING TRANSFORMATIONS

If we know from theoretical considerations or previous experience that some simple transformation will approximately normalize the particular kind of data in hand, then, both for convenience and in the interest of efficiency, we may prefer to use normal-based standard techniques on the transformed data, rather than use distribution-free techniques on the data in their original form. For example, certain kinds of data are quite definitely known to be approximately normal in logs, and the use of a log transformation in these cases may become routine. Indeed, this transformation is the subject of an entire book which is devoted to its theoretical and empirical bases, and its uses and usefulness in a wide variety of situations; see Aitchison and Brown.[3]

Table 20-1 gives a selection of transformations that are capable of normalizing a wide variety of non-normal types. They are arranged in groups according as the range of variation of the original variable X is from 0 to ∞, from 0 to 1, or from -1 to $+1$. Their "normalizing power" is exemplified in Figure 20-1. For the theoretical bases of these and other normalizing transformations, the advanced reader is referred to the papers of Curtiss[4] and Johnson.[5]

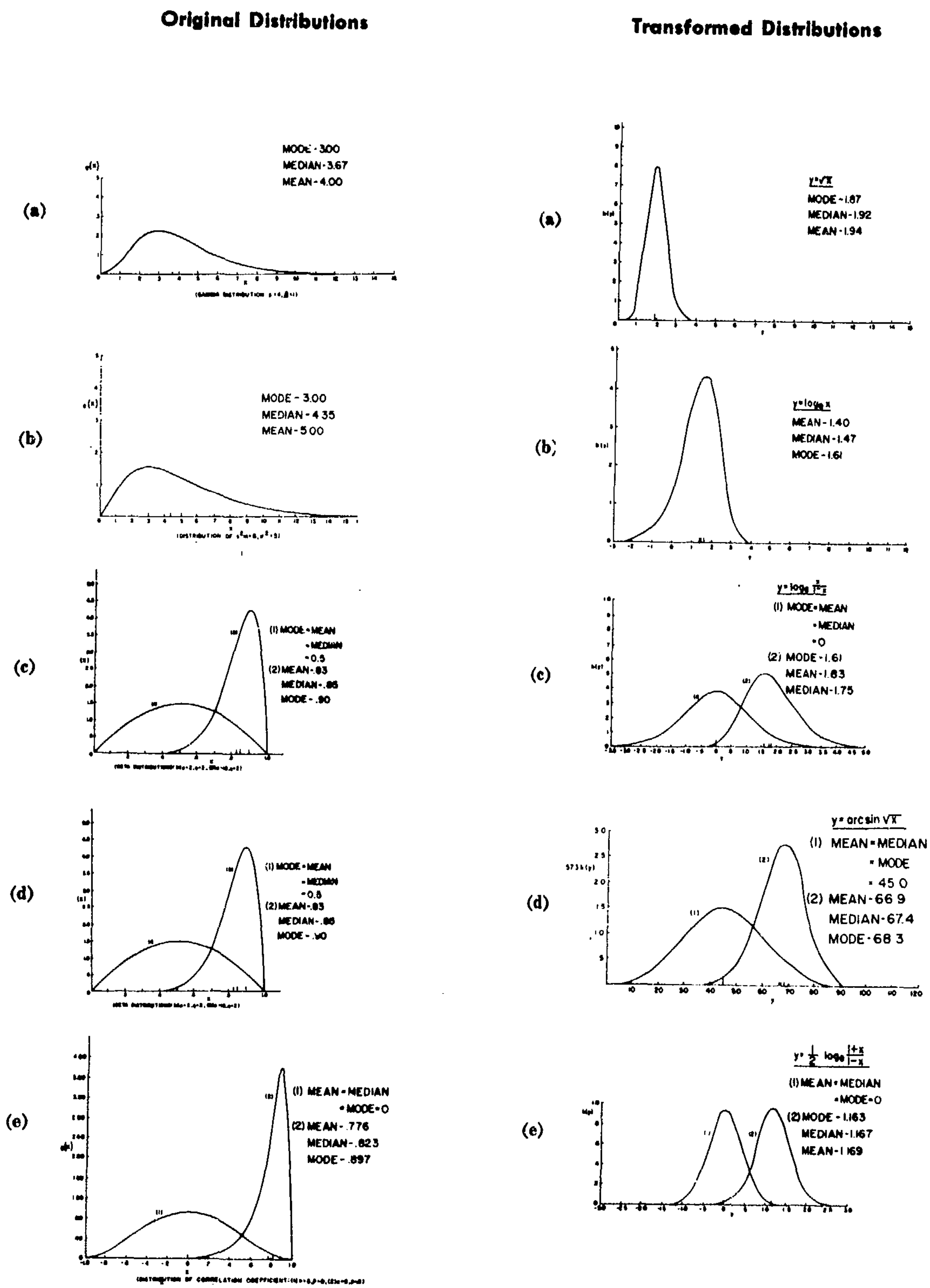


Figure 20-1. Normalizing effect of some frequently used transformations.

20-3 INEQUALITY OF VARIANCES, AND VARIANCE-STABILIZING TRANSFORMATIONS

20-3.1 IMPORTANCE OF EQUALITY OF VARIANCES

Many standard statistical techniques for the analysis and comparison of two or more materials, products, or processes with respect to average performance depend on equality of variability within groups. When the magnitude of the common within-groups variance σ^2 is unknown, it is customary (as in Procedures of Paragraphs 3-3.1.1, 3-3.2.1, and 3-4) to combine the sample evidence on variability of performance within the respective groups, to obtain a pooled estimate of σ^2. The advantages of pooling are: the resultant pooled estimate s^2 is a more precise estimate of σ^2 than is available from data of any of the individual groups alone; it leads to narrower confidence intervals for each of the individual group means, and for differences between them; and hence, it leads to more powerful tests of significance for differences between group means. If, however, the assumption of equality of within-group variances is false, then the resultant pooled s^2 does not provide a valid estimate of the standard error of any of the group averages, or of any of the differences between them. When marked inequalities exist among the true within-group variances, the standard errors of individual group averages and of differences between them, derived from a pooled s^2, may be far from the true values; and confidence intervals and tests of significance based on the pooled s^2 may be seriously distorted.

Thus, in Chapter 3, we emphasized that the standard t-tests for the comparison of averages of two groups of unpaired observations (Paragraphs 3-3.1.1 and 3-3.2.1) are based on the assumption of equal variances within the two groups. Furthermore, we noted that if the two samples involved are of equal size, or of approximately equal size, then the significance levels of the two sided t-test of the difference of two means (Paragraph 3-3.1.1) will not be seriously increased (Figure 3-9, curve (A)); but the power of the test may be somewhat lessened if the two variances are markedly unequal. Similarly, two-sided confidence intervals derived from t for the difference between the two population means will tend to be somewhat narrower than if proper allowance were made for the inequality of the variances, but the effective confidence coefficient will not be seriously less than the value intended. These remarks carry over without change to one-sided t-tests (Paragraph 3-3.2.1) and to the corresponding one-sided confidence intervals. In other words, the comparison of averages of two groups by means of the standard two sample t-test procedures and associated confidence intervals results only in some loss of efficiency when the samples from the two groups are of equal size, and the reduction in efficiency will be comparatively slight unless the two variances are markedly different.

In contrast, if the samples from the two groups differ appreciably in size, then not only may the significance levels of standard two-sample t-tests be seriously affected (Figure 3-9, curve (B)) but their power (i.e., the entire OC curve) also may be altered considerably, especially if the smallest sample comes from the group having the larger variance. Hence, in the case of samples of unequal size, inequality of variances may invalidate not only a standard two-sample t-test for comparison of averages, but also the associated confidence-interval procedures for estimating the difference between the corresponding population means.

The foregoing remarks carry over without modification to the Studentized-range techniques given in Paragraph 3-4 for the comparison of averages of several groups, and in AMCP 706-112, Section 3, Chapters 12 and 13, for the comparison of averages and groups of averages in complex and more specialized forms of comparative experiments. In all of these cases, if the true within-group variances differ appreciably from one group to another (or from subgroups to subgroups), there ordinarily will be a loss of efficiency in the estimation of, say, product means, or treatment differences. Similarly, there will be a loss of power in tests of significance. If the samples from the respective groups are of unequal sizes and the true within-group variances are markedly unequal, these losses may be substantial. Some of the estimates of group means and differences between group means may have much smaller or much larger standard errors than others, so that pair-wise t-tests, or Studentized-range tests, derived from a pooled standard-deviation estimate s may correspond to significance levels far from those intended; and the actual effective confidence coefficients associated with the corresponding confidence intervals may differ substantially from one another, and from their nominal values.

20-3.2 TYPES OF VARIANCE INHOMOGENEITY

The situations in which variance inhomogeneity may present a problem can be divided into two types:

(a) Situations in which there is a functional dependence of the variance of an observation on the mean of the group to which it belongs. Functional dependence of the variance of an observation on its mean or expected value is an intrinsic characteristic of many non-normal distributions. The second column of Table 20-1 gives some specific examples. Or, it may be a basic property of the phenomena under investigation quite apart from the form of the underlying distribution involved. Thus, in studies of various types of "growth" phenomena, the amount of variation present at any given stage of the "growth," as measured by the standard deviation of observations at that stage, is apt to be proportional to the average size characteristic of that stage.

TABLE 20-1. SOME FREQUENTLY USED TRANSFORMATIONS

Transformation $Y = f(X)$	Appropriate Situation: Range of Variable	Appropriate Situation: Characteristics of Distribution	Approximate Variance on Transformed Scale $\sigma_Y^2 \simeq [f'(m)]^2 \sigma_X^2$	Examples of Appropriate Distributions: Distribution and Its Parameters	Examples of Appropriate Distributions: Approximate Variance on Transformed Scale
$\sqrt{X}$	$0 \le X \le \infty$	Variance proportional to the mean Var $= \lambda^2 \cdot$ mean	$\lambda^2/4$	*Continuous* Gamma distributions Mean $= p\theta$ Var $= p\theta^2$	$1/4p$
				Discrete Poisson distribution* $X = 0, 1, 2, \ldots$ Mean $= m$ Var $= m$	$1/4$
$\log_e X$ or $\log_{10} X$	$0 \le X \le \infty$	Standard deviation proportional to mean Var $= \lambda^2 (\text{mean})^2$	For $\log_e$, λ^2 For $\log_{10}$, $0.189\,\lambda^2$	Distributions of s^2 in samples of size n for normal distribution Mean $= \sigma^2$ Var $= \frac{2\sigma^4}{n-1} = \left(\frac{2}{n-1}\right)(\text{mean})^2$	For $\log_e$, $\frac{2}{n-1}$ For $\log_{10}$, $\frac{0.377}{n-1}$
$\log_e \frac{X}{1-X} = 2 \tanh^{-1}(2X-1)$ or $\log_{10} \frac{X}{1-X}$	$0 \le X \le 1$	*Type A* Mean $= m$ Var $= \lambda^2 m (1-m)$	For $\log_e$, $\frac{\lambda^2}{m(1-m)}$ For $\log_{10}$, $\frac{0.189\,\lambda^2}{m(1-m)}$	Beta distributions Mean $= \frac{p}{p+q}$ Var $= \frac{pq}{(p+q)^2 (p+q+1)}$	
		Type B Mean $= m$ Var $= \lambda^2 m^2 (1-m)^2$	For $\log_e$, λ^2 For $\log_{10}$, $0.189\,\lambda^2$	Empirical	
$\arcsin \sqrt{X}$ (radians) or $\arcsin \sqrt{X}$ (degrees)	$0 \le X \le 1$	Mean $= m$ Var $= \lambda^2 m (1-m)$	For radians, $\lambda^2/4$ For degrees, $821\,\lambda^2$	*Continuous* Beta distributions	For radians, $\frac{1}{4(p+q+1)}$ For degrees, $\frac{821}{p+q+1}$
				Discrete Binomial distributions‡ $X = 0, 1/n, 2/n, \ldots, n/n$ Mean $= p$ Var $= p(1-p)/n$	For radians, $1/4n$ For degrees, $821/n$
$\frac{1}{2} \log_e \frac{1+X}{1-X} = \tanh^{-1} X$ or $\log_{10} \frac{1+X}{1-X}$	$-1 \le X \le +1$	Var $= \lambda^2 [1 - (\text{mean})^2]^2$	For $\frac{1}{2} \log_e$, λ^2 For $\log_{10}$, $0.754\,\lambda^2$	Distribution of correlation coefficient r in samples from a normal distribution Mean $= \rho\left[1 - \frac{1-\rho^2}{2n} + \ldots\right]$ Var $= (1-\rho^2)^2 \left[\frac{1}{n} + \frac{11\rho^2}{2n^2} - \ldots\right]$	For $\frac{1}{2} \log_e$, $\frac{1}{n-3}$ For $\log_{10}$, $\frac{0.754}{n-3}$
Probit	Sensitivity Testing			See ORDP 20-111, Chapter 10	

* For $1 < m < 10$, use the Freeman-Tukey modification $Y = \frac{1}{2}(\sqrt{X} + \sqrt{X+1})$.
† Use $\log_e (X+1)$ or $\log_{10} (X+1)$ to avoid difficulties with zeros in the data.
‡ For greater accuracy, use Bartlett's modification,
for $X = 0$, $Y = \arcsin \sqrt{1/4n}$
and $X = 1$, $Y = \arcsin \sqrt{(4n-1)/4n}$.

(b) Situations in which there is present incidental desultory heterogeneity of variance, arising from inadequate control of conditions or procedure; from differences or shortcomings of equipment or personnel; from use of inhomogeneous material or inadequate sampling methods; or from other disturbing features (e.g., partial failure of one or more of the products or treatments) that tend to produce less, or greater, variability among observations in some groups than in others in an irregular manner.

Situations of the first type, in which the variance inhomogeneity present is simply the consequence of a functional dependence of the variance of an observation on its mean or expected value, are most easily handled statistically by employing an appropriate variance-stabilizing transformation. Details are given in Paragraph 20-3.3. Statistical analyses of data arising from the second irregular type of variance heterogeneity should be left to experts. Variance-stabilizing transformations are of little or no help in such situations. Helpful advice, illustrated by worked examples, can be found in two papers by Cochran.[6, 7] Recourse usually must be made to subdividing the experimental observations into approximately homogenous subgroups; or to omission of parts of the experiment that have yielded data very different from the rest. An overall analysis may be impossible. Combination of the pertinent evidence from the respective subdivisions of the data may involve complex weighting and laborious arithmetic. Various procedures for the combination of evidence from different experiments, or from separately analyzed parts of a single experiment, have been examined and evaluated in a later paper by Cochran.[8] Irregular heterogeneity of variance should be avoided whenever possible, by adequate design of experiments and careful attention to the control of conditions, procedures, etc.

20-3.3 VARIANCE-STABILIZING TRANSFORMATIONS

When experimentally determined values $X_1, X_2, \ldots$, are such that their variances $\sigma^2_{X_i}$ are functionally *dependent* on their mean values m_{X_i} in accordance with a common functional relationship, say

$$\sigma^2_{X_i} = g\,(m_{X_i}),\ (i = 1, 2, \ldots), \tag{20-1}$$

then we may gain the advantages of variance homogeneity in the statistical analysis of such data by replacing the original values $X_1, X_2, \ldots$, by transformed values $Y_1 = f\,(X_1)$, $Y_2 = f\,(X_2), \ldots$, whose variances $\sigma^2_{Y_i}$ are (at least, to a good approximation) functionally *independent* of their mean values m_{Y_i}. Five such variance-stabilizing transformations $Y = f\,(X)$ are given in the first column of Table 20-1; the "situations" (i.e., the range of X and the form of the function $g\,(m)$ in equation (20-1)) for which each is appropriate* are indicated in the second column; and the third column shows the approximate variances of the corresponding transformed values Y, as given by the approximate formula

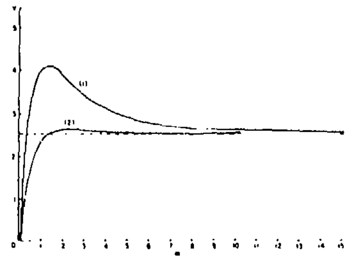
$$\sigma^2_Y \simeq [f'\,(m)]\,\sigma^2_X, \tag{20-2}$$

where $f'\,(m)$ denotes the derivative of the function $y = f\,(X)$ evaluated at $X = m$, the mean value of the original variable X.

Figure 20-2 presents comparisons of the actual values of the variances σ^2_Y of the transformed values Y and the corresponding approximate values given by formula (20-2), for four of the transformations listed in Table 20-1.

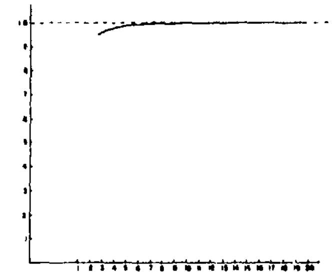
* The third transformation in the first column of Table 20-1, $\log \frac{X}{1-X}$, is variance-stabilizing only for "situations" of type B.

Panel (1)



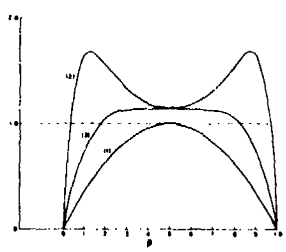
Dependence of the variances of two functions of a sample value X from a Poisson distribution on the Poisson parameter, m. (1) Variance of $\sqrt{X}$; (2) Variance of $\frac{1}{2}\left\{\sqrt{X}+\sqrt{X+1}\right\}$.

Panel (2)



The ratio of the variance of $\log_e s^2$ to its approximate value $2/(n-1)$ in samples of size n from a normal distribution.

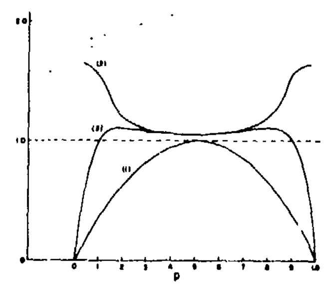
Panel (3)



Dependence of the variances of three functions of a sample proportion X/n on the population proportion p when the *sample size is 10*. (1) 40 Var (X/n); (2) 40 Var $(\sin^{-1}\sqrt{X/n})$; (3) 40 Var (φ_β), where

$$\varphi_\beta = \begin{cases} \sin^{-1}\sqrt{1/4\,n} & \text{for } X = 0 \\ \sin^{-1}\sqrt{X/n} & \text{for } X = 1, 2, \ldots, n-1 \\ \sin^{-1}\sqrt{(4\,n-1)/4\,n} & \text{for } X = n \end{cases}$$

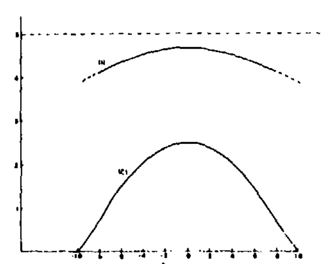
Panel (4)



Dependence of the variances of three functions of a sample proportion X/n on the population proportion p when the *sample size is 20*. (1) 80 Var (X/n); (2) 80 Var $(\sin^{-1}\sqrt{X/n})$; (3) 80 Var (φ_β), where

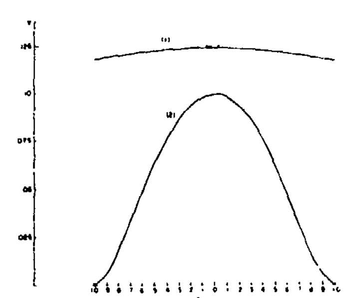
$$\varphi_\beta = \begin{cases} \sin^{-1}\sqrt{1/4\,n} & \text{for } X = 0 \\ \sin^{-1}\sqrt{X/n} & \text{for } X = 1, 2, \ldots, n-1 \\ \sin^{-1}\sqrt{(4\,n-1} & \text{for } X = n \end{cases}$$

Panel (5)



Dependence of the variance of the sample correlation coefficient r and of the variance of the transformation $z' = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$ on the true correlation coefficient ρ for *sample size* $n = 5$. (1) Variance of z'; (2) Variance of r.

Panel (6)



Dependence of the variance of the sample correlation coefficient r and of the variance of the transformation $z' = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$ on the true correlation coefficient ρ for *sample size* $n = 11$. (1) Variance of z'; (2) Variance of r.

Figure 20-2. Variance-stabilizing effect of some frequently used transformations.

The logari transformation log s^2 is "variance-stabilizing" for all values of n, since the variance of log s is functionally independent of its mean for all values of n; and, as is evident from panel 2 of Figure 20-2, the variance of log s^2 is close to its limiting value $\frac{2}{n-1}$ for all values of $n \geq 5$, say. For further details on this transformation, see Bartlett and Kendall.[9]

The other four transformations depicted in Figure 20-2 are variance-stabilizing (to a good approximation at least), only for *favorable* combinations of the parameters concerned. Thus, in the case of the Poisson distribution (panel 1), we see that the variance of $\sqrt{X}$ is independent of m_X to a good approximation only for $m \geq 10$, say; but the variance of the more sophisticated transformation $\frac{1}{2}(\sqrt{X} + \sqrt{X+1})$, devised by Freeman and Tukey[10], is nearly constant for $n \geq 3$, say. A table to facilitate the use of this transformation has been published by Mosteller *et al*[11, 12]. Similarly, for the binomial distribution: from panel 3 we see that when $n = 10$, the variance of arcsin $\sqrt{X/n}$ is no more stable as p ranges from 0 to 1, than is the variance of X/n itself; but *with Bartlett's modifications*[13], [14] *for $X = 0$ and $X = 1$*, the variance is essentially constant $\left(\text{at } \frac{1.13}{4n}\right)$ from $p = 0.25$ to $p = 0.75$. On the other hand, when $n = 20$ (panel 4), the variance of the unmodified transformation is nearly constant from $p = 0.25$ to $p = 0.75$, so that the unmodified transformation is quite adequate for this range of p. However, by adopting Bartlett's modifications, the range of variance constancy $\left(\text{at } \frac{1.06}{4n}\right)$ can be extended to $p = 0.12$ and $p = 0.88$. When $n = 30$, the unmodified transformation is adequate from $p = 0.18$ to $p = 0.82$, and with Bartlett's modifications, nearly constant variance $\left(\text{at } \frac{1.035}{4n}\right)$ is achieved from $p = 0.08$ to $p = 0.92$. Finally, panels 5 and 6 show the variance-stabilizing power of the log $\frac{1+r}{1-r}$ transformation of the correlation coefficient r, due to Fisher,[15] for $n = 5$ and $n = 10$.

Figure 20-2 and the foregoing discussion serve to bring out a very important feature of variance stabilizing transformations: over any range of *favorable circumstances* for which a particular variance-stabilizing transformation Y has an essentially constant and known variance σ_Y^2, we also have, in addition to the advantages of variance constancy, all of the attendant advantages of "σ-known" techniques. However, in practice, before proceeding on the assumption that σ_Y^2 has a particular theoretical value, we should always evaluate an estimate of σ_Y^2, say s_Y^2, from the data on hand, and check to see whether s_Y^2 is consistent with the presumed theoretical value of σ_Y^2. If it is, then "σ-known" techniques should be used in the interest of greater efficiency. On the other hand, if the magnitude of s_Y^2 indicates that the effective value of σ_Y^2 is substantially greater than its theoretical value, then "σ-unknown" techniques, based on s_Y, must be used. In such cases, the excess of s_Y^2 over the theoretical value of σ_Y^2 indicates the amount of additional variation present in the data, which, in principle at least, could be eliminated in future experiments of the same kind by improved experiment design and measurement-error control.

20-4 LINEARITY, ADDITIVITY, AND ASSOCIATED TRANSFORMATIONS

20-4.1 DEFINITION AND IMPORTANCE OF LINEARITY AND ADDITIVITY

Experimental data are much easier to interpret when the effects of the variables concerned are *linear* and *additive*.

When only a single independent variable x is involved, then *linearity* of the phenomena under investigation means that the response y corresponding to input x can be expressed in the form

$$y = \beta_0 + \beta_1 x \tag{20-3}$$

when x and y are expressed on appropriate scales. Equation (20-3) is the equation of a straight line in the x, y-plane. The analysis and interpretation of such linear relationships derived from experimental data are considered in detail in Chapter 5.

In the case of two independent or input variables, say x and z, if the dependence of the response y on these two variables is of the form

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz \\ &= (\beta_0 + \beta_2 z) + (\beta_1 + \beta_3 z)x \\ &= (\beta_0 + \beta_1 x) + (\beta_2 + \beta_3 x)z \end{aligned} \tag{20-4}$$

then clearly the response y depends linearly on x for fixed values of z, and linearly on z for fixed values of x; but the effect of changes in x and z will be additive if and only if the cross product term is missing (i.e., $\beta_3 = 0$). Only in this case will a given change in x, say δx, produce the same change in y regardless of the value of z, and a given change in z, say δz, produce the same change in y regardless of the value of x; and hence together produce the same total change in y, irrespective of the "starting values" of x and z. In other words, for ***linearity and additivity*** in the case of two independent variables, the response surface must be of the form

$$y = \beta_0 + \beta_1 x + \beta_2 z \tag{20-5}$$

which is the equation of a plane in the three-dimensional x, z, y-space.

Similar remarks extend to the case of three or more independent variables, in which case for *linearity and additivity* the response surface must be of the form

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 u + \beta_4 v + \beta_5 w + \ldots \tag{20-6}$$

which is the equation of a hyperplane in the $(x, z, u, v, w, \ldots, y)$-space.

When, as in equation (20-4), the cross-product term $\beta_3 xz$ is present, the effect of a given change in x, say δx, will depend upon the corresponding value of z; the effect of a given change in z, say δz, will depend upon the corresponding value of x; and the joint effect of δx *and* δz will depend on the "starting values" of x and z. In such cases, we say that there is an *interaction* between the factors x and z with respect to their effect on the response y. Hence, in the contrary case, when the changes in y resulting from changes in the two variables x and z *are additive*, it is customary to say that there is *no interaction* between x and z with respect to their effect on the response y.

Many of the standard techniques of statistical analysis, especially analysis-of-variance techniques, depend explicitly on linearity and additivity of the phenomenon under investigation. Thus, the usual analysis of randomized-block experiments (Paragraph 13-3.2) is based on the assumption that the response y_{ij} of the ith treatment in the jth block can be expressed in the form

$$y_{ij} = \varphi_i + \beta_j, \tag{20-7}$$

where φ_i serves to characterize the expected response of the ith treatment, and may be regarded as the average response of the ith treatment over all of the blocks of the experiment; and β_j characterizes the effect of the jth block, and is the amount by which the response in the jth block of *any* *one* of the treatments may be expected to differ from its average response over all blocks. Similarly, in the analysis of Latin-square experiments (Paragraph 13-5.2.1), it usually is assumed that y_{ijm}, the response of the ith treatment under conditions corresponding to the jth row and the mth column, can be represented in the form

$$y_{ijm} = \varphi_i + \rho_j + \kappa_m , \qquad (20\text{-}8)$$

where, as before, φ_i serves to characterize the ith treatment, and may be regarded as the expected response for the ith treatment averaged over all combinations of conditions (corresponding to the rows and columns) included in the experiment; ρ_j serves to characterize the jth row, and may be regarded as the amount by which the response of *any one* of the treatments may be expected to differ under the conditions of the jth row from *its* response averaged over all of the experiment; and κ_m serves to characterize the mth column, and may be regarded as the amount by which the response of *any* treatment may be expected to differ under the conditions of the mth column from *its* response averaged over the entire experiment.

In the case of factorial-type experiments involving many factors (Chapter 12), complete additivity as defined by equation (20-6) is rarely realistic. However, if *internal estimates* of experimental error are to be obtainable from the experimental data in hand (Paragraph 12-1.2.1), then at least some of the higher-order interaction terms, involving, say, three or more factors (e.g., terms in xzw, xzu, . . . ; $xzwu$, . . . ; $xzwuv$, . . .) must either be absent or at least of negligible magnitude in comparison to σ, the actual standard deviation of the measurements involved.

Thus the importance of additivity in the analysis and interpretation of randomized-block, Latin-square, and other multi-factor experiments is seen to be twofold: first, only when the effects of treatments and blocks, or treatments and rows and columns, etc., are strictly additive can we use a single number φ_i to represent the effect of the ith treatment under the range of conditions included in the experiment; and second, only when strict additivity prevails will the residual deviations of the observed responses Y from response surfaces of the form of equation (20-5), (20-6), (20-7), or (20-8), provide unbiased estimates s^2 of the actual experimental-error variance σ^2 associated with the experimental setup concerned. In the absence of strict additivity, for example, when "interaction" cross-product terms $(\varphi\beta)_{ij}$ need to be added to equation (20-7), the actual effect of the ith treatment will depend upon the conditions corresponding to the particular block concerned, being $\varphi_i + (\varphi\beta)_{i1}$ for the first block, $\varphi_i + (\varphi\beta)_{i2}$ for the second block, etc. Furthermore, if the experimental data are analyzed on the supposition that equation (20-7) holds, whereas the cross-product terms actually are necessary to describe the situation accurately, then the resulting residual sum of squares will contain a component due to the sum of the squares of the interaction terms $(\varphi\beta)_{ij}$. Consequently, the resulting variance estimates s^2 will tend to exceed the true experimental-error variance σ^2, to reduce the apparent "significance" of experimental estimates of the actual treatment effects φ_i, and to yield unnecessarily wide confidence interval estimates for the φ_i, and for differences between them. Worse, the customary distribution theory will no longer be strictly applicable, so that the resulting tests for significance and confidence interval estimates will, at best, be only approximately valid.

Therefore, it is highly desirable that the effects of treatments and other factors involved in a complex experiment, if not additive, at least have negligible interactions, in the sense that the corresponding terms needed to depict the situation accurately be individually and collectively negligible in comparison with the corresponding main effects (φ_i, β_j, etc.) and also with respect to the true experimental-error variance σ^2.

20-4.2 TRANSFORMATION OF DATA TO ACHIEVE LINEARITY AND ADDITIVITY

It should be noted that in connection with the linear relationship in equation (20-3) we added a qualifying phrase "when x and y are expressed on appropriate scales." This qualification was added because, if a response y depends *non-linearly* on the corresponding input x and the form of this non-linear relationship is known, then sometimes it is possible to make a transformation of one or both of the variables so that the relationship between the transformed variables y' and x' is of the form of equation (20-3) with y' in place of y and x' in place of x. A number of such linearizing transformations are considered in Paragraph 5-4.4, and are summarized in Table 5-4.

The art of transformation of data to achieve additivity is far less well developed than are the arts of transformation to achieve normality, constancy of variance, and linearity. The only situation that comes to mind for which the exact transformation needed to achieve additivity is obvious, is the case where, say, treatment, row, and column effects are multiplicative in the original units, so that instead of equation (20-8) we have

$$y_{ijm} = \varphi_i \rho_j \kappa_m . \tag{20-9}$$

On taking logarithms this becomes

$$\log y_{ijm} = \log \varphi_i + \log \rho_j + \log \kappa_m , \tag{20-10}$$

which clearly is of the form given in equation (20-8) in terms of the variables

$$y'_{ijm} = \log y_{ijm},\ \varphi'_i = \log \varphi_i,\ \rho'_j = \log \rho_j, \text{ and } \kappa'_m = \log \kappa_m .$$

Fortunately, it often happens that a transformation chosen for the purpose of achieving constancy of variance also improves the situation to some extent with respect to linearity and additivity. But, this will not always be the case. In some situations, if we can find a transformation that improves linearity or additivity we may choose to forego the advantages of constancy of variance. Such is the case, for example, when we adopt the *probit transformation* (Chapter 10) in order to achieve linearity, with the consequent necessity of performing weighted analyses of the transformed data to allow for non-constancy of variance. In other cases, variance constancy may be so advantageous that we are willing to proceed on the assumption that additivity also is achieved by the transformation to stabilize variance — a situation explored by Cochran[16] for the cases of binomial or Poisson-distributed data.

20-5 CONCLUDING REMARKS

One important characteristic of all of the transformations given in Table 20-1 is that they all are *order preserving:* the relative rank order (with respect to magnitude) of the original *individual measurements* $X_1, X_2, \ldots$ is strictly preserved in their transforms $Y_1 = f(X_1)$, $Y_2 = f(X_2) \ldots$. Consequently, the relative rank order of subgroup means $\bar{X}_{(1)}, \bar{X}_{(2)}, \ldots$ of the original measurements will usually — but not necessarily* — be preserved in the corresponding subgroup means $\bar{Y}_{(1)}, \bar{Y}_{(2)}, \ldots$ evaluated from the transformed data. When these subgroup means on the Y-scale are transformed back to the X-scale by the inverse transformation $X = g(Y)$, their transforms $\bar{X}'_{(1)} = g(\bar{Y}_{(1)})$, $\bar{X}'_{(2)} = g(\bar{Y}_{(2)}), \ldots$ will always be in the same relative rank order as the subgroup

* For example, let the original data consist of the following two groups of two observations each: 1, 10 and 5, 5. Then, $\bar{X}_1 = 5.5$, $\bar{X}_2 = 5$, and $\bar{X}_1 > \bar{X}_2$. If now we change to $Y = \log_{10} X$, the data become 0, 1 and 0.699, 0.699 so that $\bar{Y}_1 = 0.5$, $\bar{Y}_2 = 0.699$, and $\bar{Y}_1 < \bar{Y}_2$.

means $\bar{Y}_{(1)}$, $\bar{Y}_{(2)}$, ... on the Y-scale; and hence, usually — but not always — in the same relative rank order as the original subgroup means $\bar{X}_{(1)}$, $\bar{X}_{(2)}$, ... on the X-scale. In other words, by using one of these transformations, we ordinarily will not seriously distort the relative magnitudes of treatment effects, of block effects, etc.

The "transformed-back" subgroup means $\bar{X}'_{(1)}$, $\bar{X}'_{(2)}$, ..., will, of course, not have the same meaning as the "straight-forward" subgroup means $\bar{X}_{(1)}$, $\bar{X}_{(2)}$, ... Thus, in the case of the logarithmic transformation $y = \log X$, if the subgroup means $\bar{Y}_{(1)}$, $\bar{Y}_{(2)}$, ... on the transformed scale (Y) are *arithmetic means* of the corresponding Y values, then the "transformed-back" subgroup means $\bar{X}'_{(1)} = \text{anti-log } \bar{Y}_{(1)}$, $\bar{X}'_{(2)} = \text{anti-log } \bar{Y}_{(2)}$, ..., are estimates, *not* of the corresponding population arithmetic means $\mu_{(1)}$, $\mu_{(2)}$, ..., *but rather* of the corresponding population *geometric means* $\gamma_{(1)}$, $\gamma_{(2)}$, ... On the other hand, if instead of considering subgroup means, we were to consider subgroup medians $\tilde{X}_{(1)}$, $\tilde{X}_{(2)}$, ..., then the corresponding subgroup medians $\tilde{Y}_{(1)}$, $\tilde{Y}_{(2)}$, ..., on the Y-scale will always be in the same relative rank order as the original subgroup medians on the X-scale; and the "transformed-back" subgroup medians $\tilde{X}'_{(1)} = g(\tilde{Y}_{(1)})$, $\tilde{X}'_{(2)} = g(\tilde{Y}_{(2)})$, ..., will be *identically equal to* the original subgroup medians $\tilde{X}_{(1)}$, $\tilde{X}_{(2)}$, ... Consequently, if there is some danger of distortion through the use of a transformation to achieve normality, constancy of variance, linearity, or additivity, then consideration should be given to:

(a) whether, for the technical purposes at issue, discussion might not be equally or perhaps even more conveniently conducted in terms of the transformed values Y, thus obviating the necessity of transforming back to the original X-scale; or,

(b) whether, for purposes of discussion, population medians rather than population means might well be equally or perhaps more meaningful.

In this connection, it must be pointed out that confidence limits for means, differences between means (medians, differences between medians) etc., evaluated in terms of the transformed values Y can be "transformed back" directly into confidence limits for the corresponding magnitudes* on the original X-scale. On the other hand, estimated standard errors of means (medians), differences between means (differences between medians), etc., evaluated on the transformed scale Y cannot be "transformed-back" directly into standard errors of the corresponding "transformed-back" magnitudes on the original scale X. Hence, if standard errors of *final results* are to be given as a way of indicating their respective imprecisions, such standard errors must be evaluated for, and stated as being applicable only to, *final results* expressed on the transformed scale Y.

As so eloquently remarked by Acton[17, pp 221-222]:

"These three reasons for transforming ... [i.e., to achieve normality, constancy of variance, or additivity] have no obvious mathematical compulsion to be compatible; a transformation to introduce additivity might well throw out normality and mess up the constancy of variance beyond all recognition. Usually, the pleasant cloak of obscurity hides knowledge of all but one property from us — and so we cheerfully transform to vouchsafe unto ourselves this one desirable property while carefully refraining from other considerations about which query is futile. But there is a brighter side to this picture. The gods who favor statisticians have frequently ordained that the world be well behaved, and so we often find that a transformation to obtain one of these *desiderata* in fact achieves them all (well, *almost* achieves them!)."

Nevertheless, the following sobering advice from Tippett[18, pp 344-345] should not go unheeded: —

"If a transformed variate [y], having convenient statistical properties, can be substituted for x in the technical arguments from the results and in their applications, there is everything to be said for making the transformation. But otherwise the situation can become obscure. Suppose, for example, that there is an interaction between treatments and looms when the measure is warp breakage rate and that the interaction disappears for the logarithm of the warp breakage rate. It requires some clear thinking to decide what this signifies technically; and the situation becomes somewhat obscure when, as so often happens, the effects are not overwhelmingly significant, and it is remembered that a verdict 'no significant interaction' is not equivalent to 'no interaction.' If the technical interpretation has to be in terms of the untransformed variate x, and after the statistical analysis has been performed on [y], means and so on have to be converted back to x, statistical difficulties arise and the waters deepen. Readers are advised not to make transformations on statistical grounds alone unless they are good swimmers and have experience of the currents."

* E.g., for *geometric means* on the X-scale, if the transformation involved is $Y = \log X$ and *arithmetic means* are employed on the Y-scale.

REFERENCES

1. C. Eisenhart, "The Assumptions Underlying the Analysis of Variance," *Biometrics*, Vol. 3, No. 1, pp. 1-21, March 1947.

2. M. S. Bartlett, "The Use of Transformations," *Biometrics*, Vol. 3, No. 1, pp. 39-52, March 1947.

3. J. Aitchison and J. A. C. Brown, *The Lognormal Distribution*, Cambridge University Press, Cambridge, England, 1957.

4. J. H. Curtiss, "On Transformations Used in the Analysis of Variance," *Annals of Mathematical Statistics*, Vol. 14, No. 2, pp. 107-122, June 1943.

5. N. L. Johnson, "Systems of Frequency Curves Generated by Methods of Translation," *Biometrika*, Vol. 36, Pts. 1 and 2, pp. 149-176, June 1949.

6. W. G. Cochran, "Some Difficulties in the Statistical Analysis of Replicated Experiments," *Empire Journal of Experimental Agriculture*, Vol. VI, No. 22, pp. 157-175, April 1938.

7. W. G. Cochran, "Some Consequences When the Assumptions for the Analysis of Variance Are Not Satisfied," *Biometrics*, Vol. 3, No. 1, pp. 22-38, March 1947.

8. W. G. Cochran, "The Combination of Estimates from Different Experiments," *Biometrics*, Vol. 10, No. 1, pp. 101-129, March 1954.

9. M. S. Bartlett and D. G. Kendall, "The Statistical Analysis of Variance — Heterogeneity and the Logarithmic Transformation," *Journal of the Royal Statistical Society, Supplement* Vol. VIII, No. 1, pp. 128-138, 1946.

10. M. F. Freeman and J. W. Tukey, "Transformations Related to the Angular and the Square Root," *Annals of Mathematical Statistics*, Vol. 21, No. 4, pp. 607-611, December 1950.

11. F. Mosteller and R. R. Bush, "Selected Quantitative Techniques," Ch. 8 of *Handbook of Social Psychology* (Edited by G. Lindzey), Addison-Wesley Publishing Co., Inc., Reading, Mass., 1954.

12. F. Mosteller and C. Youtz, "Tables of the Freeman-Tukey Transformations for the Binomial and Poisson Distributions," *Biometrika*, Vol. 48, Pts. 3 and 4, pp. 433-440, December 1961.

13. M. S. Bartlett, "Some Examples of Statistical Methods of Research in Agriculture and Applied Biology," *Journal of the Royal Statistical Society, Supplement* Vol. IV, No. 2, pp. 137-170, 1937; especially second footnote on p. 168.

14. C. Eisenhart, "Inverse Sine Transformation of Proportions," Ch. 16 of *Selected Techniques of Statistical Analysis* (Edited by C. Eisenhart, M. W. Hastay, and W. A. Wallis), McGraw-Hill Book Co., Inc., New York, N. Y., 1947.

15. R. A. Fisher, "On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample," *Metron*, Vol. 1, No. 4, pp. 1-32, 1921.

16. W. G. Cochran, "The Analysis of Variance When Experimental Errors Follow the Poisson or Binomial Laws," *Annals of Mathematical Statistics*, Vol. 11, No. 3, pp. 335-347, September 1940.

17. F. S. Acton, *Analysis of Straight-Line Data*, John Wiley and Sons, Inc., New York, N. Y., 1959.

18. L. H. C. Tippett, *The Methods of Statistics* (4th edition), John Wiley and Sons, Inc., New York, N. Y., 1952.

CHAPTER 21

THE RELATION BETWEEN CONFIDENCE INTERVALS AND TESTS OF SIGNIFICANCE*

21-1 INTRODUCTION

Several chapters in this Handbook are concerned with statistical tests of significance — see, for example, AMCP 706-110, Chapters 3 and 4. In Paragraph 3-2.1.1, the problem is that of deciding whether the average of a new product differs from the known or specified average m_0 of the standard product. The test procedure involves computing a quantity u and comparing u with the difference between the observed average $\bar{X}$ and the standard average m_0. This comparison is the test of significance. A further step in the procedure, however, notes that the interval $\bar{X} \pm u$ is in fact a confidence interval estimate of the true mean of the new product.

In AMCP 706-111, Chapter 8, the problem of comparing an observed proportion with a standard proportion is done directly in terms of the confidence interval for the observed proportion, completely omitting the test-of-significance step given in Chapter 3 for comparisons involving quantitative data. Tables and charts that give confidence intervals for an observed proportion are used, and we "test" whether the observed proportion differs from the standard by noting whether or not the standard proportion is included in the appropriate interval.

Many statistical consultants, when analyzing an experiment for the purpose of testing a statistical hypothesis, e.g., when comparing means of normal populations, find that they prefer to present results in terms of the appropriate confidence interval.

It must be noted of course that not every statistical test can be put in the form of a confidence interval. In general, tests that are direct tests of the value of a parameter of the parent population can be expressed in terms of confidence intervals.

When the results of a statistical test can alternatively be stated in terms of a confidence interval for a parameter, why would we prefer the confidence interval statement? Some authorities have stressed the point that experimenters are not typically engaged in disproving things, but are looking for evidence for affirmative conclusions; after rejecting the null hypothesis, the experimenter will look for a reasonable hypothesis to accept. The relation between confidence intervals and tests of significance is mentioned only briefly in most textbooks, and ordinarily no insight is given as to which conclusion might be more appropriate. (A notable exception is Wallis and Roberts[1].)

* Adapted with permission from *The American Statistician*, Vol. 14, No. 1, 1960, from article entitled "The Relation Between Confidence Intervals and Tests of Significance — A Teaching Aid" by Mary G. Natrella.

21-2 A PROBLEM IN COMPARING AVERAGES

In this Chapter, we review both procedures with reference to a numerical example, which was given in Paragraph 3-2.1.1.

For a certain type of shell, specifications state that the amount of powder should average 0.735 lb. In order to determine whether the average for the present stock meets the specification, twenty shells are taken at random and the weight of powder is determined. The sample average $\bar{X}$ is 0.710 lb. The estimated standard deviation s is 0.0504 lb. The question to be answered is whether or not the average of present stock differs from the specification value. In order to use a two-sided test of significance at the $(1 - \alpha)$ probability level, we compute a critical value, to be called u. Let

$$u = \frac{t^* s}{\sqrt{n}}$$

where t^* is the positive number exceeded by $100 \left(\frac{\alpha}{2}\right) \%$ of the t-distribution with $n - 1$ degrees of freedom. (See Table A-4.)

In the above example with $\alpha = .05$, t^* equals 2.09 and u equals 0.0236 lb. The test of significance says that if $|\bar{X} - 0.735| > u$, we decide that the average for present stock differs from the specified average. Since

$$|0.710 - 0.735| > 0.0236,$$

we decide that there *is* a difference.

From the same data, we also can compute a 95% confidence interval for the average of present stock. This confidence interval is $\bar{X} \pm u = 0.710 \pm 0.0236$, or the interval from 0.686 to 0.734 lb. The confidence interval can be used for a test of significance; since it *does NOT include* the standard value 0.735, we conclude that the average for the present stock *DOES differ* from the standard.

Comparisons of two materials (see Paragraph 3-3.1.1 for the case of both means unknown and equal variances) may be made similarly. In computing a test of significance, we compare the difference between sample means, $|\bar{X}_A - \bar{X}_B|$, with a computed critical quantity, again called u. If $|\bar{X}_A - \bar{X}_B|$ is larger than u, we declare that the means differ significantly at the chosen level. We also note that the interval

$$(\bar{X}_A - \bar{X}_B) \pm u$$

is a confidence interval for the difference between the true means $(m_A - m_B)$; if the computed interval does not include zero, we conclude from the experiment that the two materials differ in mean value.

21-3 TWO WAYS OF PRESENTING THE RESULTS

Here then are two ways to answer the original question. We may present the result of a test of significance, or we may present a confidence interval. The significance test is a go no-go decision. We compute a critical value u, and we compare it with an observed difference. If the observed difference exceeds u, we announce a significant difference; if it does not, we announce that there is *NO difference.* If we had no OC curve for the test, our decision would be a yes-no proposition with no shadowland of indifference. The significance test may have said *NO*, but only the OC curve can qualify this by showing that this particular experiment had only a ghost of a chance of saying *YES* to this particular question. For example, see Figure 21-1. If the true value of $d = \left|\frac{m - m_0}{\sigma}\right|$ were equal to 0.5, a sample of 10 is not likely to detect a difference, but a sample of 100 is almost certain to detect such a difference.

Using a rejection criterion alone is not the proper way to think of a significance test; we should always think of the associated OC curve as part and parcel of the test. Unfortunately, this has not always been the case. As a matter of fact, many experimenters who use significance tests are using them as though there were no such thing as an OC curve.

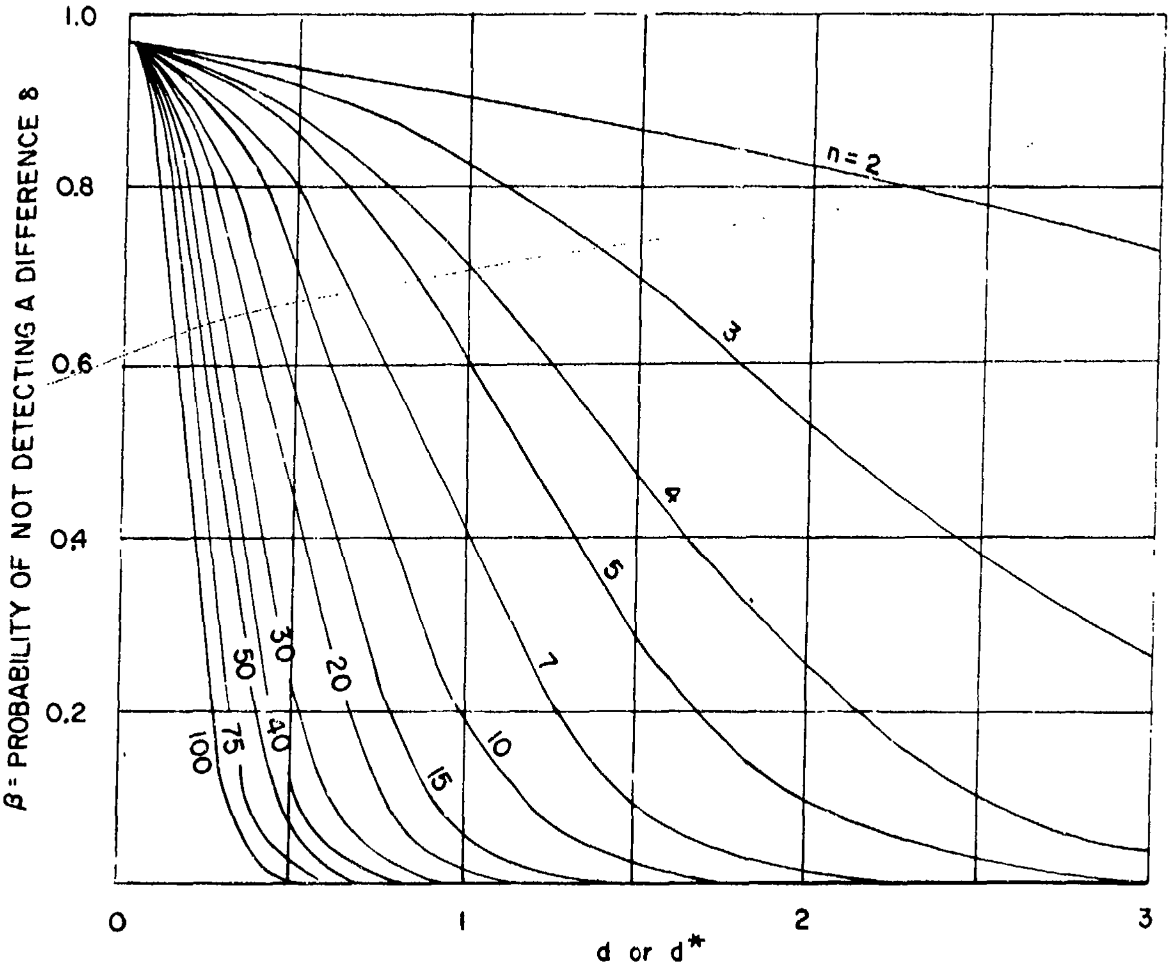


Figure 21-1. Reprint of Figure 3-1. OC curves for the two-sided t-test (α = .05).

Adapted with permission from *Annals of Mathematical Statistics*, Vol. 17, No. 2, June 1946, pp. 178-197, from article entitled "Operating Characteristics for the Common Statistical Tests of Significance" by C. D. Ferris, F. E. Grubbs, and C. L. Weaver.

21-4 ADVANTAGES OF THE CONFIDENCE-INTERVAL APPROACH

A confidence-interval procedure contains information similar to the appropriate OC curve, and, at the same time, is intuitively more appealing than the combination of a test of significance and its OC curve. If the standard value is contained in the confidence interval, we can announce *NO difference*. The *width* of the confidence interval gives a good idea of how firm is the Yes or No answer; however, there is a caution in this regard which is explained in the following paragraphs.

Suppose that the standard value for some property is known to be 0.735, and that a $100(1 - \alpha)\%$ confidence interval for the same property of a possibly different material is determined to be the interval from 0.600 to 0.800. It is true that the standard value does lie within this interval, and that we would declare *no difference*. All that we really know about the new product, however, is that its mean probably is between 0.6 and 0.8. If a much more extensive experiment gave a $100(1 - \alpha)\%$ confidence interval of 0.60 to 0.70 for the new mean, our previous decision of no difference would be reversed.

On the other hand, if the computed confidence interval for the same confidence coefficient had been 0.710 to 0.750, our answer would still have been *no difference*, but we would have said *NO* more loudly and firmly. The confidence interval not only gives a Yes or No answer, but, by its width, also gives an indication of whether the answer should be whispered or shouted.

This is certainly true when the width of the interval for a given confidence coefficient is a function only of n and the appropriate dispersion parameter (e.g., known σ). When the width itself is a random variable (e.g., is a fixed multiple of s, the estimate of σ from the sample), we occasionally can be misled by unusually short or long intervals. But the *average width* of the entire *family* of intervals associated with a given confidence-interval procedure is a definite function of the appropriate dispersion parameter, so that *on the average* the random widths do give similar information. For a graphical illustration of confidence intervals computed from 100 random samples of $n = 4$ (actually random normal deviates), see Figure 21-2. Despite the fluctuation in size and position of the individual intervals, a proportion of intervals which is remarkably close to the specified proportion do include the known population average. If σ were known rather than estimated from the individual sample, the intervals would fluctuate in position only, of course.

The significance test gives the same answer, and a study of the OC curve of the test indicates how firm is the answer. If the test is dependent on the value of σ, the OC curve has to be given in terms of the unknown σ. In such a situation, we must use an upper bound for σ in order to interpret the OC curve, and again we may be misled by a poor choice of this upper bound. On the other hand, the *width* of the confidence interval is part and parcel of the information provided by that method. No *a priori* estimates need be made of σ as would be necessary to interpret the OC curve. Furthermore, a great advantage of confidence intervals is that the width of the interval is in the same units as the parameter itself. The experimenter finds this information easy to grasp, and easy to compare with other information he may have.

The most striking illustration of information provided by confidence intervals is shown in the charts of confidence limits for a binomial parameter. In this case, the limits depend only on n and the parameter itself, and we cannot be misled in an individual sample.

Suppose that a new item is being tested for comparison with a standard. We observe two defectives in a sample of 10, and we estimate the proportion defective for the new item as 0.20. The 95% confidence interval given in Table A-22, corresponding to an observed proportion of 0.20 ($n = 10$), is 0.04 to 0.60. Assume that the known proportion defective for the standard P_0 is 0.10. Our experiment with a sample of 10 gives a confidence interval which includes P_0; and, therefore, we announce *no difference* between the new item and the standard in this regard. Intuitively, however, we feel that the interval 0.04 to 0.60 is so wide that our experiment was not very indicative. Suppose that we test 100 new items and observe 20 defectives. The observed proportion defective again is 0.20. The confidence interval from Table A-24 is 0.13 to 0.29, and does *not* include $P_0 = 0.10$. This time, we are forced to announce that the new item *is different* from the standard; and the narrower width of the confidence interval (0.13 to 0.29) gives us some confidence in doing so.

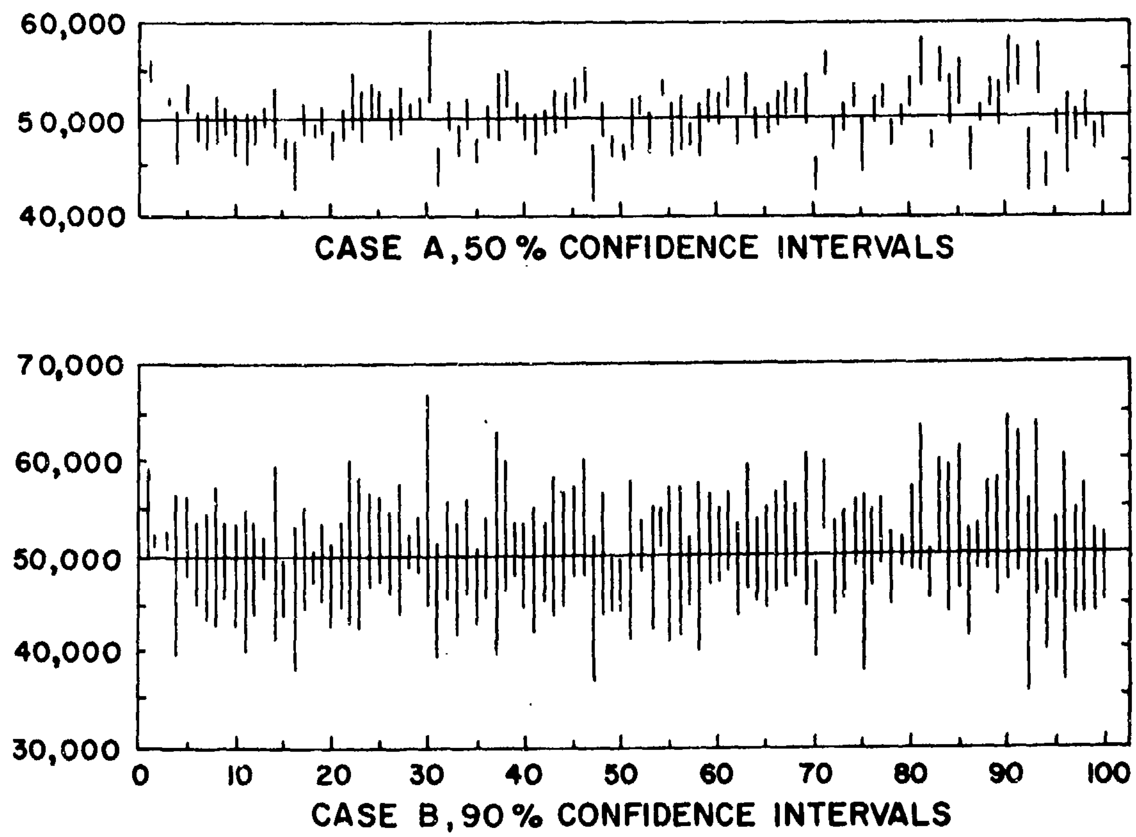


Figure 21-2. Reprint of Figure 1-8. Computed confidence intervals for 100 samples of size 4 drawn at random from a normal population with m = 50,000 psi, σ = 5,000 psi. Case A shows 50% confidence intervals; Case B shows 90% confidence intervals.

21-5 DEDUCTIONS FROM THE OPERATING CHARACTERISTIC (OC) CURVE

The foregoing paragraphs have shown that it is possible to have some notion of the discriminatory power of the test from the size of confidence intervals. Is it also possible, in reverse, to deduce from the OC curve what kind of confidence interval we would get for the new mean? Although we cannot deduce the exact width of the confidence interval, we can infer the order of magnitude. Suppose that: we have measured 100 items; we have performed a two-sided t-test (does the average m differ from m_0?); and we have obtained a significant result. Look at the curve for $n = 100$ in Figure 21-1, which plots the probability of accepting H_0 (the null hypothesis) against $d = \left|\frac{m - m_0}{\sigma}\right|$. From the curve, we see that when d is larger than 0.4, the probability of accepting the null hypothesis is practically zero. Since our significance test *did* reject the null hypothesis, we may reasonably assume that our $d = \left|\frac{m - m_0}{\sigma}\right|$ is larger than 0.4, and may perhaps infer a bound for the true value of $|m - m_0|$ — in other words, some "confidence interval" for m.

On the other hand, suppose that only 10 items were tested and a significant result was obtained. If we look at the curve for $n = 10$, we see that the value of d which is practically certain to be picked up on a significance test is $d = 1.5$ or larger. As expected, a significant result from an experiment which tested only 10 items corresponds to a wider confidence interval for m than the interval inferred from the test of 100 items. A rough comparison of the relative widths may be made. More quantitative comparisons could be made, but the purpose here is to show a broad general relationship.

21-6 RELATION TO THE PROBLEM OF DETERMINING SAMPLE SIZE

The problem of finding the sample size required to detect differences between means can be approached in two ways also. We can specify tolerable risks of making either kind of *wrong* decision (errors of the first or the second kind) — thereby fixing two points on the OC curve of the pertinent test. Matching these two points with computed curves for various n, enables us to pick the proper sample size for the experiment.

Alternatively, we can specify the magnitude of difference between means which is of importance. We then compute the sample size required to give a confidence interval of fixed length equal to the specified difference.

21-7 CONCLUSION

Presentation of results in terms of confidence intervals can be more meaningful than is the presentation of the usual tests of significance (if the test result is not considered in connection with its OC curve). Things are rarely black or white; decisions are rarely made on one-shot tests, but usually in conjunction with other information. Confidence intervals give a feeling of the uncertainty of experimental evidence, and (very important) give it in the same units, metric or otherwise, as the original observations. A recent development in statistical theory that stems from the intuitive preference for confidence intervals is given in Birnbaum[2].

REFERENCES

1. W. A. Wallis and H. V. Roberts, *Statistics, A New Approach*, pp. 461-463, The Free Press, Glencoe, Ill., 1956.

2. A. Birnbaum, "Confidence Curves: An Omnibus Technique for Estimation and Testing Statistical Hypotheses," *Journal of the American Statistical Association*, Vol. 56, No. 294, pp. 246-249, June 1961.

CHAPTER 22

NOTES ON STATISTICAL COMPUTATIONS

22-1 CODING IN STATISTICAL COMPUTATIONS

Coding is the term used when arithmetical operations are applied to the original data in order to make the numbers easier to handle in computation. The possible coding operations are:

(a) Multiplication (or its inverse, division) to change the order of magnitude of the recorded numbers for computing purposes.

(b) Addition (or its inverse, subtraction) of a constant — applied to recorded numbers which are nearly equal, to reduce the number of figures which need be carried in computation.

When the recorded results contain non-significant zeros, (e.g., numbers like .000121 or like 11,100), coding is clearly desirable. There obviously is no point in copying these zeros a large number of times, or in adding additional useless zeros when squaring, etc. Of course, these results could have been given as 121×10^{-4} or 11.1×10^{3}, in which case coding for order of magnitude would not be necessary.

The purpose of coding is to save labor in computation. On the other hand, the process of coding and decoding the results introduces more opportunities for error in computation. The decision of whether to code or not must be considered carefully, weighing the advantage of saved labor against the disadvantage of more likely mistakes. With this in mind, the following five rules are given for coding and decoding.

1. The whole set of observed results must be treated alike.

2. The possible coding operations are the two general types of arithmetic operations:

   (a) addition (or subtraction); and,

   (b) multiplication (or division). Either (a) or (b), or both together, may be used as necessary to make the original numbers more tractable.

3. Careful note must be kept of how the data have been coded.

4. The desired computation is performed on the coded data.

5. The process of decoding a computed result depends on the computation that has been performed, and is indicated separately for several common computations, in the following Paragraphs (a) through (d).

(a) *The mean* is affected by every coding operation. Therefore, we must apply the inverse operation and reverse the order of operations used in coding, to put the coded mean back into original units. For example, if the data have been coded by first multiplying by 10,000 and then subtracting 120, decode the mean by adding 120 and then dividing by 10,000.

	Observed Results		Coded Results
	.0121		1
	.0130		10
	.0125		5
Mean	.0125	Coded mean	5

$$\text{Decoding:}\quad \text{Mean} = \frac{\text{Coded mean} + 120}{10{,}000} = \frac{125}{10{,}000} = .0125$$

(b) *A standard deviation* computed on coded data is affected by multiplication or division only. The standard deviation is a measure of dispersion, like the range, and is not affected by adding or subtracting a constant to the whole set of data. Therefore, if the data have been coded by addition or subtraction only, no adjustment is needed in the computed standard deviation. If the coding has involved multiplication (or division), the inverse operation must be applied to the computed standard deviation to bring it back to original units.

(c) *A variance* computed on coded data must be: multiplied by the square of the coding factor, if division has been used in coding; or divided by the square of the coding factor, if multiplication was used in coding.

(d) *Coding which involves loss of significant figures:* The kind of coding thus far discussed has involved no loss in significant figures. There is another method of handling data, however, that involves both *coding* and *rounding*, and is also called "coding". This operation is sometimes used when the original data are considered to be too finely-recorded for the purpose.

For example, suppose that the data consist of weights (in pounds) of shipments of some bulk material. If average weight is the characteristic of interest, and if the range of the data is large, we might decide to work with weights coded to the nearest hundred pounds, as follows:

Observed Weights Units: lbs.	Coded Data Units: 100 lbs.
7,123	71
10,056	101
100,310	1003
5,097	51
543	5
.	.
.	.
.	.
etc.	etc.

Whether or not the resulting average of the coded data gives us sufficient information will depend on the range of the data and the intended use of the result. It should be noted that this "coding" requires a higher order of judgment than the strictly ari[illegible]al coding discussed in previous exampl[illegible]use some loss of information does occur. [illegible]e decision to "code" in this way should be made by someone who understands the source of the data and the intended use of the computations. The grouping of data in a frequency distribution is coding of this kind.

22-2 ROUNDING IN STATISTICAL COMPUTATIONS

22-2.1 ROUNDING OF NUMBERS

Rounded numbers are inherent in the process of reading and recording data. The readings of an experimenter are rounded numbers to start with, because all measuring equipment is of limited accuracy. Often he records results to even less accuracy than is attainable with the available equipment, simply because such results are completely adequate for his immed[illegible] purpose. Computers often are requir[illegible] round numbers — either to simplify [illegible] metic calculations, or because i[illegible] be avoided, as when 3.1416 is used f[illegible] .414 is used for $\sqrt{2}$.

When a number is to be rounded to a specific number of significant figures, the rounding procedure should be carried out in accordance with the following three rules.

1. When the figure next beyond the last place to be retained is less than 5, the figure in the last place retained should be kept unchanged.

   For example, .044 is rounded to .04.

2. When the figure next beyond the last figure or place to be retained is greater than 5, the figure in the last place retained should be increased by 1.

   For example, .046 is rounded to .05.

[illegible] When the figure next beyond the last fig[illegible] retained is 5, and,

[illegible]ere are no figures or are only zeros [illegible] this 5, an odd figure in the last place to be retained should be increased by 1, an even figure should be kept unchanged.

For example, .045 or .0450 is rounded to .04; .055 or .0550 is rounded to .06.

(b) if the 5 is followed by any figures oth than zero, the figure in the last place to b retained should be increased by 1, whether odd or even.

For example, in rounding to two decimals, .0451 is rounded to .05.

A number should always be rounded off in one step to the number of figures that are to be recorded, and should not be rounded in two or more steps of successive roundings.

22-2.2 ROUNDING THE RESULTS OF SINGLE ARITHMETIC OPERATIONS

Nearly all numerical calculations arising in the problems of everyday life are in some way approximate. The aim of the computer should be to obtain results consistent with the data, with a minimum of labor. We can be guided in the various arithmetical operations by some basic rules regarding significant figures and the rounding of data:

1. *Addition.* When several approximate numbers are to be added, the sum should be rounded to the number of decimal places (not significant figures) no greater than in the addend which has the smallest number of decimal places.

Although the result is determined by the least accurate of the numbers entering the operation, one more decimal place in the more-accurate numbers should be retained, thus eliminating inherent errors in the numbers.

For example:

$$\begin{array}{r} 4.01 \\ .002 \\ .623 \\ \hline 4.635 \end{array}$$

The sum should be rounded to and recorded as 4.64.

2. *Subtraction.* When one approximate number is to be subtracted from another, they must both be rounded off to the same place before subtracting.

Errors arising from the subtraction of nearly-equal approximate numbers are frequent and troublesome, often making the computation practically worthless. Such errors can be avoided when the two nearly-equal numbers can be approximated to more significant digits.

3. *Multiplication.* If the less-accurate of two approximate numbers contains n significant digits, their product can be relied upon for n digits at most, and should not be written with more.

As a practical working plan, carry intermediate computations out in full, and round off the final result in accordance with this rule.

4. *Division.* If the less-accurate of either the dividend or the divisor contains n significant digits, their quotient can be relied upon for n digits at most, and should not be written with more.

Carry intermediate computations out in full, and round off the final result in accordance with this rule.

5. *Powers and Roots.* If an approximate number contains n significant digits, its power can be relied upon for n digits at most; its root can be relied upon for at least n digits.

6. *Logarithms.* If the mantissa of the logarithm in an n-place log table is not in error by more than two units in the last significant figure, the antilog is correct to $n - 1$ significant figures.

The foregoing statements are working rules only. More complete explanations of the rules, together with procedures for determining explicit bounds to the accuracy of particular computations, are given in Scarborough[1], and the effects of rounding on statistical analyses of large numbers of observations are discussed in Eisenhart[2].

22-2.3 ROUNDING THE RESULTS OF A SERIES OF ARITHMETIC OPERATIONS

Most engineers and physical scientists are well acquainted with the rules for reporting a result to the proper number of significant figures. From a computational point of view, they know these rules too well. It is perfectly true, for example, that a product of two numbers should be reported to the same number of significant figures as the least-accurate of the two numbers. It is not so true that the two numbers should be rounded to the same number of significant figures before multiplication. A better rule is to round the more-accurate number to one more figure than the less-accurate number, and then to round the product to the same number of figures as the less-accurate one. The great emphasis against reporting more figures than are reliable has led to a prejudice against carrying enough figures in computation.

Assuming that the reader is familiar with the rules of the preceding Paragraph 22-2.2, regarding significant figures in a single arithmetical operation, the following paragraphs will stress the less well-known difficulties which arise in a computation consisting of a long series of different arithmetic operations. In such a computation, strict adherence to the rules at each stage can wipe out all meaning from the final results.

For example, in computing the slope of a straight line fitted to observations containing three significant figures, we would not report the slope to seven significant figures; but, if we round to three significant figures after each necessary step in the computation, we might end up with no significant figures in the value of the slope.

It is easily demonstrated by carrying out a few computations of this nature that there is real danger of losing all significance by too-strict adherence to rules devised for use at the final stage. The greatest trouble of this kind comes where we must subtract two nearly-equal numbers, and many statistical computations involve such subtractions.

The rules generally given for rounding-off, were given in a period when the average was the only property of interest in a set of data. Reasonable rounding does little damage to the average. Now, however, we almost always calculate the standard deviation, and this statistic does suffer from too-strict rounding. Suppose we have a set of numbers:

$$\begin{array}{r} 3.1 \\ 3.2 \\ 3.3 \\ \hline \text{Avg.} = 3.2 \end{array}$$

If the three numbers are rounded off to one significant figure, they are all identical. The average of the rounded figures is the same as the rounded average of the original figures, but all information about the variation in the original numbers is lost by such rounding.

The generally recommended procedure is to carry two or three extra figures throughout the computation, and then to round off the final reported answer (e.g., standard deviation, slope of a line, etc.) to a number of significant figures consistent with the original data. However, in some special computations such as the fitting of equations by least squares methods given in AMCP 706-110, Chapters 5 and 6, one should carry extra decimals in the intermediate steps — decimals sufficiently in excess of the number considered significant to insure that the computational errors in the final solutions are negligible in relation to their statistical imprecision as measured by their standard errors. For example, on a hand-operated computing machine, use its total capacity and trim the figures off as required in the final results. (See Chapter 23.)

REFERENCES

1. J. B. Scarborough, *Numerical Mathematical Analysis*, Chapter 1, (3d edition), The Johns Hopkins Press, Baltimore, Md., 1955.
2. C. Eisenhart, *Techniques of Statistical Analysis*, Chapter 4, McGraw-Hill Book Co., New York, N. Y., 1947.

CHAPTER 23

EXPRESSION OF THE UNCERTAINTIES OF FINAL RESULTS

23-1 INTRODUCTION

Measurement of some property of a thing in practice always takes the form of a sequence of steps or operations that yield as an end result a number that serves to represent the amount or quantity of some particular property of a thing — a number that indicates how much of this property the thing has, for someone to use for a specific purpose. The end result may be the outcome of a single reading of an instrument, with or without corrections for departures from prescribed conditions. More often, it is some kind of average; e.g., the arithmetic mean of a number of independent determinations of the same magnitude, or the final result of a least squares "reduction" of measurements of a number of different magnitudes that bear known relations with each other in accordance with a definite experimental plan. In general, the purpose for which the answer is needed determines the precision or accuracy of measurement required, and ordinarily also determines the method of measurement employed.

Although the accuracy required of a reported value depends primarily on the use, or uses, for which it is intended, we should not ignore the requirements of other uses to which the reported value is likely to be put. A certified or reported value whose accuracy is entirely unknown is worthless.

Strictly speaking, the actual *error* of a reported value, that is, the magnitude and sign of its deviation from the truth, is usually unknowable. Limits to this error, however, can usually be inferred — with some risk of being incorrect — from the *precision* of the measurement process by which the reported value was obtained, and from reasonable limits to the possible *bias* of the measurement process. The *bias*, or *systematic error*, of a measurement process is the magnitude and direction of its tendency to measure something other than what was intended; its *precision* refers to the typical *closeness together* of successive independent measurements of a single magnitude generated by repeated applications of the process under specified conditions; and, its *accuracy* is determined by the *closeness to the true value* characteristic of such measurements.

Precision and *accuracy* are inherent characteristics of the measurement process employed, and not of the particular end result obtained. From experience with a particular measurement process and knowledge of its sensitivity to uncontrolled factors, we can often place reasonable bounds on its likely systematic error (bias). It also is necessary to know how well the particular value in hand is likely to agree with other values that the same measurement process might have provided in this instance, or might yield on remeasurement of the same magnitude on another occasion. Such information is provided by the *standard error* of the reported value, which measures the characteristic disagreement of repeated determinations of the same quantity by the same method, and thus serves to indicate the precision (strictly, the *imprecision*) of the reported value.

The uncertainty of a reported value is indicated by giving credible limits to its likely inaccuracy. No single form of expression for these limits is universally satisfactory. In fact, different forms of expression are recommended, the choice of which will depend on the relative magnitudes of the imprecision and likely bias; and on their relative importance in relation to the intended use of the reported value, as well as to other possible uses to which it may be put.

Four distinct cases need to be recognized:

1. *Both systematic error and imprecision negligible* in relation to the requirements of the intended and likely uses of the result.

2. *Systematic error not negligible*, but *imprecision negligible*, in relation to the requirements.

3. *Neither systematic error nor imprecision negligible* in relation to the requirements.

4. *Systematic error negligible*, but *imprecision not negligible* in relation to the requirements.

Specific recommendations are made below with respect to each of these four cases, supplemented by further discussion of each case in Paragraphs 23-2 through 23-5. These recommendations may be summarized as follows:

(a) Two numerics, respectively expressing the imprecision and bounds to the systematic error of the result, should be used whenever: (1) the margin is narrow between ability to measure and the accuracy or precision requirements of the situation; or, (2) the imprecision and the bounds to the systematic error are nearly equal in indicating possible differences from the *true value*. Such instances come under Case 3.

(b) A quasi-absolute type of statement with one numeric, placing bounds on the inaccuracy of the result, should be used whenever: (1) a wide or adequate margin exists between ability to measure and the accuracy requirements of the situation (Case 1); (2) the imprecision is negligibly small in comparison with the bounds placed on the systematic error (Case 2); or, (3) the control is so satisfactory that the extent of error is known.

(c) A single numeric expressing the imprecision of the result should be used whenever the systematic error is either zero by definition or negligibly small in comparison with the imprecision (Case 4).

(d) Expressions of uncertainty should be given in sentence form whenever feasible.

(e) The form "$a \pm b$" should be avoided as much as possible; and never used without explicit explanation of its connotation.

23-2 SYSTEMATIC ERROR AND IMPRECISION BOTH NEGLIGIBLE (CASE 1)

In this case, the certified or reported result should be given correct to the number of significant figures consistent with the accuracy requirements of the situation, together with an explicit statement of its accuracy or correctness.

For example:

> . . . the wavelengths of the principal visible lines of mercury 198 have been measured relative to the 6057.802106 A (Angstrom units) line of krypton 98, and their values in vacuum are certified to be
>
> 5792.2685 A
> 5771.1984 A
> 5462.2706 A
> 4359.5625 A
> 4047.7146 A
>
> correct to eight significant figures.

It must be emphasized that when no statement of accuracy or precision accompanies a certified or reported number, then, in accordance with the usual conventions governing rounding, this number will be interpreted as being accurate within $\pm\frac{1}{2}$ unit in the last significant figure given; i.e., it will be understood that its inaccuracy before rounding was less than ± 5 units in the next place.

23-3 SYSTEMATIC ERROR NOT NEGLIGIBLE, IMPRECISION NEGLIGIBLE (CASE 2)

In such cases:

(a) Qualification of a certified or reported result should be limited to a single quasi-absolute type of statement that places bounds on its inaccuracy;

(b) These bounds should be stated to no more than two significant figures;

(c) The certified or reported result itself should be given (i.e., rounded) to the last place affected by the stated bounds, unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that the result may possess for certain particular uses;

(d) Accuracy statements should be given in sentence form in all cases, except when a number of results of different accuracies are presented, e.g., in tabular arrangement. If it is necessary or desirable to indicate the respective accuracies of a number of results, the results should be given in the form $a \pm b$ (or $a \begin{smallmatrix}+b\\-c\end{smallmatrix}$, if necessary) with an appropriate explanatory remark (as a footnote to the table, or incorporated in the accompanying text) to the effect that the $\pm b$, or $\begin{smallmatrix}+b\\-c\end{smallmatrix}$, signify bounds to the errors to which the a's may be subject.

The particular form of the quasi-absolute type of statement employed in a given instance ordinarily will depend upon personal taste, experience, current and past practice in the field of activity concerned, and so forth. Some examples of good practice are:

. . . is (are) not in error by more than 1 part in (x).

. . . is (are) accurate within $\pm$ (x units) (or $\pm$ (x)%).

. . . is (are) believed accurate within (. . . . . . .).

Positive wording, as in the first two of these quasi-absolute statements, is appropriate only when the stated bounds to the possible inaccuracy of the certified or reported value are themselves reliably established. On the other hand, when the indicated bounds are somewhat conjectural, it is desirable to signify this fact (and thus put the reader on guard) by inclusion of some modifying expression such as "believed", "considered", "estimated to be", "thought to be", and so forth, as exemplified by the third of the foregoing examples.

Results should never be presented in the form "$a \pm b$", without explanation. If no explanation is given, many persons will automatically take $\pm b$ to signify bounds to the inaccuracy of a. Others may assume that b is the *standard error* or the *probable error* of a, and hence that the uncertainty of a is at least $\pm 3b$, or $\pm 4b$, respectively. Still others may take b to be an indication merely of the imprecision of the individual measurements; that is, to be the *standard deviation*, the *average deviation*, or the *probable error of a SINGLE observation*. Each of these interpretations reflects a practice of which instances can be found in current scientific literature. As a step in the direction of reducing this current confusion, we urge that the use of "$a \pm b$" in presenting results in official documents be limited to that sanctioned under (d) above.

The term *uncertainty*, with the quantitative connotation of limits to the likely departure from the truth, and not simply connoting vague lack of certainty, may sometimes be used effectively to achieve a conciseness of expression otherwise difficult or impossible to attain. Thus, we might make a statement such as:

> The uncertainties in the above values are not more than ±0.5 degree in the range 0° to 1100°C, and then increase to ±2 degrees at 1450°C.;

or,

> The uncertainty in this value does not exceed . . . . excluding (or, including) the uncertainty of . . . . in the value . . . . adopted for the reference standard involved.

Finally, the following forms of quasi-absolute statements are considered poor practice, and should be avoided:

The accuracy of ...... is 5 percent.
The accuracy of ...... is ±2 percent.

These statements are presumably intended to mean that the result concerned is not inaccurate, i.e., not in error, by more than 5 percent or 2 percent, respectively; but they explicitly state the opposite.

23-4 NEITHER SYSTEMATIC ERROR NOR IMPRECISION NEGLIGIBLE (CASE 3)

In such cases:

(a) A certified or reported result should be qualified by: (1) a quasi-absolute type of statement that places bounds on its systematic error; and, (2) a separate statement of its standard error or its probable error, explicitly identified, as a measure of its imprecision;

(b) The bounds to its systematic error and the measure of its imprecision should be stated to no more than two significant figures;

(c) The certified or reported result itself should be stated, at most, to the last place affected by the finer of the two qualifying statements, unless it is desired to indicate and preserve such relative accuracy or precision of a higher order that the result may possess for certain particular uses;

(d) The qualification of a certified or reported result, with respect to its imprecision and systematic error, should be given in sentence form, except when results of different precision or with different bounds to their systematic errors are presented in tabular arrangement. If it is necessary or desirable to indicate their respective imprecisions or bounds to their respective systematic errors, such information may be given in a parallel column or columns, with appropriate identification.

Here, and in Paragraph 23-5, the term *standard error* is to be understood as signifying *the standard deviation of the reported value itself*, not as signifying *the standard deviation of a single determination* (unless, of course, the reported value is the result of a single determination only).

The above recommendations should not be construed to exclude the presentation of a quasi-absolute type of statement placing bounds on the inaccuracy, i.e., on the overall uncertainty, of a certified or reported value, provided that separate statements of its imprecision and its possible systematic error are included also. Bounds indicating the overall uncertainty of a reported value should not be numerically less than the corresponding bounds placed on the systematic error outwardly increased by at least two times the standard error. The fourth of the following examples of good practice is an instance at point:

The standard errors of these values do not exceed 0.000004 inch, and their systematic errors are not in excess of 0.00002 inch.

The standard errors of these values are less than (x units), and their systematic errors are thought to be less than ± (y units).

... with a standard error of (x units), and a systematic error of not more than ± (y units).

... with an overall uncertainty of ±3 percent based on a standard error of 0.5 percent and an allowance of ±1.5 percent for systematic error.

When a reliably established value for the relevant standard error is available, based on considerable recent experience with the measurement process or processes involved, and the dispersion of the present measurements is in keeping with this experience, then this established value of the standard error should be used. When experience indicates that the relevant standard error is subject to fluctuations greater than the intrinsic variation of such a measure, then an appropriate upper bound should be given, e.g., as in the first two of the above examples, or by changing "a standard error . . ." in the third and fourth examples to "an upper bound to the standard error . . .".

When there is insufficient recent experience with the measurement processes involved, an estimate of the standard error must of necessity be computed, by recognized statistical procedures, from the same measurements as the certified or reported value itself. It is essential that such computations be carried out according to an agreed-upon standard procedure, and that the results thereof be presented in sufficient detail to enable the reader to form his own judgment and make his own allowances for their inherent uncertainties. To avoid possible misunderstanding in such cases:

(a) the term *computed standard error* should be used;

(b) the estimate of the standard error employed should be that obtained from the relation

estimate of standard error

$$= \sqrt{\frac{\text{sum of squared residuals}}{n\nu}},$$

where n is the (effective) number of completely independent determinations of which a is the arithmetic mean (or, other appropriate least squares adjusted value) and ν is the number of degrees of freedom involved in the sum of squared residuals (i.e., the number of residuals minus the number of fitted constants and/or other independent constraints); and,

(c) the number of degrees of freedom ν should be explicitly stated.

If the reported value a is the arithmetic mean, then:

$$\text{estimate of standard error} = \sqrt{\frac{s^2}{n}}$$

where s^2 is computed as shown in AMCP 706-110, Chapter 2, Paragraph 2-2.2, and n is the number of completely independent determinations of which a is the arithmetic mean.

For example:

The computed probable error (or, standard error) of these values is (x units), based on (ν) degrees of freedom, and the systematic error is estimated to be less than $\pm$ (y units).

. . . which is the arithmetic mean of (n) independent determinations and has a computed standard error of . . . .

. . . with an overall uncertainty of ± 5.2 km/sec based on a standard error of 1.5 km/sec and bounds of ± 0.7 km/sec on the systematic error. (The figure 5.2 equals 0.7 plus 3 times 1.5).

Or, if based on a computed standard error:

. . . with an overall uncertainty of ± 7 km/sec derived from bounds of ± 0.7 km/sec on the systematic error and a computed standard error of 1.5 km/sec based on 9 degrees of freedom. (The figure 7 is approximately equal to $0.7 + 4.3\,(1.5)$, where 4.3 is the two-tail 0.002 probability value of Student's t for 9 degrees of freedom. As $\nu \rightarrow \infty$, $t_{.002}(\nu) \rightarrow 3.090$.)

23-5 SYSTEMATIC ERROR NEGLIGIBLE, IMPRECISION NOT NEGLIGIBLE (CASE 4)

In such cases:

(a) Qualification of a certified or reported value should be limited to a statement of its standard error or of an upper bound thereto, whenever a reliable determination of such value or bound is available. Otherwise, a computed value of the standard error so designated should be given, together with a statement of the number of degrees of freedom on which it is based;

(b) The standard error or upper bound thereto, should be stated to not more than two significant figures;

(c) The certified or reported result itself should be stated, at most, to the last place affected by the stated value or bound to its imprecision, unless it is desired to indicate and preserve such relative precision of a higher order that the result may possess for certain particular uses;

(d) The qualification of a certified or reported result with respect to its imprecision should be given in sentence form, except when results of different precision are presented in tabular arrangement and it is necessary or desirable to indicate their respective imprecisions, in which event such information may be given in a parallel column or columns, with appropriate identification.

The above recommendations should not be construed to exclude the presentation of a quasi-absolute type of statement placing bounds on its possible inaccuracy, provided that a separate statement of its imprecision is included also. Such bounds to its inaccuracy should be numerically equal to at least two times the stated standard error. The fourth of the following examples of good practice is an instance at point:

> The standard errors of these values are less than (x units).
>
> . . . with a standard error of (x units).
>
> . . . with a computed standard error of (x units) based on (ν) degrees of freedom.
>
> . . . with an overall uncertainty of ± 4.5 km/sec derived from a standard error of 1.5 km/sec. (The figure 4.5 equals 3 times 1.5).

Or, if based on a computed standard error:

> . . . with an overall uncertainty of ± 6.5 km/sec derived from a computed standard error of 1.5 km/sec (based on 9 degrees of freedom). (The figure 6.5 equals 4.3 times 1.5, where 4.3 is the two-tail 0.002 probability value of Student's t for 9 degrees of freedom. As $\nu \to \infty$, $t_{.002}(\nu) \to 3.090$.

The remarks with regard to a computed standard error in Paragraph 23-4 apply with equal force to the last two of the above examples.

INDEX

The following alphabetical listing includes all topics from the entire Handbook, Sections 1 through 5, inclusive.

Page references have double numerals, identifying both the chapter and page within the chapter. Page numbers for Section 5 (A-Tables) are identified as T-1, T-2, etc., as numbered in that Section to avoid confusion with the pages of Chapter 5. Roman numeral page references indicate pages of the same number in every Section.

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

T

V

W

X

Y

Z