The Design and Analysis of Clinical Experiments

JOSEPH L. FLEISS

Division of Biostatistics
School of Public Health
Columbia University

WILEY CLASSICS LIBRARY EDITION PUBLISHED 1999

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This page intentionally left blank

The Design and Analysis of Clinical Experiments

This page intentionally left blank

The Design and Analysis of Clinical Experiments

JOSEPH L. FLEISS

Division of Biostatistics
School of Public Health
Columbia University

WILEY CLASSICS LIBRARY EDITION PUBLISHED 1999

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This text is printed on acid-free paper. ♾

Published simultaneously in Canada.

Wiley Classics Library Edition Published 1999.

Library of Congress Cataloging in Publication Data:

Fleiss, Joseph L.
The design and analysis of clinical experiments.

Includes bibliographical references and indexes.
1. Clinical trials—Statistical methods.
2. Medicine, Clinical—Problems, exercises, etc.
I. Title. [DNLM: 1. Clinical Trials—methods.
2. Mathematics. 3. Research Design. 4. Statistics.
W 20.5 F596d]

R853.C55F54 1985 615'0724 85-17830
ISBN 0-471-82047-4
ISBN 0-471-34991-7 (Wiley Classics Paperback Edition)

10 9 8 7 6 5 4

TO MY WIFE, ISABEL,
AND OUR CHILDREN, ART, DEB, LIZ, AND MENACHEM

This page intentionally left blank

Preface

Experimental design is concerned with the arrangement of one's experimental units and the assignment to them of treatments in such a way that the comparisons among the treatments are unbiased and as precise and powerful as possible. A score or more of books on the design of experiments are still in print but none, to my knowledge, is devoted to those principles and techniques that are especially relevant in biomedical experiments involving human subjects. In my teaching and consulting, I have referred students and colleagues to the two texts I cite most frequently in this book: *Experimental designs* (second edition) (Cochran and Cox, 1957) and *Planning of experiments* (Cox, 1958), both published by Wiley. The former was often criticized as not being sufficiently applicable to clinical studies and the latter as not providing sufficient guidance with respect to the analysis of the data. I hope that my book will prove more useful in clinical applications than Cochran and Cox's and more helpful statistically than D. R. Cox's.

I have restricted attention to *bona fide* experimental comparisons of treatments, that is, to studies in which treatments are assigned to subjects at random. I therefore do not consider the challenging problems posed by those nonexperimental studies in which the assignment of treatments to subjects was out of the investigator's control (e.g., by being left to the individual clinician). Anderson, Auquier, Hauck, et al. (*Statistical methods for comparative studies: Techniques for bias reduction*. New York: Wiley, 1980); Campbell and Cook (*Quasi-experimentation: Design and analysis for field settings*. Chicago: Rand McNally, 1979) and Cochran (*Planning and analysis of observational studies*. New York: Wiley, 1983) are useful references to the design and analysis of nonexperimental studies.

This book complements the several monographs on clinical trials that have appeared since 1980. Some slight overlap necessarily exists with them, such as in the discussion of methods for carrying out randomization, of problems with multicenter trials, and of the validity of the

crossover design. Overall, however, they tend to be more concerned with issues such as improving patient compliance, satisfying regulatory requirements, methods for assuring double-blindedness, quality control, and ethical constraints in conducting experiments on patients. This book concentrates more on the technical aspects of design and statistical analysis.

The book is aimed primarily at clinical investigators and biostatisticians in biomedical research centers and the pharmaceutical industry who are responsible for designing clinical experiments and for analyzing the resulting data. With rare exceptions, real examples from such specialties as cardiology, dentistry, gerontology, neurology, pediatrics, and psychiatry are used to illustrate the designs and their analyses. The book is also intended to serve as the text for a second-year, graduate-level course on the design of experiments. Each chapter concludes with a set of problems. Some are numerical; others, algebraic. Copious hints and signposts are provided. Mathematical and statistical derivations that involve more than simple algebra also appear as problems (they have been relegated to the end of each chapter so the text can concentrate, with a minimum of digression, on more practical matters).

There are several other features of the book that should make it useful to its intended readers:

1. There is a discussion of the untoward consequences of imprecise measurement, including bias, and a presentation of methods for improving precision.

2. Methods for carrying out a randomized assignment of treatments to subjects are illustrated using tables of random permutations.

3. A distinction is made between blocking and stratification as methods to control for prognostic factors, and the rare occasions when the former is superior to the latter are identified.

4. Techniques that are appropriate for ordered categorical response variables (e.g., poor, fair, or good response) are given prominence.

5. Data analyses are usually illustrated first for the general case of unequal sample sizes and only secondarily for the special case of equal sample sizes.

6. The three most popular methods for analyzing the data from a multicenter study are reviewed, and criteria are proposed for deciding which is appropriate.

7. Problems that may arise in the study of change, and some possible solutions, are discussed.

8. Some recent suggestions for analyzing the data from a two-period crossover study are reviewed and shown capable of producing biased results.

9. In the chapters on factorial studies, the emphasis is more on the estimation and interpretation of factorial effects than on the tests for their statistical significance.

10. For those occasions when the data must be analyzed by computer, criteria are suggested for choosing the appropriate package and the appropriate set of options from that package.

11. The appendix is devoted to easily programmed methods for determining the sample sizes needed to assure specified power.

A knowledge of statistics at the level of Armitage's *Statistical methods in medical research* (New York: Wiley, 1971) or Snedecor and Cochran's *Statistical methods*, seventh edition (Ames, Iowa: Iowa State University Press, 1980) is assumed. Familiarity with matrix algebra will help the reader understand a few sections of the book, especially Sections 3.4, 6.4, and 8.2, and part of 7.2. A knowledge of calculus will help the reader solve some of the more mathematical problems at the ends of some chapters. Otherwise, a good knowledge of high school algebra is the only mathematical prerequisite.

I am pleased to acknowledge the help, advice, encouragement, and criticism I received from several people. J. Thomas Bigger, Jr., Albert Kingman, and Linda Rolnitzky kindly provided me with some of their unpublished data; others also did so, but asked to remain anonymous. Charles Dunnett graciously gave me permission to reproduce some new and as yet unpublished critical values for his multiple comparison criterion. Students in my course on experimental design at Columbia University saw draft copies of the manuscript and pointed out several typographical errors. John Fertig, Rupert Miller, and Sylvan Wallenstein read the penultimate draft and made suggestions that I always took seriously but sometimes chose not to follow. The influence of John Fertig, my predecessor as professor and head of biostatistics at Columbia University, was more profound than that of a critical reader. I learned the design of experiments and the analysis of experimental data from him as my professor, and I learned the practice of biostatistics from him as a role model *par excellence*. He died while I was putting the finishing touches on the book. I shall miss him.

Molly Park and Michael Parides carried out the computer analyses that are reported in Sections 5.4 and 6.4. The typing of the initial drafts was ably performed by Alice Arana and Anntrene Wilson. Gerda Burian

Cordova and my son Art helped with the editing and indexes. My editor at Wiley-Interscience, Bea Shube, was constantly supportive, encouraging, and a morale booster. My wife, Isabel, was all of these but also ever patient and a source of inspiration.

JOSEPH L. FLEISS

New York, New York
September 1985

Contents

CHAPTER

1. Reliability of Measurement 1

   1.1. A Statistical Model for Reliability 2
   1.2. Some Consequences of Unreliability 3
   1.3. The Simple Replication Reliability Study 8
   1.4. The Control of Unreliability by Replication 14
   1.5. The Interexaminer Reliability Study 17
   Problems 28
   References 31

2. Simple Linear Regression Analysis 33

   2.1. The Linear Regression Model 33
   2.2. Inferences About the Slope and Intercept 37
   2.3. Estimating Input from Output 41
   Problems 43
   References 45

3. The Parallel Groups Design 46

   3.1. Randomization in the Parallel Groups Design 47
   3.2. The Analysis of Variance and Multiple Comparisons 51
   3.3. Equality of Variance, Normality, and Transformations 59
   3.4. The Analysis of Several Variables 68

3.5. A Non-Normally Distributed Response Variable 73
3.6. Ridit Analysis for Ordered Categorical Data 80
Problems 84
References 88

4. Special Cases of the Parallel Groups Study 91

4.1. Several Treatments Versus a Control 92
4.2. The 2×2 Factorial Experiment 96
4.3. The Bonferroni Criterion for Multiple Comparisons 103
4.4. A Quantitative Experimental Factor 107
Problems 115
References 119

5. Blocking to Control for Prognostic Variables 120

5.1. The Randomized Blocks Experiment 121
5.2. The Analysis of Variance for Randomized Blocks 125
5.3. Nonparametric Analyses 130
5.4. Missing Values 135
Problems 143
References 148

6. Stratification to Control for Prognostic Variables 149

6.1. The Comparison of Two Treatments 150
6.2. Treatment-by-Stratum Interaction 161
6.3. Pre- Versus Post-Stratification 164
6.4. The Comparison of More than Two Treatments 165
6.5. Multicenter Studies 176
Problems 180
References 183

7. Analysis of Covariance and the Study of Change 186

7.1. The Measurement of Change 187
7.2. The Algebra of Analysis of Covariance 194
7.3. Nonparallel Regression Lines 203
7.4. More Complicated Designs 208
7.5. Describing Treatment Effects for a Deteriorating Condition 210

Problems 215
References 219

8. REPEATED MEASUREMENTS STUDIES 220

8.1. The Analysis of Variance of Repeated Measurements 220
8.2. The Multivariate Analysis of Repeated Measurements 228
8.3. Multiple Comparisons Involving Time 232
Problems 236
References 239

9. LATIN AND GRECO-LATIN SQUARES 241

9.1. The Single $g \times g$ Latin Square 242
9.2. Replicated Latin Squares 249
9.3. Variations on the Latin Square 255
Problems 261
References 262

10. THE CROSSOVER STUDY 263

10.1 The Two-Period Crossover Study 264
10.2. A Non-Normally Distributed Response Variable 275
10.3. More than Two Treatments 281
Problems 286
References 289

11. BALANCED INCOMPLETE BLOCK DESIGNS 291

11.1. Application to an Interexaminer Reliability Study 292
11.2. A BIBD As a Two-Period Crossover Study 300
Problems 302
References 305

12. FACTORIAL EXPERIMENTS 306

12.1. The 2^p Factorial Study, Unequal Sample Sizes 307
12.2. The 2^p Factorial Study, Equal Sample Sizes 319
12.3. A 3×4 Factorial Experiment 330

12.4. Fractional Replication of a 2^p Study 335
Problems 341
References 347

13. Split-Plot Designs and Confounding 348

13.1. Split-Plot Experiments 349
13.2. General Confounding 355
Problems 365
References 368

Appendix. Sample-Size Determination 369

A.1. The Comparison of Two Treatments 369
A.2. The Comparison of Several Treatments 371

Author Index 419

Subject Index 423

The Design and Analysis of Clinical Experiments

This page intentionally left blank

CHAPTER 1

Reliability of Measurement

The most elegant design of a clinical study will not overcome the damage caused by unreliable or imprecise measurement. The requirement that one's data be of high quality is at least as important a component of proper study design as the requirement for randomization, double blinding, controlling when necessary for prognostic factors, and so on. Larger sample sizes than otherwise necessary, biased estimates, and even biased samples are some of the untoward consequences of unreliable measurement that will be demonstrated.

Investigators in the mental disorders traditionally have been more concerned with the reliability of their measures, and with the design of informative reliability studies (Grove, Andreasen, McDonald-Scott, et al., 1981) than have their colleagues in other medical specialties. All clinical investigators should be as concerned: as shown by Koran (1975), reliability appears to be equally good (or equally poor) in all the specialties in which reliability data have been collected and published.

Attention is restricted in this chapter to the reliability of quantitative data. In many clinical studies, the response variable will be qualitative: a familiar example is improved—no change—worse. Cohen's kappa statistic (Cohen, 1960) is the appropriate measure of reliability for such data. The reader is referred to Davies and Fleiss (1982), Fleiss (1981, Chapter 13), and Landis and Koch (1977) for applications and generalizations of Cohen's kappa.

In Section 1.1, a statistical framework is provided for the formal definition and measurement of reliability. Some of the consequences of unreliability are described in Section 1.2. In Section 1.3, methods for making inferences about reliability are presented when one's reliability study calls for independent replicate measurements to be made on each of a sample of subjects. Section 1.4 is devoted to replication as a method

for improving reliability and indicates how the cost of measurement may be taken into account. In Section 1.5, methods for estimating and improving reliability are presented when the measurements are made by each of the same set of examiners.

Some of the statistical concepts alluded to in this chapter (e.g., the Scheffé and Bonferroni criteria for multiple comparisons) are not defined until later. Therefore, it might seem more appropriate to have placed this chapter later in the book after all the concepts discussed in it had been introduced. The idea that good measurement is fundamental to good design is so important, though, that it seemed preferable to begin the book with a development of this theme even at the risk of some readers' having to check on some ideas in later chapters.

1.1. A STATISTICAL MODEL FOR RELIABILITY

Let X represent the observed value for an individual on some variable. No matter what the variable and no matter how it is obtained (by physical examination or by interview or by laboratory assay), it is measured unreliably in the sense that, were the individual to be measured again under similar conditions, the second value would differ to some extent from the first. Imagine a subject's being repeatedly measured on the variable of interest under as close to uniform conditions as possible, and let T denote the mean of the many hypothetical replicate measurements on him. T is referred to in psychometrics as the subject's "true score" (Lord and Novick, 1968), but less image-laden expressions are "error-free score," "steady-state value," and "signal." A single measurement X will differ from T for any number of reasons: random coding errors, misunderstanding by the subject of the interviewer's questions or by the interviewer of the subject's responses, inherent lability of the characteristic, or imperfect calibration of a measuring device. If e represents the difference between a single observation on a subject, X, and its underlying mean value, T, the classical linear model for an observed score is obtained,

$$X = T + e. \tag{1.1}$$

In a population of subjects, the error-free score T will vary about some mean value μ with a variance of σ_T^2. For a single subject, the random error e will vary about a mean of zero. Under the assumption that the distribution of the errors is independent of the value of T, e has a variance of σ_e^2 no matter what the value of T, and therefore the

variance of X is

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2. \tag{1.2}$$

In words, (1.2) expresses the phenomenon that there are two components to the variability among a series of measurements on different subjects, variability among their steady-state values plus the variability of the random errors.

A single quantity that usefully expresses the relative magnitude of the two components of the variance of X is the *intraclass correlation coefficient of reliability* (the *reliability*, for short),

$$R = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2} \tag{1.3}$$

(Bartko, 1966; Ebel, 1951; Fisher, 1921; Shrout and Fleiss, 1979). As σ_e^2/σ_T^2 decreases, error constitutes a decreasing portion of what is observed, reliability therefore increases, and R approaches its maximum value of unity. As σ_e^2/σ_T^2 increases, error constitutes an increasing portion of what is observed, reliability therefore decreases, and R approaches its minimum value of zero. Problem 1.1 calls for a proof that R is a bona fide correlation coefficient. Notice that R, unlike the traditional product-moment correlation coefficient, is directly interpretable as a proportion of variance. It is the proportion of the variance of an observation due to subject-to-subject variability in error-free scores.

1.2. SOME CONSEQUENCES OF UNRELIABILITY

What makes the parameter R so important is that most of the untoward effects of unreliability are expressible as functions of it.

1.2.1. Attenuated Correlations

Suppose that a study is designed to estimate the correlation between two variables T and U, but that what are measurable are

$$X = T + e$$

and

$$Y = U + f,$$

with e and f being random measurement errors uncorrelated with each other and uncorrelated with T and U. Finally, suppose that the correlation between T and U is ρ_{TU}. The correlation between the two

observable quantities is then

$$\rho_{XY} = \rho_{TU}\sqrt{R_X R_Y}, \tag{1.4}$$

where R_X and R_Y are the reliabilities of X and Y (see Problem 1.2). Because $\sqrt{R_X R_Y}$ is always less than unity, ρ_{XY} will always be closer to zero than ρ_{TU} is. The effect of unreliability, therefore, is to *attenuate* correlations.

Suppose, for example, that $\rho_{TU} = 0.50$ but that $R_X = 0.7$ and $R_Y = 0.6$. The observable correlation is then only $\rho_{XY} = 0.50\sqrt{0.7 \times 0.6} = 0.32$. One consequence of attenuation is that a sample estimate of the observable correlation may fail to reach statistical significance, whereas a sample estimate of the correlation between the error-free scores might be significant. A more serious substantive consequence is that the proportion of shared variance between the two variables may be seriously underestimated. Instead of its being found to be $0.50^2 = 0.25$ in the present example, it would be calculated as only $0.32^2 = 0.10$.

The phenomenon of attenuated correlations is often cited as an example of the limitation on "validity" imposed by unreliability (Lord and Novick, 1968, p. 72). There is no gainsaying this limitation, but there are other equally or more serious consequences of unreliability that are less widely appreciated.

1.2.2. Increased Sample Sizes

Consider designing a simple comparative study involving two groups of patients, and suppose that a mean difference on the response variable of $\delta = \mu_1 - \mu_2$ is considered on clinical grounds to be so important that, if δ is the true underlying mean difference, the investigator wants the chances to be high that a significant difference between the groups will be declared. The required sample sizes in the two groups may be determined as follows.

Suppose the significance level of the test comparing the two means is α, and assume for simplicity that the sample sizes are large enough for the t ratio to be referable to the standard normal distribution. Assuming that a two-tailed test is employed, significance will be declared if the absolute value of the t ratio exceeds $z_{\alpha/2}$, the standard normal curve value cutting off the proportion $\alpha/2$ in the upper tail. For example, $z_{\alpha/2} = 1.96$ for $\alpha = 0.05$. When the true difference between the means is δ, suppose the desired *power* (i.e., the chance of finding a significant difference) is $1 - \beta$. Let z_β denote the standard normal curve value cutting off the proportion β in the upper tail. For example, if 95% power is demanded, $1 - \beta = 0.95$ so that $\beta = 0.05$ and $z_\beta = 1.645$. Finally,

assume that the variances of the responses in the two groups are equal. If the responses were measured without error, the common variance would be σ_T^2. The required sample size in each group is then given by

$$n^* = \frac{2\sigma_T^2(z_{\alpha/2} + z_\beta)^2}{\delta^2} \tag{1.5}$$

(see the Appendix or Armitage, 1971, p. 186). If, however, random error intrudes into the measurements, the required sample size becomes

$$n = \frac{2(\sigma_T^2 + \sigma_e^2)(z_{\alpha/2} + z_\beta)^2}{\delta^2} = \frac{n^*}{R}, \tag{1.6}$$

which is always larger than n^*.

Suppose, for example, that change in diastolic blood pressure is to be used in the comparison of two independent treatment groups, that a two-tailed significance level of 0.05 is to be employed, and that a power of 80% is demanded if the difference in mean change between the groups is as large as 5 millimeters of mercury. Suppose finally that the standard deviation of error-free changes is 8 millimeters of mercury. Then,

$$z_{\alpha/2} = z_{0.025} = 1.96,$$

$$z_\beta = z_{0.20} = 0.842,$$

$$\delta = 5,$$

$$\sigma_T^2 = 8^2 = 64,$$

and

$$n^* = \frac{2(64)(1.96 + 0.842)^2}{5^2} = 40$$

patients are required in each group for a total sample size of 80.

Suppose, however, that the reliability with which change in diastolic blood pressure is measured is $R = 0.67$. The required number of patients per group becomes $n = 40/0.67 = 60$ for a total sample size of 120, a 50% increase over the earlier total. If the reliability were as high as $R = 0.80$, $n = 40/0.80 = 50$ for a total sample size of 100, a 25% increase. Thus unreliable measurement of the response variable increases the sample size necessary to detect an important treatment difference with a specified probability and therefore adds to the cost of the study.

1.2.3. Biased Sample Selection for Clinical Studies

A popular and valid strategy for selecting patients for comparative clinical trials and other kinds of studies is to recruit into the study only

patients who, inter alia, score above a minimum value at baseline on a given variable. One of the many good reasons for such a requirement is that the patients who enter the study should be sufficiently ill for the treatment to exhibit an effect. Let A denote the value of the threshold criterion, μ the mean value of the variable in the population of patients from which the sample will be drawn, and σ_T the standard deviation of error-free scores. The intent is to admit only those patients whose error-free score, T, exceeds A, but in actuality patients will be admitted if their observed score, $X = T + e$, exceeds A; some of these patients will have an error-free score less than A, and will exceed the threshold because of random error.

Therefore, the resulting sample will contain some patients who technically should not have been included. If there is random assignment of patients to treatment, no bias will be introduced into the comparison of treatments by these so-called *false positives*. Rather, the precision of treatment comparisons will be adversely affected by the bias in the sample as a whole because some patients will have been treated who were not severely ill enough to exhibit much response (Goldman, 1976). The biased nature of the sample is a special case of *regression to the mean*, the tendency for subjects whose observed values on some variable are above or below the mean of their population to have error-free values closer to the mean than the observed values (Davis, 1976).

The false-positive rate is the proportion of all patients, among those whose observed score exceeds A, whose error-free score is actually less than A. Define $C = (A - \mu)/\sigma_T$, the number of standard deviation units that the threshold criterion is away from the population mean. Table 1.1 tabulates the false-positive rate as a function of C and of R under the assumption that T and e are normally distributed. (The reader is asked in Problem 1.3 to derive the equation for the false-positive rate. The equation was solved to produce Table 1.1 using the tables of the bivariate normal distribution published by the National Bureau of Standards, 1959.)

For a fixed reliability R, the false-positive rate is seen in Table 1.1 to increase as C increases. This makes intuitive sense when it is realized that C is an indicator of where the threshold criterion A lies relative to the mean of the population from which the sample will be drawn. When C is negative, the threshold lies below the mean and only a minority of all patients will have error-free scores below A and will therefore be subject to being erroneously scored above A. When C is positive, the threshold lies above the mean and a majority of all patients will have error-free scores below A and therefore will be subject to being erroneously scored above A. In summary, the rarer the extreme group

Table 1.1. False-positive rate for the selection of a sample on the basis of a score's exceeding a specified value A

	Reliability (R)				
C^a	.50	.75	.85	.90	.95
−2.0	.01	.01	.01	0	0
−1.5	.03	.02	.02	.01	.01
−1.0	.07	.05	.04	.03	.02
−0.5	.14	.09	.07	.06	.04
0	.25	.17	.12	.10	.07
0.5	.40	.26	.20	.16	.11
1.0	.55	.37	.28	.22	.15
1.5	.70	.49	.39	.30	.21
2.0	.82	.61	.47	.38	.26

$^aC = (A - \mu)/\sigma_T$, where μ and σ_T are the mean and standard deviation of the error-free scores.

one intends to draw from the population, the larger the false-positive rate.

For a fixed value of C, on the other hand, the false-positive rate decreases as the reliability increases. This, too, makes intuitive sense, but what is distressing is how slowly the false-positive rate approaches zero as a function of R. Consider, for example, the value $C = 1.0$, corresponding to the intended selection of patients in the upper 16% of the distribution. When the reliability is 0.90, over a fifth of the patients included in the sample should not have been. Even for a reliability as high as 0.95, the false-positive rate is 15%.

It is clear from the preceding examples that no universally applicable standards are possible for what constitutes poor, fair, or good reliability. In general, values of R below 0.4 or so may be taken to represent poor reliability, values above 0.75 or so may be taken to represent excellent reliability, and values between 0.4 and 0.75 may be taken to represent fair to good reliability.

Several other untoward consequences of unreliability have been documented (Cochran, 1968; Fleiss and Shrout, 1977; Shrout and Fleiss 1981; see also Sections 2.1 and 7.1 in this book), all presupposing knowledge of the value of the reliability coefficient R. Sections 1.3 and 1.5 describe the two most important kinds of reliability studies that permit one to estimate R. The appropriate time to conduct a reliability study is before one's major research study is undertaken, not during or after it. As shown in Sections 1.3–1.5, the results of the former may and

should be used in the design of the latter. The reliability study need not involve a large number of subjects. Usually 15–20 will be enough for a quantitative variable, but more will be required for estimating the reliability of a categorical variable. No matter how reliable a measure has been found to be in the past, reliability should be assessed again prior to a new study. There is no guarantee, after all, that reliability will continue to be high for a new group of examiners obtaining measurements on a new sample of patients.

1.3. THE SIMPLE REPLICATION RELIABILITY STUDY

Suppose that each of a sample of N subjects in a reliability study is measured several times on the variable under investigation. For example, several blood samples may be drawn from a patient and each sample subjected to a laboratory assay for the activity of a certain enzyme. Or, a patient may be evaluated on a rating scale by a few nurses selected at random from a larger pool of available nurses. Or, a 24-hour recording of the electrical functioning of a patient's heart may be obtained and subjected to several independent computer analyses. In each of these examples it is arbitrary which measurement on a patient is designated the first, which the second, and so on. There is no structure to the replicate measurements in the sense that nothing ties the first or second measurement on one patient to the first or second on another. In the terminology of the analysis of variance, the study conforms to a one-way *random effects* model (Armitage, 1971, p. 198).

The results of this simple kind of reliability or reproducibility study may be summarized as in Table 1.2. For a typical subject, say the ith, k_i

Table 1.2. Layout of data from a simple replication reliability study

Subject	Number of Measurements	Mean	Variance
1	k_1	$\bar{X}_1$	s_1^2
⋮			
i	k_i	$\bar{X}_i$	s_i^2
⋮			
N	k_N	$\bar{X}_N$	s_N^2
Total	K	$\bar{X}$	S^2

is the number of replicate measurements on him or her, $\bar{X}_i$ is the mean of the k_i measurements, and s_i^2 is their variance. Thus, if $X_{i1}, X_{i2}, \ldots, X_{ik_i}$ represent the k_i measurements on Subject i,

$$\bar{X}_i = \frac{1}{k_i} \sum_{j=1}^{k_i} X_{ij} \tag{1.7}$$

and

$$s_i^2 = \frac{1}{k_i - 1} \sum_{j=1}^{k_i} (X_{ij} - \bar{X}_i)^2. \tag{1.8}$$

In the final row of Table 1.2, K is the total number of measurements,

$$K = \sum_{i=1}^{N} k_i; \tag{1.9}$$

$\bar{X}$ is the overall mean,

$$\bar{X} = \frac{1}{K} \sum_i \sum_j X_{ij} = \frac{1}{K} \sum_i k_i \bar{X}_i; \tag{1.10}$$

and S^2 is the overall variance,

$$S^2 = \frac{1}{K - 1} \sum_i \sum_j (X_{ij} - \bar{X})^2. \tag{1.11}$$

Table 1.3 presents the results for a random sample of 10 patients out

Table 1.3. Results of a simple replication reliability study ($k = 2$ measurements per patient)

Patient	Mean	Variance
1	0.235	0.0265
2	0.115	0.0005
3	0.140	0.0008
4	0	0
5	0.385	0.0061
6	2.655	0.0005
7	0.065	0.0013
8	0.375	0.0085
9	0.580	0.0002
10	3.900	0.0338
Total	0.845	1.6711

of a total of 63 whose 24-hour Holter tape recordings of the heart's electrical functioning were read and analyzed two independent times by computer (Clark, Rolnitzky, Miller, et al., 1981). The variable being analyzed in Table 1.3 is ln(VPD + 1), the natural logarithm of one plus the computer-calculated number of ventricular premature depolarizations (VPDs) per hour. The fact that only one patient had a variance of zero means that for only that patient's tape were the two computer analyses in agreement.

The variability among the means appears to be appreciably greater than the average of the several within-patient variances, suggesting that within-patient variability (which estimates σ_e^2) is much smaller than between-patient variability (which is informative about σ_T^2). The analysis of variance provides a quantitative rather than qualitative description of the two components of variability.

The general analysis of variance table for data arrayed as in Table 1.2 appears in the left-hand portion of Table 1.4, and the quantities obtained by applying the formulas to the data in Table 1.3 appear in the right-hand portion. The column headings are abbreviations of *degrees of freedom*, *sum of squares*, *mean square* (the ratio of the sum of squares to the corresponding number of degrees of freedom), and *expected mean square* (the underlying statistical quantity that the mean square estimates). The constant k_0 that appears in the expected value of the between-patient mean square is equal to

$$k_0 = \bar{k} - \frac{s_k^2}{N\bar{k}}, \tag{1.12}$$

where $\bar{k}$ and s_k^2 are the mean and variance of the numbers of replicate

Table 1.4. Analysis of variance for the results of a simple replication reliability study

Source of Variation	In General				For the Data in Table 1.3		
	df	SS	MS	E(MS)	df	SS	MS
Between patients	$N-1$	$\sum k_i(\bar{X}_i - \bar{X})^2$	BMS	$\sigma_e^2 + k_0\sigma_T^2$	9	31.6726	3.5192
Within patients	$K-N$	$\sum (k_i - 1)s_i^2$	WMS	σ_e^2	10	0.0782	0.0078
Total	$K-1$	$(K-1)S^2$			19	31.7508	

measurements,

$$\bar{k} = \frac{K}{N} \tag{1.13}$$

and

$$s_k^2 = \frac{1}{N-1} \sum_i (k_i - \bar{k})^2 \tag{1.14}$$

(see Problem 1.4). If, as in Table 1.3, these numbers are constant (i.e., if $k_1 = \ldots = k_N = k$, say), then $\bar{k} = k$, $s_k^2 = 0$, and $k_0 = k$; otherwise, k_0 will be slightly less than $\bar{k}$. (Problem 1.5 calls for analyzing a set of data in which the k_i's vary.)

The within-subject mean square (WMS) is seen to be an unbiased estimator of σ_e^2, the component of variance due to random error. It is therefore convenient for current purposes to let s_e^2 designate WMS. The quantity

$$s_T^2 = \frac{\text{BMS} - \text{WMS}}{k_0} \tag{1.15}$$

is seen to be an unbiased estimator of σ_T^2, the component of variance due to error-free variability among subjects. An estimator of the intraclass correlation coefficient is then

$$\hat{R} = \frac{s_T^2}{s_T^2 + s_e^2} = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (k_0 - 1)\,\text{WMS}}. \tag{1.16}$$

Even though $\hat{R}$ is the ratio of two unbiased estimators, it itself is a slightly biased estimator of the parameter R (Olkin and Pratt, 1958). The bias becomes negligible as N increases.

For the data at hand, the estimated component of variance due to random measurement error is

$$s_e^2 = \text{WMS} = 0.0078. \tag{1.17}$$

Because the constant number of replicate readings per patient is $k = 2$, therefore $k_0 = 2$ and the estimated component of variance due to the variability of error-free scores is

$$s_T^2 = \frac{\text{BMS} - \text{WMS}}{k} = \frac{3.5192 - 0.0078}{2} = 1.7557. \tag{1.18}$$

The square root of WMS, s_e, is referred to as the *standard error of measurement*. One of its most important uses follows. For a single subject, the error-free score T may be considered an unknown parameter. If a single observed measurement X were available on that

subject, an approximate 95% confidence interval for his or her underlying T is $X \pm 2s_e$. For the current example, $s_e = \sqrt{0.0078} = 0.09$. If, for example, a subject has an observed value of $X = 1.54$ (i.e., the calculated number of ventricular premature depolarizations per hour is VPD = 3.67, so $X = \ln(3.67 + 1) = 1.54$), an approximate 95% confidence interval for that subject's error-free value T is $1.54 \pm 2 \times 0.09$, or the interval from 1.36 to 1.72. The corresponding limits for the associated error-free value of VDP are $\exp(1.36) - 1 = 2.90$ and $\exp(1.72) - 1 = 4.58$. If the subject is measured m times and the mean of the replicate measurements $\bar{X}$ is taken, an approximate 95% confidence interval for T is $\bar{X} \pm 2s_e/\sqrt{m}$. The value of m may be less than or greater than any of the k_i's, which are particular to the reliability study. For example, if a subject has a mean of $X = 2.05$ based on $m = 3$ replicate readings, an approximate 95% confidence interval for that subject's error-free value is $2.05 \pm 2 \times 0.09/\sqrt{3}$, or the interval from 1.95 to 2.15. The limits for the error-free value of VPD are $\exp(1.95) - 1 = 6.03$ and $\exp(2.15) - 1 = 7.58$.

The estimated intraclass correlation coefficient for the variable being analyzed is

$$\hat{R} = \frac{s_T^2}{s_T^2 + s_e^2} = \frac{1.7557}{1.7557 + 0.0078} = 0.996, \tag{1.19}$$

a value indicating nearly perfect reliability. If the random quantities T and e are normally distributed, an approximate one-sided $100(1 - \alpha)\%$ confidence interval for R is

$$R \geq \frac{\dfrac{\text{BMS}}{\text{WMS}} - F_{N-1,K-N,\alpha}}{\dfrac{\text{BMS}}{\text{WMS}} + (k_0 - 1)F_{N-1,K-N,\alpha}}, \tag{1.20}$$

where $F_{\nu_1,\nu_2,p}$ denotes the tabulated value of the F distribution with ν_1 and ν_2 degrees of freedom cutting off the proportion p in the upper tail. When the k_i's are equal, the interval in (1.20) is exact (Wald, 1940). For a 95% confidence interval in the current example, $F_{9,10,0.05} = 3.02$ and

$$R \geq \frac{\dfrac{3.5192}{0.0078} - 3.02}{\dfrac{3.5192}{0.0078} + (2 - 1) \times 3.02} = 0.987. \tag{1.21}$$

The confidence interval for R serves at least two purposes. One, which is usually the less important, is to provide a test of the hypothesis that the underlying value of R is zero (i.e., that the measurements are so unreliable that differences between subjects are due exclusively to ran-