

Sample Size Determination and Power

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Harvey Goldstein, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Joseph B. Kadane, Jozef L. Teugels*

A complete list of the titles in this series appears at the end of this volume.

Sample Size Determination and Power

THOMAS P. RYAN

Institute for Statistics Education, Arlington, Virginia and
Northwestern University, Evanston, Illinois

WILEY

Cover design: John Wiley & Sons, Inc.

Cover image: © Thomas P. Ryan

Copyright © 2013 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Ryan, Thomas P., 1945–

Sample size determination and power / Thomas P. Ryan.

p. ; cm.

Includes bibliographical references and index.

ISBN 978-1-118-43760-5 (cloth)

I. Title.

[DNLM: 1. Sample Size. 2. Clinical Trials as Topic. 3. Mathematical Computing. 4. Regression Analysis. 5. Sampling Studies. WA 950]

615.5072'4—dc23

2013000329

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface	xv
1 Brief Review of Hypothesis Testing Concepts/Issues and Confidence Intervals	1
1.1 Basic Concepts of Hypothesis Testing,	1
1.2 Review of Confidence Intervals and Their Relationship to Hypothesis Tests,	5
1.3 Sports Applications,	9
1.4 Observed Power, Retrospective Power, Conditional Power, and Predictive Power,	9
1.5 Testing for Equality, Equivalence, Noninferiority, or Superiority,	10
1.5.1 Software,	11
References,	12
Exercises,	14
2 Methods of Determining Sample Sizes	17
2.1 Internal Pilot Study Versus External Pilot Study,	20
2.2 Examples: Frequentist and Bayesian,	24
2.2.1 Bayesian Approaches,	30
2.2.2 Probability Assessment Approach,	31
2.2.3 Reproducibility Probability Approach,	32
2.2.4 Competing Probability Approach,	32
2.2.5 Evidential Approach,	32
2.3 Finite Populations,	32

- 2.4 Sample Sizes for Confidence Intervals, 33
 - 2.4.1 Using the Finite Population Correction Factor, 36
 - 2.4.1.1 Estimating Population Totals, 38
- 2.5 Confidence Intervals on Sample Size and Power, 39
- 2.6 Specification of Power, 39
- 2.7 Cost of Sampling, 40
- 2.8 Ethical Considerations, 40
- 2.9 Standardization and Specification of Effect Sizes, 42
- 2.10 Equivalence Tests, 43
- 2.11 Software and Applets, 45
- 2.12 Summary, 47
 - References, 47
 - Exercises, 53

3 Means and Variances

57

- 3.1 One Mean, Normality, and Known Standard Deviation, 58
 - 3.1.1 Using the Coefficient of Variation, 65
- 3.2 One Mean, Standard Deviation Unknown, Normality Assumed, 66
- 3.3 Confidence Intervals on Power and/or Sample Size, 67
- 3.4 One Mean, Standard Deviation Unknown, Nonnormality Assumed, 70
- 3.5 One Mean, Exponential Distribution, 71
- 3.6 Two Means, Known Standard Deviations—Independent Samples, 71
 - 3.6.1 Unequal Sample Sizes, 74
- 3.7 Two Means, Unknown but Equal Standard Deviations—Independent Samples, 74
 - 3.7.1 Unequal Sample Sizes, 76
- 3.8 Two Means, Unequal Variances and Sample Sizes—Independent Samples, 77
- 3.9 Two Means, Unknown and Unequal Standard Deviations—Independent Samples, 77
- 3.10 Two Means, Known and Unknown Standard Deviations—Dependent Samples, 78
- 3.11 Bayesian Methods for Comparing Means, 81
- 3.12 One Variance or Standard Deviation, 81
- 3.13 Two Variances, 83
- 3.14 More Than Two Variances, 84

- 3.15 Confidence Intervals, 84
 - 3.15.1 Adaptive Confidence Intervals, 85
 - 3.15.2 One Mean, Standard Deviation Unknown—With Tolerance Probability, 85
 - 3.15.3 Difference Between Two Independent Means, Standard Deviations Known and Unknown—With and Without Tolerance Probability, 88
 - 3.15.4 Difference Between Two Paired Means, 90
 - 3.15.5 One Variance, 91
 - 3.15.6 One-Sided Confidence Bounds, 92
- 3.16 Relative Precision, 93
- 3.17 Computing Aids, 94
- 3.18 Software, 94
- 3.19 Summary, 95
 - Appendix, 95
 - References, 96
 - Exercises, 99

4 Proportions and Rates

103

- 4.1 One Proportion, 103
 - 4.1.1 One Proportion—With Continuity Correction, 107
 - 4.1.2 Software Disagreement and Rectification, 108
 - 4.1.3 Equivalence Tests and Noninferiority Tests for One Proportion, 109
 - 4.1.4 Confidence Interval and Error of Estimation, 110
 - 4.1.5 One Proportion—Exact Approach, 113
 - 4.1.6 Bayesian Approaches, 115
- 4.2 Two Proportions, 115
 - 4.2.1 Two Proportions—With Continuity Correction, 119
 - 4.2.2 Two Proportions—Fisher's Exact Test, 121
 - 4.2.3 What Approach Is Recommended?, 122
 - 4.2.4 Correlated Proportions, 123
 - 4.2.5 Equivalence Tests for Two Proportions, 124
 - 4.2.6 Noninferiority Tests for Two Proportions, 125
 - 4.2.7 Need for Pilot Study?, 125
 - 4.2.8 Linear Trend in Proportions, 125
 - 4.2.9 Bayesian Method for Estimating the Difference of Two Binomial Proportions, 126
- 4.3 Multiple Proportions, 126

- 4.4 Multinomial Probabilities and Distributions, 129
- 4.5 One Rate, 130
 - 4.5.1 Pilot Study Needed?, 132
- 4.6 Two Rates, 132
- 4.7 Bayesian Sample Size Determination Methods for Rates, 135
- 4.8 Software, 135
- 4.9 Summary, 136
 - Appendix, 136
 - References, 140
 - Exercises, 144

5 Regression Methods and Correlation 145

- 5.1 Linear Regression, 145
 - 5.1.1 Simple Linear Regression, 146
 - 5.1.2 Multiple Linear Regression, 150
 - 5.1.2.1 Application: Predicting College Freshman Grade Point Average, 155
- 5.2 Logistic Regression, 155
 - 5.2.1 Simple Logistic Regression, 156
 - 5.2.1.1 Normally Distributed Covariate, 158
 - 5.2.1.2 Binary Covariate, 162
 - 5.2.2 Multiple Logistic Regression, 163
 - 5.2.2.1 Measurement Error, 165
 - 5.2.3 Polytomous Logistic Regression, 165
 - 5.2.4 Ordinal Logistic Regression, 166
 - 5.2.5 Exact Logistic Regression, 167
- 5.3 Cox Regression, 167
- 5.4 Poisson Regression, 169
- 5.5 Nonlinear Regression, 172
- 5.6 Other Types of Regression Models, 172
- 5.7 Correlation, 172
 - 5.7.1 Confidence Intervals, 174
 - 5.7.2 Intraclass Correlation, 175
 - 5.7.3 Two Correlations, 175
- 5.8 Software, 176
- 5.9 Summary, 177
 - References, 177
 - Exercises, 180

6 Experimental Designs**183**

- 6.1 One Factor—Two Fixed Levels, 184
 - 6.1.1 Unequal Sample Sizes, 186
- 6.2 One Factor—More Than Two Fixed Levels, 187
 - 6.2.1 Multiple Comparisons and Dunnett's Test, 192
 - 6.2.2 Analysis of Means (ANOM), 193
 - 6.2.3 Unequal Sample Sizes, 195
 - 6.2.4 Analysis of Covariance, 196
 - 6.2.5 Randomized Complete Block Designs, 197
 - 6.2.6 Incomplete Block Designs, 198
 - 6.2.7 Latin Square Designs, 199
 - 6.2.7.1 Graeco-Latin Square Designs, 202
- 6.3 Two Factors, 203
- 6.4 2^k Designs, 205
 - 6.4.1 2^2 Design with Equal and Unequal Variances, 206
 - 6.4.2 Unreplicated 2^k Designs, 206
 - 6.4.3 Software for 2^k Designs, 208
- 6.5 2^{k-p} Designs, 209
- 6.6 Detecting Conditional Effects, 210
- 6.7 General Factorial Designs, 211
- 6.8 Repeated Measures Designs, 212
 - 6.8.1 Crossover Designs, 215
 - 6.8.1.1 Software, 217
- 6.9 Response Surface Designs, 218
- 6.10 Microarray Experiments, 219
 - 6.10.1 Software, 220
- 6.11 Other Designs, 220
 - 6.11.1 Plackett–Burman Designs, 220
 - 6.11.2 Split-Plot and Strip-Plot Designs, 222
 - 6.11.3 Nested Designs, 224
 - 6.11.4 Ray designs, 225
- 6.12 Designs for Nonnormal Responses, 225
- 6.13 Designs with Random Factors, 227
- 6.14 Zero Patient Design, 228
- 6.15 Computer Experiments, 228
- 6.16 Noninferiority and Equivalence Designs, 229
- 6.17 Pharmacokinetic Experiments, 229
- 6.18 Bayesian Experimental Design, 229

- 6.19 Software, 230
- 6.20 Summary, 232
 - Appendix, 233
 - References, 234
 - Exercises, 239

7 Clinical Trials 243

- 7.1 Clinical Trials, 245
 - 7.1.1 Cluster Randomized Trials, 247
 - 7.1.2 Phase II Trials, 247
 - 7.1.2.1 Phase II Cancer Trials, 247
 - 7.1.3 Phase III Trials, 247
 - 7.1.4 Longitudinal Clinical Trials, 248
 - 7.1.5 Fixed Versus Adaptive Clinical Trials, 248
 - 7.1.6 Noninferiority Trials, 249
 - 7.1.7 Repeated Measurements, 249
 - 7.1.8 Multiple Tests, 250
 - 7.1.9 Use of Internal Pilot Studies for Clinical Trials, 250
 - 7.1.10 Using Historical Controls, 250
 - 7.1.11 Trials with Combination Treatments, 251
 - 7.1.12 Group Sequential Trials, 251
 - 7.1.13 Vaccine Efficacy Studies, 251
- 7.2 Bioequivalence Studies, 251
- 7.3 Ethical Considerations, 252
- 7.4 The Use of Power in Clinical Studies, 252
- 7.5 Preclinical Experimentation, 253
- 7.6 Pharmacodynamic, Pharmacokinetic, and Pharmacogenetic Experiments, 253
- 7.7 Method of Competing Probability, 254
- 7.8 Bayesian Methods, 255
- 7.9 Cost and Other Sample Size Determination Methods for Clinical Trials, 256
- 7.10 Meta-Analyses of Clinical Trials, 256
- 7.11 Miscellaneous, 257
- 7.12 Survey Results of Published Articles, 259
- 7.13 Software, 260
- 7.14 Summary, 263

References, 263

Exercises, 275

8 Quality Improvement 277

- 8.1 Control Charts, 277
 - 8.1.1 Shewhart Measurement Control Charts, 278
 - 8.1.2 Using Software to Determine Subgroup Size, 281
 - 8.1.2.1 \bar{X} -Chart, 282
 - 8.1.2.2 S -Chart and S^2 -Chart, 284
 - 8.1.3 Attribute Control Charts, 286
 - 8.1.4 CUSUM and EWMA Charts, 289
 - 8.1.4.1 Subgroup Size Considerations for CUSUM Charts, 290
 - 8.1.4.2 CUSUM and EWMA Variations, 291
 - 8.1.4.3 Subgroup Size Determination for CUSUM and EWMA Charts and Their Variations, 291
 - 8.1.4.4 EWMA Applied to Autocorrelated Data, 293
 - 8.1.5 Adaptive Control Charts, 293
 - 8.1.6 Regression and Cause-Selecting Control Charts, 293
 - 8.1.7 Multivariate Control Charts, 295
- 8.2 Medical Applications, 296
- 8.3 Process Capability Indices, 297
- 8.4 Tolerance Intervals, 298
- 8.5 Measurement System Appraisal, 300
- 8.6 Acceptance Sampling, 300
- 8.7 Reliability and Life Testing, 301
- 8.8 Software, 301
- 8.9 Summary, 302
 - References, 302
 - Exercises, 305

9 Survival Analysis and Reliability 307

- 9.1 Survival Analysis, 307
 - 9.1.1 Logrank Test, 308
 - 9.1.1.1 Freedman Method, 311
 - 9.1.1.2 Other Methods, 312
 - 9.1.2 Wilcoxon–Breslow–Gehan Test, 313
 - 9.1.3 Tarone–Ware Test, 313

- 9.1.4 Other Tests, 314
- 9.1.5 Cox Proportional Hazards Model, 314
- 9.1.6 Joint Modeling of Longitudinal and Survival Data, 315
- 9.1.7 Multistage Designs, 316
- 9.1.8 Comparison of Software and Freeware, 316
- 9.2 Reliability Analysis, 317
- 9.3 Summary, 318
- References, 319
- Exercise, 321

10 Nonparametric Methods 323

- 10.1 Wilcoxon One-Sample Test, 324
 - 10.1.1 Wilcoxon Test for Paired Data, 327
- 10.2 Wilcoxon Two-Sample Test (Mann–Whitney Test), 327
 - 10.2.1 van Elteren Test—A Stratified Mann–Whitney Test, 331
- 10.3 Kruskal–Wallis One-Way ANOVA, 331
- 10.4 Sign Test, 331
- 10.5 McNemar’s Test, 334
- 10.6 Contingency Tables, 334
- 10.7 Quasi-Likelihood Method, 334
- 10.8 Rank Correlation Coefficients, 335
- 10.9 Software, 335
- 10.10 Summary, 336
 - References, 336
 - Exercises, 339

11 Miscellaneous Topics 341

- 11.1 Case–Control Studies, 341
- 11.2 Epidemiology, 342
- 11.3 Longitudinal Studies, 342
- 11.4 Microarray Studies, 343
- 11.5 Receiver Operating Characteristic ROC Curves, 343
- 11.6 Meta-Analyses, 343
- 11.7 Sequential Sample Sizes, 343
- 11.8 Sample Surveys, 344
 - 11.8.1 Vegetation Surveys, 344

11.9	Cluster Sampling, 345	
11.10	Factor Analysis, 346	
11.11	Multivariate Analysis of Variance and Other Multivariate Methods, 346	
11.12	Structural Equation Modeling, 348	
11.13	Multilevel Modeling, 349	
11.14	Prediction Intervals, 349	
11.15	Measures of Agreement, 350	
11.16	Spatial Statistics, 350	
11.17	Agricultural Applications, 350	
11.18	Estimating the Number of Unseen Species, 351	
11.19	Test Reliability, 351	
11.20	Agreement Studies, 351	
11.21	Genome-wide Association Studies, 351	
11.22	National Security, 352	
11.23	Miscellaneous, 352	
11.24	Summary, 353	
	References, 354	
Answers to Selected Exercises		363
Index		369

Preface

Determining a good sample size to use in a scientific study is of utmost importance, especially in clinical studies with some participants receiving a placebo or nothing at all and others taking a drug whose efficacy has not been established. It is imperative that a large enough sample be used so that an effect that is large enough to be of practical significance has a high probability of being detected from the study. That is, the study should have sufficient *power*. It is also important that sample sizes not be larger than necessary so that the cost of a study not be any larger than necessary and to minimize risk to human subjects in drug studies.

Compared to other subjects in the field of statistics, there is a relative paucity of books on sample size determination and power, especially general purpose books. The classic book on the subject has for decades been Jacob Cohen's *Statistical Power Analysis for the Behavioral Sciences*, the second edition of which was published in 1988. That book is oriented, as the title indicates, toward the behavioral sciences, with the statistical methodology being quite useful in the behavioral sciences. The second edition has 567 numbered pages, 208 of which are tables, reflecting the "noncomputer" age in which the two editions of the book were written. In contrast, the relatively recent book by Patrick Dattalo, *Determining Sample Size: Balancing Power, Precision, and Practicality* (2008), which is part of the series in Pocket Guides to Social Work Research Methods, is 167 pages with more than 20% consisting of tables and screen displays reflecting the now heavy reliance on software for sample size determination. An even smaller book is *Sample Size Methodology* (1990) by Desu and Raghavarao at 135 pages, while *How Many Subjects: Statistical Power Analysis in Research* (1987) by Kraemer and Thiemann is just 120 pages and was stated in a review as being an extension of a 1985 journal article by Kraemer. *Sample-Size Determination* (1964) by Mace is larger at 226 pages and *Sample Size Choice: Charts for Experimenters*, 2nd ed. (1991) by Odeh and Fox is 216 pages. Thus, some rather small books have been published on the subject, with almost all of these books having been published over 20 years ago.

At the other extreme in terms of size, focus, and mathematical sophistication, there are books on sample determination for clinical studies, such as *Sample Size Calculations in Clinical Research*, 2nd ed. (2008) by Chow, Shao, and Wang, that are mathematically sophisticated, with the title of this book perhaps suggesting that. A similar recent book is *Sample Sizes for Clinical Trials* (2010) by Julious, whereas *Sample Size Calculations: Practical Methods for Engineers and Scientists* (2010) by Mathews is oriented toward engineering and industrial applications.

There are additional statistical methods that are useful in fields other than behavioral sciences, social sciences, and clinical trials, however, and during the past two decades new needs for sample size determination have arisen in fields that are part of the advancement of science, such as microarray experiments.

Although many formulas are given in Cohen's book, they are not derived in either the chapters or chapter appendices, so the inquisitive reader is left wondering how the formulas came about.

Software is also not covered in Cohen's book, nor is software discussed in the books by Mathews, Julious or Chow, Shao, and Wang. Software and Java applets for sample size determination are now fairly prevalent and, of course, are more useful than tables since theoretically there are an infinite number of values that could be entered for one or more parameter values. There was a need for a book that has a broader scope than Cohen's book and that gives some of the underlying math for interested readers, as well as having a strong software focus, along the lines of Dattalo's book, but is not too mathematical for a general readership. No such book met these requirements at the time of writing, which is why this book was written.

This book can be used as a reference book as well as a textbook in special topics courses. Software discussion and illustration is integrated with the subject matter, and there is also a summary section on software at the end of most chapters. Mixing software discussion with subject matter may seem unorthodox, but I believe this is the best way to cover the material since almost every experimenter faced with software determination will probably feel the need to use software and should know what is available in terms of various software and applets. So the book is to a significant extent a software guide, with considerable discussion about the capabilities of each software package. There is also a very large number of references, considerably more than in any other book on the subject.

THOMAS P. RYAN