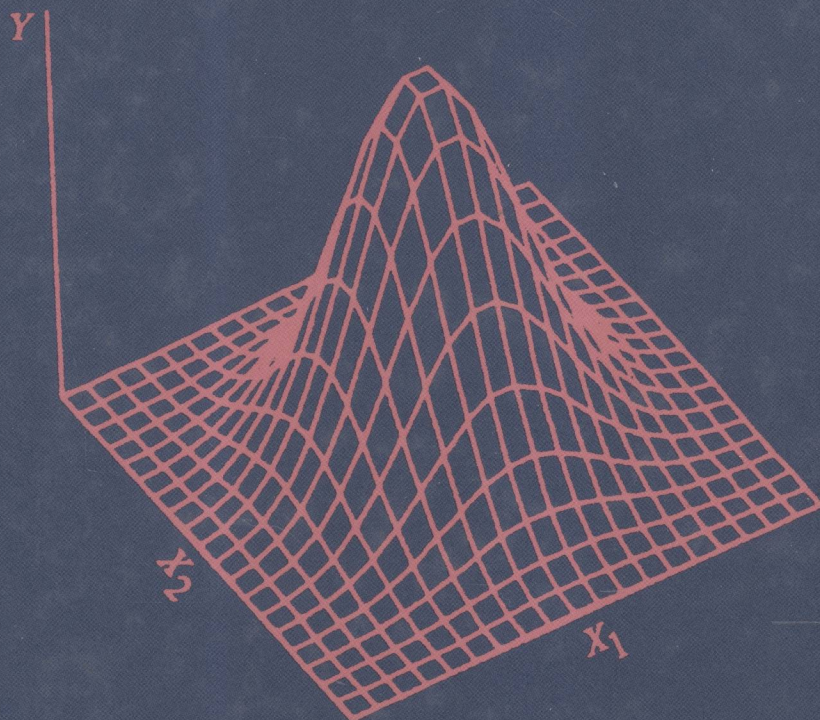


# **HANDBOOK OF REGRESSION AND MODELING**

Applications for the Clinical and  
Pharmaceutical Industries



**Daryl S. Paulson**



Chapman & Hall/CRC  
Taylor & Francis Group



Chapman & Hall/CRC Biostatistics Series

R971-62

P332

# **Handbook of Regression and Modeling**

## **Applications for the Clinical and Pharmaceutical Industries**

**Daryl S. Paulson**

BioScience Laboratories, Inc.

Bozeman, Montana, U.S.A.



E2007000855



**Chapman & Hall/CRC**

Taylor & Francis Group

Boca Raton London New York

---

Chapman & Hall/CRC is an imprint of the  
Taylor & Francis Group, an informa business

Chapman & Hall/CRC  
Taylor & Francis Group  
6000 Broken Sound Parkway NW, Suite 300  
Boca Raton, FL 33487-2742

© 2007 by Taylor & Francis Group, LLC  
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works  
Printed in the United States of America on acid-free paper  
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-57444-610-X (Hardcover)  
International Standard Book Number-13: 978-1-57444-610-4 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC) 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

---

#### Library of Congress Cataloging-in-Publication Data

---

Paulson, Daryl S., 1947-

Handbook of regression and modeling : applications for the clinical and pharmaceutical industries / Daryl S. Paulson.

p. ; cm. -- (Biostatistics ; 18)

Includes index.

ISBN-13: 978-1-57444-610-4 (hardcover : alk. paper)

ISBN-10: 1-57444-610-X (hardcover : alk. paper)

1. Medicine--Research--Statistical methods--Handbooks, manuals, etc. 2. Regression analysis--Handbooks, manuals, etc. 3. Drugs--Research--Statistical methods--Handbooks, manuals, etc. 4. Clinical trials--Statistical methods--Handbooks, manuals, etc. I. Title. II. Series: Biostatistics (New York, N.Y.) ; 18.

[DNLM: 1. Clinical Medicine. 2. Regression Analysis. 3. Biometry--methods. 4. Drug Industry. 5. Models, Statistical. WA 950 P332h 2007]

R853.S7P35 2007  
610.72'7--dc22

2006030225

---

Visit the Taylor & Francis Web site at  
<http://www.taylorandfrancis.com>

and the CRC Press Web site at  
<http://www.crcpress.com>



Chapman & Hall/CRC Biostatistics Series

# **Handbook of Regression and Modeling**

**Applications for the Clinical and  
Pharmaceutical Industries**

Series Editor

**Shein-Chung Chow, Ph.D.**

*Professor*

*Department of Biostatistics and Bioinformatics*

*Duke University School of Medicine*

*Durham, North Carolina, U.S.A.*

*Department of Statistics*

*National Cheng-Kung University*

*Tainan, Taiwan*

1. *Design and Analysis of Animal Studies in Pharmaceutical Development*, Shein-Chung Chow and Jen-pei Liu
2. *Basic Statistics and Pharmaceutical Statistical Applications*, James E. De Muth
3. *Design and Analysis of Bioavailability and Bioequivalence Studies, Second Edition, Revised and Expanded*, Shein-Chung Chow and Jen-pei Liu
4. *Meta-Analysis in Medicine and Health Policy*, Dalene K. Stangl and Donald A. Berry
5. *Generalized Linear Models: A Bayesian Perspective*, Dipak K. Dey, Sujit K. Ghosh, and Bani K. Mallick
6. *Difference Equations with Public Health Applications*, Lemuel A. Moyé and Asha Seth Kapadia
7. *Medical Biostatistics*, Abhaya Indrayan and Sanjeev B. Sarmukaddam
8. *Statistical Methods for Clinical Trials*, Mark X. Norleans
9. *Causal Analysis in Biomedicine and Epidemiology: Based on Minimal Sufficient Causation*, Mikel Aickin
10. *Statistics in Drug Research: Methodologies and Recent Developments*, Shein-Chung Chow and Jun Shao
11. *Sample Size Calculations in Clinical Research*, Shein-Chung Chow, Jun Shao, and Hansheng Wang
12. *Applied Statistical Design for the Researcher*, Daryl S. Paulson
13. *Advances in Clinical Trial Biostatistics*, Nancy L. Geller
14. *Statistics in the Pharmaceutical Industry, 3rd Edition*, Ralph Buncher and Jia-Yeong Tsay
15. *DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments*, David B. Allison, Grier P. Page, T. Mark Beasley, and Jode W. Edwards
16. *Basic Statistics and Pharmaceutical Statistical Applications, Second Edition*, James E. De Muth
17. *Adaptive Design Methods in Clinical Trials*, Shein-Chung Chow and Mark Chang
18. *Handbook of Regression and Modeling: Applications for the Clinical and Pharmaceutical Industries*, Daryl S. Paulson

---

# Preface

In 2003, I wrote a book, *Applied Statistical Designs for the Researcher* (Marcel Dekker, Inc.), in which I covered experimental designs commonly encountered in the pharmaceutical, applied microbiological, and healthcare-product-formulation industries. It included two sample evaluations, analysis of variance, factorial, nested, chi-square, exploratory data analysis, nonparametric statistics, and a chapter on linear regression. Many researchers need more than simple linear regression methods to meet their research needs. It is for those researchers that this regression analysis book is written.

Chapter 1 is an overview of statistical methods and elementary concepts for statistical model building.

Chapter 2 covers simple linear regression applications in detail.

Chapter 3 deals with a problem that many applied researchers face when collecting data of time–serial correlation (the actual response values of  $y$  are correlated with one another). This chapter lays the foundation for the discussion on multiple regression in Chapter 8.

Chapter 4 introduces multiple linear regression procedures and matrix algebra. The knowledge of matrix algebra is not a prerequisite, and Appendix II presents the basics in matrix manipulation. Matrix notation is used because those readers without specific statistical software that contains “canned” statistical programs can still perform the statistical analyses presented in this book. However, I assume that the reader will perform most of the computations using statistical software such as SPSS, SAS, or MiniTab. This chapter also covers strategies for checking the contribution of each  $x_i$  variable in a regression equation to assure that it is actually contributing. Partial  $F$ -tests are used in stepwise, forward selection, and backward elimination procedures.

Chapter 5 focuses on aspects of correlation analysis and those of determining the contribution of  $x_i$  variables using partial correlation analysis.

Chapter 6 discusses common problems encountered in multiple linear regression and the ways to deal with them. One problem is multiple collinearity, in which some of the  $x_i$  variables are correlated with other  $x_i$  variables and the regression equation becomes unstable in applied work. A number of procedures are explained to deal with such problems and a biasing method called ridge regression is also discussed.

Chapter 7 describes aspects of polynomial regression and its uses.

Chapter 8 aids the researcher in determining outlier values of the variables  $y$  and  $x$ . It also includes residual analysis schemas, such as standardized, Studentized, and jackknife residual analyses. Another important feature of

this chapter is leverage value identification, or identifying values,  $y$ s and  $x$ s, that have undue influence.

Chapter 9 applies indicator or dummy variables to an assortment of analyses.

Chapter 10 presents forward and stepwise selections of  $x_i$  variables, as well as backward elimination, in terms of statistical software.

Chapter 11 introduces covariance analysis, which combines regression and analysis of variance into one model.

The concepts presented in this book have been used for the past 25 years, in the clinical trials and new product development and formulation areas at BioScience Laboratories, Inc. They have also been used in analyzing data supporting studies submitted to the Food and Drug Administration (FDA) and the Environmental Protection Agency (EPA), and in my work as a statistician for the Association of Analytical Chemists (AOAC) in projects related to EPA regulation and Homeland Security.

This book has been two years in the making, from my standpoint. Certainly, it has not been solely an individual process on my part. I thank my friend and colleague, John A. Mitchell, PhD, also known as doctor for his excellent and persistent editing of this book, in spite of his many other duties at BioScience Laboratories, Inc. I also thank Tammy Anderson, my assistant, for again managing the entire manuscript process of this book, which is her sixth one for me. I also want to thank Marsha Paulson, my wife, for stepping up to the plate and helping us with the grueling final edit.

**Daryl S. Paulson, PhD**



---

# Author

**Daryl S. Paulson** is the president and chief executive officer of BioScience Laboratories, Inc., Bozeman, Montana. Previously, he was the manager of laboratory services at Skyland Scientific Services (1987–1991), Belgrade, Montana. A developer of statistical models for clinical trials of drugs and cosmetics, he is the author of more than 40 articles on clinical evaluations, software validations, solid dosage validations, and quantitative management science. In addition, he has also authored several books, including *Topical Antimicrobial Testing and Evaluation*, the *Handbook of Topical Antimicrobials*, *Applied Statistical Designs for the Researcher* (Marcel Dekker, Inc.), *Competitive Business*, *Caring Business: An Integral Business Perspective for the 21<sup>st</sup> Century* (Paraview Press), and *The Handbook of Regression Analysis* (Taylor & Francis Group). Currently, his books *Biostatistics and Microbiology: A Survival Manual* (Springer Group) and the *Handbook of Applied Biomedical Microbiology: A Biofilms Approach* (Taylor & Francis Group) are in progress. He is a member of the American Society for Microbiology, the American Society for Testing and Materials, the Association for Practitioners in Infection Control, the American Society for Quality Control, the American Psychological Association, the American College of Forensic Examiners, and the Association of Analytical Chemists.

Dr. Paulson received a BA (1972) in business administration and an MS (1981) in medical microbiology and biostatistics from the University of Montana, Missoula. He also received a PhD (1988) in psychology from Sierra University, Riverside, California; a PhD (1992) in psychoneuroimmunology from Saybrook Graduate School and Research Center, San Francisco, California; an MBA (2002) from the University of Montana, Missoula; and a PhD in art from Warnborough University, United Kingdom. He is currently working toward a PhD in both psychology and statistics and performs statistical services for the AOAC and the Department of Homeland Security.



---

# Series Introduction

The primary objectives of the *Biostatistics Book Series* are to provide useful reference books for researchers and scientists in academia, industry, and government, and also to offer textbooks for undergraduate and graduate courses in the area of biostatistics. This book series will provide comprehensive and unified presentations of statistical designs and analyses of important applications in biostatistics, such as those in biopharmaceuticals. A well-balanced summary is given of current and recently developed statistical methods, and interpretations for both statisticians and researchers or scientists with minimal statistical knowledge and engaged in the field of applied biostatistics. The series is committed to providing easy-to-understand, state-of-the-art references and textbooks. In each volume, statistical concepts and methodologies are illustrated through real-world examples.

Regression and modeling are commonly employed in pharmaceutical research and development. The purpose is not only to provide a valid and fair assessment of the pharmaceutical entity under investigation before regulatory approval, but also to assure that the pharmaceutical entity possesses good characteristics with the desired accuracy and reliability. In addition, it is to establish a predictive model for identifying patients who are most likely to respond to the test treatment under investigation. This volume is a condensation of various useful statistical methods that are commonly employed in pharmaceutical research and development. It covers important topics in pharmaceutical research and development such as multiple linear regression, model building or model selection, and analysis of covariance. This handbook provides useful approaches to pharmaceutical research and development. It would be beneficial to biostatisticians, medical researchers, and pharmaceutical scientists who are engaged in the areas of pharmaceutical research and development.

**Shein-Chung Chow**

---

# Table of Contents

<b>Chapter 1</b>	Basic Statistical Concepts . . . . .	1
	Meaning of Standard Deviation . . . . .	2
	Hypothesis Testing . . . . .	3
	Upper-Tail Test . . . . .	4
	Lower-Tail Test . . . . .	4
	Two-Tail Test . . . . .	4
	Confidence Intervals . . . . .	9
	Applied Research and Statistics . . . . .	9
	Experimental Validity . . . . .	10
	Empirical Research . . . . .	13
	Biases . . . . .	14
	Openness . . . . .	14
	Discernment . . . . .	15
	Understanding ( <i>Verstehen</i> ) . . . . .	15
	Experimental Process . . . . .	16
	Other Difficulties in Research . . . . .	20
	Experimental Error . . . . .	20
	Confusing Correlation with Causation . . . . .	20
	Complex Study Design . . . . .	21
	Basic Tools in Experimental Design . . . . .	22
	Statistical Method Selection: Overview . . . . .	23
<b>Chapter 2</b>	Simple Linear Regression . . . . .	25
	General Principles of Regression Analysis . . . . .	26
	Regression and Causality . . . . .	26
	Meaning of Regression Parameters . . . . .	27
	Data for Regression Analysis . . . . .	29
	Regression Parameter Calculation . . . . .	29
	Properties of the Least-Squares Estimation . . . . .	31
	Diagnostics . . . . .	34
	Estimation of the Error Term . . . . .	38
	Regression Inferences . . . . .	39
	Computer Output . . . . .	43
	Confidence Interval for $\beta_1$ . . . . .	43
	Inferences with $\beta_0$ . . . . .	44
	Power of the Tests for $\beta_0$ and $\beta_1$ . . . . .	47
	Estimating $\hat{y}$ via Confidence Intervals . . . . .	49

Confidence Interval of $\hat{y}$ . . . . .	50
Prediction of a Specific Observation. . . . .	53
Confidence Interval for the Entire Regression Model . . . . .	54
ANOVA and Regression . . . . .	57
Linear Model Evaluation of Fit of the Model . . . . .	62
Reduced Error Model . . . . .	65
Exploratory Data Analysis and Regression . . . . .	71
Pattern A. . . . .	72
Pattern B. . . . .	72
Pattern C. . . . .	72
Pattern D. . . . .	72
Data That Cannot Be Linearized by Reexpression . . . . .	73
Exploratory Data Analysis to Determine the Linearity of a Regression Line without Using the $F_c$ Test for Lack of Fit . . . . .	73
Correlation Coefficient . . . . .	76
Correlation Coefficient Hypothesis Testing. . . . .	79
Confidence Interval for the Correlation Coefficient. . . . .	81
Prediction of a Specific $x$ Value from a $y$ Value . . . . .	83
Predicting an Average $\hat{x}$ . . . . .	85
$D$ Value Computation . . . . .	86
Simultaneous Mean Inferences of $\beta_0$ and $\beta_1$ . . . . .	87
Simultaneous Multiple Mean Estimates of $y$ . . . . .	89
Special Problems in Simple Linear Regression . . . . .	91
Piecewise Regression. . . . .	91
Comparison of Multiple Simple Linear Regression Functions . . . . .	93
Evaluating Two Slopes ( $b_{1a}$ and $b_{1b}$ ) for Equivalence in Slope Values . . . . .	95
Evaluating the Two $y$ Intercepts ( $\beta_0$ ) for Equivalence . . . . .	101
Multiple Regression. . . . .	105
More Difficult to Understand . . . . .	105
Cost-Benefit Ratio Low. . . . .	105
Poorly Thought-Out Study . . . . .	106
Conclusion . . . . .	106
 <b>Chapter 3</b> Special Problems in Simple Linear Regression: Serial Correlation and Curve Fitting. . . . .	107
Autocorrelation or Serial Correlation . . . . .	107
Durbin-Watson Test for Serial Correlation . . . . .	109
Two-Tail Durbin-Watson Test Procedure . . . . .	119
Simplified Durbin-Watson Test . . . . .	119
Alternate Runs Test in Time Series. . . . .	120
Measures to Remedy Serial Correlation Problems . . . . .	123

Transformation Procedure (When Adding More Predictor $x_i$ Values Is Not an Option) . . . . .	124
Cochrane–Orcutt Procedure . . . . .	126
Lag 1 or First Difference Procedure . . . . .	133
Curve Fitting with Serial Correlation . . . . .	136
Remedy. . . . .	144
Residual Analysis $y_i - \hat{y}_i = e_i$ . . . . .	147
Standardized Residuals. . . . .	151
 <b>Chapter 4</b> Multiple Linear Regression. . . . .	153
Regression Coefficients . . . . .	153
Multiple Regression Assumptions. . . . .	154
General Regression Procedures . . . . .	155
Application . . . . .	156
Hypothesis Testing for Multiple Regression . . . . .	159
Overall Test. . . . .	159
Partial $F$ -Test. . . . .	161
Alternative to $SS_R$ . . . . .	165
The $t$ -Test for the Determination of the $\beta_i$ Contribution . . . . .	166
Multiple Partial $F$ -Tests . . . . .	168
Forward Selection: Predictor Variables Added into the Model . . . . .	173
Backward Elimination: Predictors Removed from the Model . . . . .	182
Discussion. . . . .	192
$\hat{Y}$ Estimate Point and Interval: Mean. . . . .	192
Confidence Interval Estimation of the $\beta_i$ s . . . . .	197
Predicting One or Several New Observations . . . . .	200
New Mean Vector Prediction . . . . .	202
Predicting $\ell$ New Observations. . . . .	202
Entire Regression Surface Confidence Region. . . . .	203
 <b>Chapter 5</b> Correlation Analysis in Multiple Regression . . . . .	205
Procedure for Testing Partial Correlation Coefficients. . . . .	209
$R^2$ Used to Determine How Many $x_i$ Variables to Include in the Model . . . . .	211
 <b>Chapter 6</b> Some Important Issues in Multiple Linear Regression . . . . .	213
Collinearity and Multiple Collinearity . . . . .	213
Measuring Multiple Collinearity. . . . .	214
Eigen ( $\lambda$ ) Analysis. . . . .	217
Condition Index. . . . .	219
Condition Number . . . . .	221
Variance Proportion. . . . .	221
Statistical Methods to Offset Serious Collinearity . . . . .	222

Rescaling the Data for Regression . . . . .	222
Ridge Regression . . . . .	222
Ridge Regression Procedure . . . . .	224
Conclusion . . . . .	240
<b>Chapter 7</b> Polynomial Regression . . . . .	241
Other Points to Consider . . . . .	242
Lack of Fit . . . . .	257
Splines (Piecewise Polynomial Regression) . . . . .	261
Spline Example Diagnostic . . . . .	266
Linear Splines . . . . .	269
<b>Chapter 8</b> Special Topics in Multiple Regression . . . . .	277
Interaction between the $x_i$ Predictor Variables . . . . .	277
Confounding . . . . .	280
Unequal Error Variances . . . . .	281
Residual Plots . . . . .	283
Modified Levene Test for Constant Variance . . . . .	285
Procedure . . . . .	286
Breusch–Pagan Test: Error Constancy . . . . .	293
For Multiple $x_i$ Variables . . . . .	296
Variance Stabilization Procedures . . . . .	299
Weighted Least Squares . . . . .	300
Estimation of the Weights . . . . .	302
Residuals and Outliers, Revisited . . . . .	307
Standardized Residuals . . . . .	309
Studentized Residuals . . . . .	309
Jackknife Residual . . . . .	310
To Determine Outliers . . . . .	311
Outlier Identification Strategy . . . . .	311
Leverage Value Diagnostics . . . . .	311
Cook's Distance . . . . .	313
Leverages and Cook's Distance . . . . .	323
Leverage and Influence . . . . .	325
Leverage: Hat Matrix ( $x$ Values) . . . . .	325
Influence: Cook's Distance . . . . .	333
Outlying Response Variable Observations, $y_i$ . . . . .	335
Studentized Deleted Residuals . . . . .	336
Influence: Beta Influence . . . . .	339
Summary . . . . .	340
<b>Chapter 9</b> Indicator (Dummy) Variable Regression . . . . .	341
Inguinal Site, IPA Product, Immediate . . . . .	345
Inguinal Site, IPA $\pm$ CHG Product, Immediate . . . . .	346

Inguinal Site, IPA Product, 24 h. . . . .	346
Inguinal Site, IPA $\pm$ CHG Product, 24 h. . . . .	346
Comparing Two Regression Functions . . . . .	353
Comparing the $y$ -Intercepts . . . . .	356
Test of $b_1$ s or Slopes: Parallelism. . . . .	361
Parallel Slope Test Using Indicator Variables . . . . .	364
Intercept Test Using an Indicator Variable Model . . . . .	367
Parallel Slope Test Using a Single Regression Model . . . . .	370
IPA Product . . . . .	372
IPA $\pm$ CHG Product. . . . .	372
Test for Coincidence Using a Single Regression Model. . . . .	373
Larger Variable Models . . . . .	375
More Complex Testing . . . . .	376
Global Test for Coincidence . . . . .	379
Global Parallelism . . . . .	383
Global Intercept Test. . . . .	385
Confidence Intervals for $\beta_i$ Values. . . . .	386
Piecewise Linear Regression . . . . .	387
More Complex Piecewise Regression Analysis . . . . .	391
Discontinuous Piecewise Regression . . . . .	401
<b>Chapter 10</b> Model Building and Model Selection . . . . .	409
Predictor Variables . . . . .	409
Measurement Collection . . . . .	410
Selection of the $x_i$ Predictor Variables . . . . .	410
Adequacy of the Model Fit . . . . .	411
Stepwise Regression . . . . .	414
Forward Selection . . . . .	416
Backward Elimination . . . . .	417
Best Subset Procedures . . . . .	419
$R_k^2$ and $SS_{E_k}$ . . . . .	420
Adj $R_k^2$ and $MS_{E_k}$ . . . . .	420
Mallow's $C_k$ Criteria . . . . .	421
Other Points . . . . .	421
<b>Chapter 11</b> Analysis of Covariance. . . . .	423
Single-Factor Covariance Model . . . . .	424
Some Further Considerations . . . . .	426
Requirements of ANCOVA . . . . .	428
ANCOVA Routine . . . . .	429
Regression Routine Example . . . . .	434
Treatment Effects . . . . .	437
Single Interval Estimate. . . . .	440

Scheffe Procedure—Multiple Contrasts . . . . .	440
Bonferroni Method . . . . .	442
Adjusted Average Response . . . . .	442
Conclusion . . . . .	443
 <b>Appendix I</b> . . . . .	 445
Tables A through O . . . . .	445
 <b>Appendix II</b> . . . . .	 481
Matrix Algebra Applied to Regression . . . . .	481
Matrix Operations . . . . .	483
Addition . . . . .	484
Subtraction . . . . .	484
Multiplication . . . . .	485
Inverse of Matrix . . . . .	488
 <b>References</b> . . . . .	 497
 <b>Index</b> . . . . .	 499



---

# 1 Basic Statistical Concepts

The use of statistics in clinical and pharmaceutical settings is extremely common. Because the data are generally collected under experimental conditions that result in measurements containing a certain amount of error,\* statistical analyses, though not perfect, are the most effective way of making sense of the data. The situation is often portrayed as

$$T = t + e.$$

Here, the true but unknown value of a measurement,  $T$ , consists of a sample measurement,  $t$ , and random error or variation,  $e$ . Statistical error is considered to be the random variability inherent in any system, not a mistake. For example, the incubation temperature of bacteria in an incubator might have a normal random fluctuation of  $\pm 1^\circ\text{C}$ , which is considered a statistical error. A timer might have an inherent fluctuation of  $\pm 0.01$  sec for each minute of actual time. Statistical analysis enables the researcher to account for this random error.

Fundamental to statistical measurement are two basic parameters: the population mean,  $\mu$ , and the population standard deviation,  $\sigma$ . The population parameters are generally unknown and are estimated by the sample mean,  $\bar{x}$ , and sample standard deviation,  $s$ . The sample mean is simply the central tendency of a sample set of data that is an unbiased estimate of the population mean,  $\mu$ . The central tendency is the sum of values in a set, or population, of numbers divided by the number of values in that set or population. For example, for the sample set of values 10, 13, 19, 9, 11, and 17, the sum is 79. When 79 is divided by the number of values in the set, 6, the average is  $79 \div 6 = 13.17$ . The statistical formula for average is

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

---

\*Statistical error is not a wrong measurement or a mistaken measurement. It is, instead, a representation of uncertainty concerning random fluctuations.

where the operator,  $\sum_{i=1}^n x_i$ , means to sum (add) the values beginning with  $i = 1$  and ending with the value  $n$ ; where  $n$  is the sample size.

The standard deviation for the population is written as  $\sigma$ , and for a sample as  $s$ .

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{N}},$$

where  $\sum_{i=1}^n (x_i - \mu)^2$  is the sum of the actual  $x_i$  values minus the population mean, the quantities squared; and  $N$  the total population size.

The sample standard deviation is given by

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

where  $\sum_{i=1}^n (x_i - \bar{x})^2$  is the sum of the actual sample values minus the sample mean, the quantities squared; and  $n - 1$  is the sample size minus 1, to account for the loss of one degree of freedom from estimating  $\mu$  by  $\bar{x}$ . Note that the standard deviation  $\sigma$  or  $s$  is the square root of the variance  $\sigma^2$  or  $s^2$ .

## MEANING OF STANDARD DEVIATION

The standard deviation provides a measure of variability about the mean or average value. If two data sets have the same mean, but their data range differ,\* so will their standard deviations. The larger the range, the larger the standard deviation.

For instance, using our previous example, the six data points—10, 13, 19, 9, 11, and 17—have a range of  $19 - 9 = 10$ . The standard deviation is calculated as

$$\sqrt{\frac{(10 - 13.1667)^2 + (13 - 13.1667)^2 + (19 - 13.1667)^2 + (9 - 13.1667)^2 + (11 - 13.1667)^2 + (17 - 13.1667)^2}{6 - 1}} = 4.0208.$$

Suppose the values were 1, 7, 11, 3, 28, and 29,

$$\bar{x} = \frac{1 + 7 + 11 + 3 + 28 + 29}{6} = 13.1667.$$

---

\*Range = maximum value – minimum value.