

STATISTICAL ANALYSIS IN MICROBIOLOGY

StatNotes



RICHARD A. ARMSTRONG • ANTHONY C. HILTON

STATISTICAL ANALYSIS IN MICROBIOLOGY: STATNOTES

Richard A. Armstrong and Anthony C. Hilton



 **WILEY-BLACKWELL**

A John Wiley & Sons, Inc. Publication



Copyright © 2011 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 is available

ISBN 978-0-470-55930-7

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

STATISTICAL ANALYSIS IN MICROBIOLOGY: STATNOTES



PREFACE

This book is aimed primarily at microbiologists who are undertaking research and who require a basic knowledge of statistics to analyze their experimental data. Computer software employing a wide range of data analysis methods is widely available to experimental scientists. The availability of this software, however, makes it essential that investigators understand the basic principles of statistics. Statistical analysis of data can be complex with many different methods of approach, each of which applies in a particular experimental circumstance. Hence, it is possible to apply an incorrect statistical method to data and to draw the wrong conclusions from an experiment. The purpose of this book, which has its origin in a series of articles published in the Society for Applied Microbiology journal *The Microbiologist*, is an attempt to present the basic logic of statistics as clearly as possible and, therefore, to dispel some of the myths that often surround the subject. The 28 *statnotes* deal with various topics that are likely to be encountered, including the nature of variables, the comparison of means of two or more groups, nonparametric statistics, analysis of variance, correlating variables, and more complex methods such as multiple linear regression and principal components analysis. In each case, the relevant statistical method is illustrated with examples drawn from experiments in microbiological research. The text incorporates a glossary of the most commonly used statistical terms, and there are two appendices designed to aid the investigator in the selection of the most appropriate test.

Richard Armstrong and Anthony Hilton

ACKNOWLEDGMENTS

We thank the Society for Applied Microbiology (SFAM) for permission to publish material that originally appeared in *The Microbiologist*. We would also like to acknowledge Dr. Lucy Harper, the editor of *The Microbiologist*, for help in commissioning this book, supporting its production, and for continuing encouragement.

We thank Tarja Karpanen and Tony Worthington (both of Aston University) for the use of data to illustrate Statnotes 15, 18, and 20 and Dr. Steve Smith (Aston University) for the data to illustrate Statnote 13.

We thank Graham Smith (Aston University) for drawing the figures used in Statnotes 25 and 28.

This book benefits from the teaching, research data, critical discussion, and especially the criticism of many colleagues: Dr. T. Bradwell (British Geological Survey), Dr. N. J. Cairns (Washington University, St Louis), Dr. M. Cole (Aston University), Dr. R. Cubbidge (Aston University), Dr. C. Dawkins (University of Oxford), Dr. M. C. M. Dunne (Aston University), Dr. F. Eperjesi (Aston University), Professor B. Gilmartin (Aston University), Dr. I. Healy (King's College London), Dr. E. Hilton (Aston University), Dr. P. D. Moore (King's College London), Dr. S. N. Smith (Aston University), and Dr. K. M. Wade (University of Oxford).

We dedicate the book to our families.

CONTENTS

Preface	xv
Acknowledgments	xvii
Note on Statistical Software	xix
1 ARE THE DATA NORMALLY DISTRIBUTED?	1
1.1 Introduction	1
1.2 Types of Data and Scores	2
1.3 Scenario	3
1.4 Data	3
1.5 Analysis: Fitting the Normal Distribution	3
1.5.1 How Is the Analysis Carried Out?	3
1.5.2 Interpretation	3
1.6 Conclusion	5
2 DESCRIBING THE NORMAL DISTRIBUTION	7
2.1 Introduction	7
2.2 Scenario	8
2.3 Data	8
2.4 Analysis: Describing the Normal Distribution	8
2.4.1 Mean and Standard Deviation	8
2.4.2 Coefficient of Variation	10
2.4.3 Equation of the Normal Distribution	10
2.5 Analysis: Is a Single Observation Typical of the Population?	11
2.5.1 How Is the Analysis Carried Out?	11
2.5.2 Interpretation	11
2.6 Analysis: Describing the Variation of Sample Means	12
2.7 Analysis: How to Fit Confidence Intervals to a Sample Mean	12
2.8 Conclusion	13
3 TESTING THE DIFFERENCE BETWEEN TWO GROUPS	15
3.1 Introduction	15
3.2 Scenario	16
3.3 Data	16

3.4	Analysis: The Unpaired t Test	16
3.4.1	How Is the Analysis Carried Out?	16
3.4.2	Interpretation	18
3.5	One-Tail and Two-Tail Tests	18
3.6	Analysis: The Paired t Test	18
3.7	Unpaired versus the Paired Design	19
3.8	Conclusion	19
4	WHAT IF THE DATA ARE NOT NORMALLY DISTRIBUTED?	21
4.1	Introduction	21
4.2	How to Recognize a Normal Distribution	21
4.3	Nonnormal Distributions	22
4.4	Data Transformation	23
4.5	Scenario	24
4.6	Data	24
4.7	Analysis: Mann–Whitney U test (for Unpaired Data)	24
4.7.1	How Is the Analysis Carried Out?	24
4.7.2	Interpretation	24
4.8	Analysis: Wilcoxon Signed-Rank Test (for Paired Data)	25
4.8.1	How Is the Analysis Carried Out?	25
4.8.2	Interpretation	26
4.9	Comparison of Parametric and Nonparametric Tests	26
4.10	Conclusion	26
5	CHI-SQUARE CONTINGENCY TABLES	29
5.1	Introduction	29
5.2	Scenario	30
5.3	Data	30
5.4	Analysis: 2×2 Contingency Table	31
5.4.1	How Is the Analysis Carried Out?	31
5.4.2	Interpretation	31
5.4.3	Yates' Correction	31
5.5	Analysis: Fisher's 2×2 Exact Test	31
5.6	Analysis: Rows \times Columns ($R \times C$) Contingency Tables	32
5.7	Conclusion	32
6	ONE-WAY ANALYSIS OF VARIANCE (ANOVA)	33
6.1	Introduction	33
6.2	Scenario	34
6.3	Data	34

6.4	Analysis	35
6.4.1	Logic of ANOVA	35
6.4.2	How Is the Analysis Carried Out?	35
6.4.3	Interpretation	36
6.5	Assumptions of ANOVA	37
6.6	Conclusion	37
7	POST HOC TESTS	39
7.1	Introduction	39
7.2	Scenario	40
7.3	Data	40
7.4	Analysis: Planned Comparisons between the Means	40
7.4.1	Orthogonal Contrasts	40
7.4.2	Interpretation	41
7.5	Analysis: Post Hoc Tests	42
7.5.1	Common Post Hoc Tests	42
7.5.2	Which Test to Use?	43
7.5.3	Interpretation	44
7.6	Conclusion	44
8	IS ONE SET OF DATA MORE VARIABLE THAN ANOTHER?	45
8.1	Introduction	45
8.2	Scenario	46
8.3	Data	46
8.4	Analysis of Two Groups: Variance Ratio Test	46
8.4.1	How Is the Analysis Carried Out?	46
8.4.2	Interpretation	47
8.5	Analysis of Three or More Groups: Bartlett's Test	47
8.5.1	How Is the Analysis Carried Out?	47
8.5.2	Interpretation	48
8.6	Analysis of Three or More Groups: Levene's Test	48
8.6.1	How Is the Analysis Carried Out?	48
8.6.2	Interpretation	48
8.7	Analysis of Three or More Groups: Brown–Forsythe Test	48
8.8	Conclusion	49
9	STATISTICAL POWER AND SAMPLE SIZE	51
9.1	Introduction	51
9.2	Calculate Sample Size for Comparing Two Independent Treatments	52
9.2.1	Scenario	52
9.2.2	How Is the Analysis Carried Out?	52

9.3	Implications of Sample Size Calculations	53
9.4	Calculation of the Power (P') of a Test	53
9.4.1	How Is the Analysis Carried Out?	53
9.4.2	Interpretation	54
9.5	Power and Sample Size in Other Designs	54
9.6	Power and Sample Size in ANOVA	54
9.7	More Complex Experimental Designs	55
9.8	Simple Rule of Thumb	56
9.9	Conclusion	56
10	ONE-WAY ANALYSIS OF VARIANCE (RANDOM EFFECTS MODEL): THE NESTED OR HIERARCHICAL DESIGN	57
10.1	Introduction	57
10.2	Scenario	58
10.3	Data	58
10.4	Analysis	58
10.4.1	How Is the Analysis Carried Out?	58
10.4.2	Random-Effects Model	58
10.4.3	Interpretation	60
10.5	Distinguish Random- and Fixed-Effect Factors	61
10.6	Conclusion	61
11	TWO-WAY ANALYSIS OF VARIANCE	63
11.1	Introduction	63
11.2	Scenario	64
11.3	Data	64
11.4	Analysis	64
11.4.1	How Is the Analysis Carried Out?	64
11.4.2	Statistical Model of Two-Way Design	65
11.4.3	Interpretation	65
11.5	Conclusion	66
12	TWO-FACTOR ANALYSIS OF VARIANCE	67
12.1	Introduction	67
12.2	Scenario	68
12.3	Data	68
12.4	Analysis	69
12.4.1	How Is the Analysis Carried Out?	69
12.4.2	Interpretation	70
12.5	Conclusion	70

13	SPLIT-PLOT ANALYSIS OF VARIANCE	71
13.1	Introduction	71
13.2	Scenario	72
13.3	Data	72
13.4	Analysis	73
13.4.1	How Is the Analysis Carried Out?	73
13.4.2	Interpretation	74
13.5	Conclusion	75
14	REPEATED-MEASURES ANALYSIS OF VARIANCE	77
14.1	Introduction	77
14.2	Scenario	78
14.3	Data	78
14.4	Analysis	78
14.4.1	How Is the Analysis Carried Out?	78
14.4.2	Interpretation	78
14.4.3	Repeated-Measures Design and Post Hoc Tests	80
14.5	Conclusion	80
15	CORRELATION OF TWO VARIABLES	81
15.1	Introduction	81
15.2	Naming Variables	82
15.3	Scenario	82
15.4	Data	83
15.5	Analysis	83
15.5.1	How Is the Analysis Carried Out?	83
15.5.2	Interpretation	83
15.6	Limitations of r	85
15.7	Conclusion	86
16	LIMITS OF AGREEMENT	87
16.1	Introduction	87
16.2	Scenario	88
16.3	Data	88
16.4	Analysis	88
16.4.1	Theory	88
16.4.2	How Is the Analysis Carried Out?	89
16.4.3	Interpretation	90
16.5	Conclusion	90

17	NONPARAMETRIC CORRELATION COEFFICIENTS	91
17.1	Introduction	91
17.2	Bivariate Normal Distribution	91
17.3	Scenario	92
17.4	Data	92
17.5	Analysis: Spearman's Rank Correlation (ρ , r_s)	93
17.5.1	How Is the Analysis Carried Out?	93
17.5.2	Interpretation	94
17.6	Analysis: Kendall's Rank Correlation (τ)	94
17.7	Analysis: Gamma (γ)	94
17.8	Conclusion	94
18	FITTING A REGRESSION LINE TO DATA	95
18.1	Introduction	95
18.2	Line of Best Fit	96
18.3	Scenario	97
18.4	Data	98
18.5	Analysis: Fitting the Line	98
18.6	Analysis: Goodness of Fit of the Line to the Points	98
18.6.1	Coefficient of Determination (r^2)	98
18.6.2	Analysis of Variance	99
18.6.3	t Test of Slope of Regression Line	100
18.7	Conclusion	100
19	USING A REGRESSION LINE FOR PREDICTION AND CALIBRATION	101
19.1	Introduction	101
19.2	Types of Prediction Problem	101
19.3	Scenario	102
19.4	Data	102
19.5	Analysis	102
19.5.1	Fitting the Regression Line	102
19.5.2	Confidence Intervals for a Regression Line	103
19.5.3	Interpretation	104
19.6	Conclusion	104
20	COMPARISON OF REGRESSION LINES	105
20.1	Introduction	105
20.2	Scenario	105
20.3	Data	106

20.4	Analysis	106
20.4.1	How Is the Analysis Carried Out?	106
20.4.2	Interpretation	107
20.5	Conclusion	108
21	NONLINEAR REGRESSION: FITTING AN EXPONENTIAL CURVE	109
21.1	Introduction	109
21.2	Common Types of Curve	110
21.3	Scenario	111
21.4	Data	111
21.5	Analysis	112
21.5.1	How Is the Analysis Carried Out?	112
21.5.2	Interpretation	112
21.6	Conclusion	112
22	NONLINEAR REGRESSION: FITTING A GENERAL POLYNOMIAL-TYPE CURVE	113
22.1	Introduction	113
22.2	Scenario A: Does a Curve Fit Better Than a Straight Line?	114
22.3	Data	114
22.4	Analysis	114
22.4.1	How Is the Analysis Carried Out?	114
22.4.2	Interpretation	115
22.5	Scenario B: Fitting a General Polynomial-Type Curve	115
22.6	Data	116
22.7	Analysis	117
22.7.1	How Is the Analysis Carried Out?	117
22.7.2	Interpretation	117
22.8	Conclusion	118
23	NONLINEAR REGRESSION: FITTING A LOGISTIC GROWTH CURVE	119
23.1	Introduction	119
23.2	Scenario	119
23.3	Data	120
23.4	Analysis: Nonlinear Estimation Methods	120
23.4.1	How Is the Analysis Carried Out?	120
23.4.2	Interpretation	121
23.6	Conclusion	122

24	NONPARAMETRIC ANALYSIS OF VARIANCE	123
24.1	Introduction	123
24.2	Scenario	123
24.3	Analysis: Kruskal–Wallis Test	124
24.3.1	Data	124
24.3.2	How Is the Analysis Carried Out?	124
24.3.3	Interpretation	125
24.4	Analysis: Friedmann’s Test	125
24.4.1	Data	125
24.4.2	How Is the Analysis Carried Out?	126
24.4.3	Interpretation	126
24.5	Conclusion	126
25	MULTIPLE LINEAR REGRESSION	127
25.1	Introduction	127
25.2	Scenario	128
25.3	Data	128
25.4	Analysis	129
25.4.1	Theory	129
25.4.2	Goodness-of-Fit Test of the Points to the Regression Plane	131
25.4.3	Multiple Correlation Coefficient (R)	131
25.4.4	Regression Coefficients	131
25.4.5	Interpretation	132
25.5	Conclusion	132
26	STEPWISE MULTIPLE REGRESSION	135
26.1	Introduction	135
26.2	Scenario	136
26.3	Data	136
26.4	Analysis by the Step-Up Method	136
26.4.1	How Is the Analysis Carried Out?	136
26.4.2	Interpretation	137
26.4.3	Step-Down Method	137
26.5	Conclusion	138
27	CLASSIFICATION AND DENDROGRAMS	139
27.1	Introduction	139
27.2	Scenario	140
27.3	Data	140

27.4	Analysis	140
27.4.1	Theory	140
27.4.2	How Is the Analysis Carried Out?	142
27.4.3	Interpretation	142
27.5	Conclusion	144
28	FACTOR ANALYSIS AND PRINCIPAL COMPONENTS ANALYSIS	145
28.1	Introduction	145
28.2	Scenario	146
28.3	Data	146
28.4	Analysis: Theory	147
28.5	Analysis: How Is the Analysis Carried Out?	148
28.5.1	Correlation Matrix	148
28.5.2	Statistical Tests on the Correlation Coefficient Matrix	148
28.5.3	Extraction of Principal Components	149
28.5.4	Stopping Rules	149
28.5.5	Factor Loadings	149
28.5.6	What Do the Extracted Factors Mean?	149
28.5.7	Interpretation	150
28.6	Conclusion	152
	References	153
	Appendix 1 Which Test to Use: Table	157
	Appendix 2 Which Test to Use: Key	159
	Appendix 3 Glossary of Statistical Terms and Their Abbreviations	163
	Appendix 4 Summary of Sample Size Procedures for Different Statistical Tests	167
	Index of Statistical Tests and Procedures	169

Statnote 1

ARE THE DATA NORMALLY DISTRIBUTED?

Why is knowledge of statistics necessary?

The role of statistics in an experimental investigation.

Types of data and scores.

Testing the degree of normality of the data: chi-square (χ^2) goodness-of-fit test or Kolmogorov–Smirnov (KS) test.

1.1 INTRODUCTION

Knowledge of statistical analysis is important for four main reasons. First, it is necessary to understand statistical data reported in increasing quantities in articles, reports, and research papers. Second, to appreciate the information provided by a statistical analysis of data, it is necessary to understand the logic that forms the basis of at least the most common tests. Third, it is necessary to be able to apply statistical tests correctly to a range of experimental problems. Fourth, advice will often be needed from a professional statistician with some experience of research in microbiology. Therefore, it will be necessary to communicate with a statistician, that is, to explain the problem clearly and to understand the advice given.

The scientific study of microbiology involves three aspects: (1) collecting the evidence, (2) processing the evidence, and (3) drawing a conclusion from the evidence. Statistical analysis is the most important stage of processing the evidence so that a sound

conclusion can be drawn from the data. Two types of question are often posed by scientific studies. The first type of question is a test of a hypothesis, for example, does adding a specific supplement to a culture medium increase the yield of a microorganism? The answer to this question will be either yes or no, and an experiment is often designed to elicit this answer. By convention, hypotheses are usually stated in the negative, or as *null hypotheses* (often given the symbol H_0), that is, we prefer to believe that there is no effect of the supplement until the experiment proves otherwise. The second type of question involves the estimation of a quantity. It may be established that a particular supplement increases the yield of a bacterium, and an experiment may be designed to quantify this effect. Statistical analysis of data enables H_0 to be tested and the errors involved in estimating quantities to be determined.

1.2 TYPES OF DATA AND SCORES

There are many types of numerical data or scores that can be collected in a scientific investigation, and the choice of statistical analysis will often depend on the form of the data. A major distinction between variables is to divide them into parametric and nonparametric variables. When a variable is described as *parametric*, it is assumed that the data come from a symmetrically shaped distribution known as the normal distribution, whereas *nonparametric* variables have a distribution whose shape may be markedly different from normal and are referred to as *distribution free*, that is, no assumptions are usually made about the shape of the distribution.

In this book, three types of data are commonly collected:

1. Attribute data in which the data are frequencies of events, for example, the frequencies of males and females in a hospital with a particular infectious disease. In addition, frequency data can be expressed as a proportion, for example, the proportions of patients who are resistant to various antibiotics in a hospital or community-based environment.
2. Ranked data in which a particular variable is ranked or scored on a fixed scale, for example, the abundance of fungi in different soil environments might be expressed on a scale from 0 (none) to 5 (abundant).
3. Measurements of variables that fulfill the requirements of the normal distribution. Many continuous biological variables are normally distributed and include many measurements in microbiology. Not all measurements, however, can be assumed to be normally distributed, and it may be difficult to be certain in an individual case. The decision may not be critical, however, since small departures from normality do not usually affect the validity of many of the common statistical tests (Snedecor & Cochran, 1980). In addition, many parametric tests can be carried out if the sample size is large enough. It is worth noting that tests designed to be used on normally distributed data are usually the most sensitive and efficient of those available.

Statnote 1 is concerned with the basic question of whether the data depart significantly from a normal distribution and, hence, whether parametric or nonparametric tests would be the most appropriate form of statistical analysis.