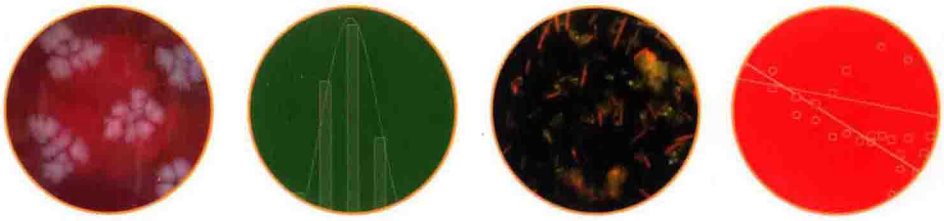
STATISTICAL ANALYSIS IN MICROBIOLOGY

StatNotes



RICHARD A. ARMSTRONG • ANTHONY C. HILTON

WILEY-BLACKWELL

society for applied sfam microbiology

STATISTICAL ANALYSIS IN MICROBIOLOGY: STATNOTES

Richard A. Armstrong and Anthony C. Hilton

WILEY-BLACKWELL

A John Wiley & Sons, Inc. Publication

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

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

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

Statistical Analysis in Microbiology: Statnotes, Edited by Richard A. Armstrong and Anthony C. Hilton

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.