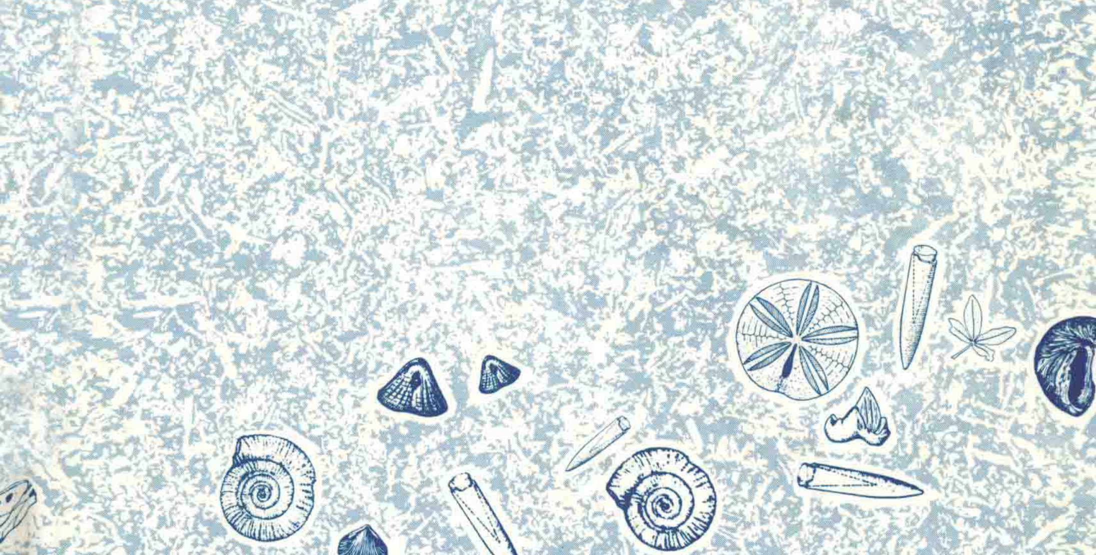
STATISTICS FOR GEOSCIENTISTS



D. MARSAL

PERGAMON PRESS

Statistics for Geoscientists

by

DIETER MARSAL

University of Stuttgart
Federal Republic of Germany

Translation editor

DANIEL F. MERRIAM

Wichita State University, USA

PERGAMON PRESS

OXFORD · NEW YORK · BEIJING · FRANKFURT
SÃO PAULO · SYDNEY · TOKYO · TORONTO

U.K.
Pergamon Press, Headington Hill Hall,
Oxford OX3 0BW, England

U.S.A.
Pergamon Press, Maxwell House, Fairview Park,
Elmsford, New York 10523, U.S.A.

PEOPLE'S REPUBLIC OF CHINA
Pergamon Press, Room 4037, Qianmen Hotel, Beijing,
People's Republic of China

FEDERAL REPUBLIC OF GERMANY
Pergamon Press, Hammerweg 6,
D-6242 Kronberg, Federal Republic of Germany

BRAZIL
Pergamon Editora, Rua Eça de Queiros, 346,
CEP 04011, Paraiso, São Paulo, Brazil

AUSTRALIA
Pergamon Press Australia, P.O. Box 544,
Potts Point, N.S.W. 2011, Australia

JAPAN
Pergamon Press, 8th Floor, Matsuoka Central Building,
1-7-1 Nishishinjuku, Shinjuku-ku, Tokyo 160, Japan

CANADA
Pergamon Press Canada, Suite No. 271,
253 College Street, Toronto, Ontario, Canada M5T 1R5

First English edition 1987

This book is a translation of *Statistische Methoden für Erdwissenschaftler* 1979, published by E. Schweizerbart'sche Verlagsbuchhaudlung (Nägele u. Obermiller) Stuttgart. ©1967 by E. Schweizerbart'sche Verlagsbuchhaudlung (Nägele u. Obermiller), Stuttgart.

Library of Congress Cataloging-in-Publication Data

Marsal, Dieter
Statistics for geoscientists.
Translation of: Statistische Methoden für Erdwissenschaftler.
Bibliography: p.
Includes indexes.
1. Earth sciences—Statistical methods.
I. Merriam, Daniel Francis. II. Title.
QE33.2.M3M3413 1987 550'.72 87-10582

British Library Cataloguing in Publication Data

Marsal, D.
Statistics for geoscientists.
1. Earth sciences—Statistical methods
I. Title II. II. Statistische Methoden für Erdwissenschaftler.
English
519.5'02455 QE33.2.M3

ISBN 0-08-026268-6 (Hardcover)
ISBN 0-08-026260-0 (Flexicover)

Reproduced, printed and bound in Great Britain by
Hazell Watson & Viney Limited,
Member of the BPCC Group,
Aylesbury, Bucks

Statistics for Geoscientists

Related Pergamon Titles

ANDERSON
The Structure of Western Europe

ANDERSON
The Structure of the British Isles, 2nd Edition

CONDIE
Plate Tectonics and Crustal Evolution, 2nd edition

EYLES
Glacial Geology: An Introduction for Engineers and Scientists

FERGUSSON
Inorganic Chemistry and the Earth

HANLEY & MERRIAM
Microcomputer Applications in Geology

To my dear old friend

Dr. Karl Dietrich Adam

Professor of Paleontology and Prehistory

Foreword to the English Edition

The use of statistics is becoming more evident in geology and related earth sciences. Witness the many texts and references on the subject as well as the research papers and special meetings such as the one at the Geological Society of America in Reno on "The Use and Abuse of Statistics." Because of the importance in understanding the background of such things as sampling, distributions, significance, correlation, classification, etc., geologists need to be introduced to the subject early in their careers. This little book is intended for just that purpose.

Simple, straightforward, and concise, this book on *Statistics for Geoscientists* by Dieter Marsal is ideal as an introduction to the subject. The first edition was published in 1967 in German and it was revised in 1979. The elegance of the presentation, however, was lost on the English-speaking geological public and the book's influence was limited. Although there are many examples, the book is not a step-by-step cookbook on how to use statistics. Subject matter covers everything from sampling, distributions, and central tendencies through significance tests, time series, Markov chains, correlation and regression, and Fourier analysis to discriminant analysis, splines, and the analysis of variance. A brief but thorough presentation, it is sprinkled liberally with examples of application, albeit European ones.

For some readers, the compact style (23 chapters) may be difficult to adjust to, and for others the lack of familiar examples may be distracting, but for most, the directness, completeness, and brevity will be welcome. The book can be used as a first approach to a difficult subject, for clarification of particular subjects, or for new ideas in applications. Some may ask why another statistics book, especially one first published so long ago? The answer is easy—people learn from different approaches and this approach is different from others that are available. Although the book was published first 20 years ago, the basic tenets and mathematics have not changed, only the adaptations to the computer which have improved during this time, especially in the presentation of output. So, it is believed the book has a place and will be appreciated, even though there are other books on the subject.

I first encountered Marsal's book soon after its original publication, and although the German was not difficult, I thought how much better it would be if translated into English. We worked toward that goal—an English edition—for several years. Finally, Dieter did the rewrite himself with the stipulation I edit the English version; I gladly agreed. Those familiar with the German edition will note the English edition is not a word-for-word translation but subject-by-subject translation. During the project it was realized that some changes were needed, and those were affected. I would like to thank my German-speaking colleague, Dr. Peter G. Sutterlin of Wichita State University, for reading the manuscript and making many helpful suggestions. His thoughtful contribution to the project was most appreciated.

So, here is yet another statistics book for earth scientists—from a European viewpoint. May geologists learn to appreciate the importance of statistics and their application in the earth sciences!

Department of Geology,
Wichita State University,
Wichita, Kansas 67208, USA

D. F. MERRIAM

Contents

CHAPTER 1

Introduction, Scope and Purpose of Applied Statistics

Applied statistics are concerned mainly with describing and analyzing the *variability* of various data gathered from a larger collection termed a *universe* or *population.* For instance, the universe might be defined as a sand dune composed of grains of various shapes and sizes that we want to study by scrutinizing a finite number of grains of a gathered *sample.*

Invariably, every statistical investigation commences with the collection of raw data and its representation in a clearly arranged mode. The raw data may have been obtained by collecting observations or measurements; they may refer to nonnumerical qualities such as the colors and morphological properties of a mineral, or they may comprise numerical quantities such as lengths and weights. All data must be stored in a *databank* or catalogued as a *list.* This *inventory* is a document representing a basic set of data from which different types of inferences may be derived by analysts.

In most situations—especially in elementary statistics—the first step of analysis consists in determining the *frequency distribution* of the collected data. It shows the frequency of occurrence of each listed value or quality, or it shows the frequency of occurrence of all listed numbers which are in a certain *interval*, also termed *a class* or *fraction.*

Many empirical frequency distributions can be described more or less accurately by functions with well-known mathematical properties. Typical examples are the *normal distribution* which seems to be the most usual mathematical distribution function, the *binomial distribution* which describes the expected frequencies of the occurrence of two alternative states, and the *distribution of Poisson* which can be shown to describe quantitatively the occurrence of rare events.

Not much can be learned directly from the initial raw-data list; however, *condensed* and systematically arranged as a frequency distribution, the variability of the considered property or properties can be seen at a glance, especially if plotted as a *histogram.* Yet it may be desirable to condense the data further. This is done by characterizing the frequency distribution by a few numbers termed *parameters* which describe properties of the distribution as a whole. Most important are average values such as the *mean* and *median*, and variability indicators such as the *variance* and *standard deviation.* These and other concepts related to them will be discussed at the appropriate places. Some of these parameters characterizing a sampling distribution especially are appropriate to the geological sciences.

To obtain additional results, the elementary methods of routine collection, tabulation, description, and condensation of data *per se*, are to be complemented by more advanced techniques.

In most situations, the number of individuals in a universe is large, and most parts of it are inaccessible. Thus, the investigator must be content with a sample whose size is small compared to the total number of individuals of the whole population. Unfortunately, the frequency distribution usually differs for each sample, and none is exactly alike to the frequency distribution of the sample universe. Hence, we cannot readily assume that the results derived from a sample are *representative* for its universe. This problem gives rise to the following questions:

1. How many samples must be drawn from a population to obtain representative results?
2. How large is the size of a representative sample? Strictly speaking: How many individuals constitute a representative sample?
3. If the objects of the universe are distributed spatially such as minerals in a rock or stars in a stellar cloud, then at what locations must the samples be collected?

In mathematical statistics these questions are dealt with on the assumption that the objects of the universe are *randomly distributed,* that is distributed according to the laws of chance. Then the problems can be solved by applying the principles of probability theory to statistics. This part of mathematical statistics is termed *sampling theory.* It turns out that the confidence we can have in the results inferred from a sample depends on the size of the sample and the universe's frequency distribution. Unfortunately, however, probability theory never gives answers that are certain; the answers are invariably of the type "With a specified certainty of less than 100%, the mean or the variance, etc., of the universe is contained in some interval, the so-called *confidence interval."* For instance, the answer may be: "With a certainty of 99%, the mean grain size of an investigated dune is between 0.11 and 0.19 centimetres." Thus the expression "with a certainty of 99%" indicates that the *estimate* of the confidence interval is correct in 99 situations out of 100 drawn samples*. In the latter situation, the mean grain size of the universe is either larger than 0.19 cm or smaller than 0.11 cm.

*The colloquial expression "with a certainty of 99%" corresponds to the technical jargon "for the *confidence coefficient* 0.99."

The probabilistic approach to sampling poses two fundamental questions:

1. What is the appropriate confidence coefficient? 0.9, or 0.95, or 0.99, or 0.999, etc.?
2. Is the basic assumption of sampling theory—randomness of the distribution—justified?

When attempting to forecast the results of a democratic election by polling a sample of people, the appropriate confidence coefficient may be determined by comparing forecasts with corresponding outcomes. This approach, however, is seldom workable in the geosciences. Thus, there is no unique answer to the first question—the investigator has to rely on experience, and the answer may be different for different geological universes. Hence, statistics as applied to the earth sciences is—although mainly a rigorous discipline—partly an art.

The answer to the second question is Yes and No. Occasionally, a geological body is rather homogeneous, and randomness is a fair assumption. In most situations, however, geological bodies are rather inhomogeneous as a whole. The mineral distributions of sediments, magmatic and metamorphic rocks furnish an inexhaustable multitude of examples and realized possibilities ranging from ideal randomness to almost regular pattern arrangements.

Another important aspect of sampling theory is the *testing of statistical hypotheses.* In the simplest situation, two different distributions are compared with each other to test whether they belong to the same universe. This is done usually by applying the *Chi-square method.* In many instances, however, it is sufficient to compare the means and variances by applying the *Student-t test* and *Snedecor's F-test* respectively or some other technique*. Qualitative frequency distributions, where the frequency is characterized by designations as "abundant," "rare," etc., are compared with each other by employing *ranking methods.* The idea behind the ranking is to allocate a placement to each frequency designation in much the same manner as in school the "1" or "A" might be assigned to note "excellent", "2" or "B" good, etc.

These methods may fail and thus the question whether two samples belong to the same population remains undecided. Then the technique of *discriminant analysis* might be applied successfully. It is based on the idea that a proper combination of several properties may be more informative than a comparison of a single property.

The advanced techniques of applied statistics permit the simultaneous

*The Student-*t* test is based on a mathematical distribution function that was published under the pseudonym "Student." This piece of work can be considered the starting point of modern statistics.

comparison of an arbitrary number of samples. Thereby the universe is tested for homogeneity, and the population's variance is split and allocated to different determining factors. This permits in the determination of the underlying factors which mainly are responsible for the variablity of the universe. Typical keywords of this powerful technique are "*analysis of variance,*" "*factorial analysis,*" and "*Latin squares.*" These and related topics are covered amply by an extensive literature and their understanding requires mathematical maturity. Hence, in an introductory text only the simpler methods can be discussed. The modern rather sophisticated approaches to factorial analysis are beyond the scope of this book.

It may be that several properties of a universe are interrelated as possibly, for instance, the largest and the smallest diameter of a mineral or the size of some fossil and geological time. Usually the interrelationship is obscured by unknown influences. As a consequence, the relationship among the properties under consideration cannot be described exactly but only grossly by simple graphs such as straight lines, tilted planes, parabolas, exponential curves etc. This relationship is known as "*correlation.*" The concept has two main aspects, a statistical and a numerical one. For instance, when plotting the diameters of *Cosmoceras* individuals versus geological time, we may obtain a set of points scattered about a straight line. Then the first task is to determine a "*regression line*" such that it represents a *best fit* of the data. Now the solution to the question "What is the best fit and how can the best fitting curve be obtained?" is determined by applying elementary numerical analysis methods. In most situations, either the *Gaussian least-square method* or the modern approach of employing *spline functions* is used.

The statistical aspect of this type of correlation is concerned with the interpretation of interrelated data. How good is the fit? Are the regression curves of a sample representative for the universe? How are the regression curves related to the frequency distribution which encompasses all correlated properties of the population?

Obviously, these questions are rather technical. The reader might be more interested in detecting causal relationships among correlated data—a topic popular in the medical sciences. Primarily, a correlation constitutes a formal measure which might suggest, but does not prove necessarily a cause-and-effect relationship. Thus the statements "The Dutch coast is sinking slowly every year" and "The population of the Netherlands is steadily increasing" certainly may constitute a correlation, but is not necessarily a causal relationship. The statistician always must be on guard when attempting to draw conclusions which are beyond the scope of statistical inference. Physico-biological interpretations of statistical results must include supporting data or complementary reasoning.

Variability of various properties is one of the outstanding features of most geological bodies. Hence statistics as the science of studying the variability of observed data should be a major tool of the geoscientist. However, statistical inference is not a method in itself and although perhaps necessary, usually is not sufficient from which to draw causal conclusions.

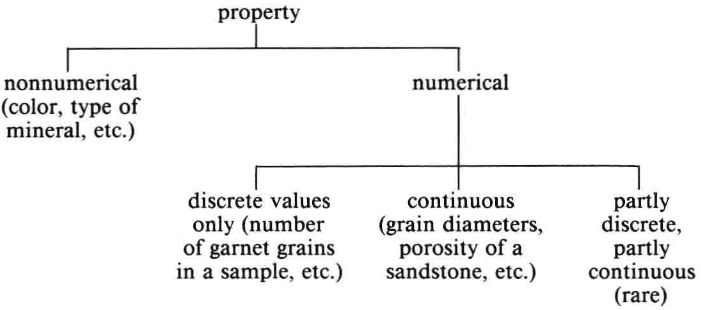
In a *tour de force* I have mentioned most topics dealt with in this book with the exception of such topics as *analysis of time series* and *banking cycles.* Advanced methods and special topics such as the analysis of random vibrations as applied to geophysics are not included. The inclusion of these would have made the book bulky and too lengthy and therefore unattractive to the beginner.

Further there is no introduction to probability theory. My reason is simple. As soon as this mathematical discipline is treated beyond the theory of tossing coins, it becomes intricate and demanding. To understand, for instance, probabilistic sampling theory requires a mathematical maturity which cannot be expected from the practicing geologist, petrographer, or paleontologist. Thus, the reader will become better acquainted with the application of statistical techniques rather than with the derivation of the equations developed from basic principles.

CHAPTER 2

Classification and Tabulation of Frequency Distributions

For a collection of data, the number of items in a given class C is termed C's *frequency*. In the geosciences, C may be a nonnumerical property such as a color, a specific type of object such as a mineral, a single numerical value, or an interval such as the set of all numbers between zero and one. The frequency may refer to a *universe* as a whole or to a *sample*. The hierarchical order of most statistical properties is covered by the following scheme:


property
nonnumerical (color, type of mineral, etc.)
numerical
discrete values only (number of garnet grains in a sample, etc.)
continuous (grain diameters, porosity of a sandstone, etc.)
partly discrete, partly continuous (rare)

When a collection of data is separated into several classes, the number of items in a given class is the *absolute frequency*, and the absolute frequency divided by the total number of items is the *relative frequency*. The sum of all relative frequencies is unity (100%). The set of all frequencies with their corresponding classes is the *absolute* (*relative*) *frequency distribution*. Occasionally, the exact frequencies are unknown but can be designated qualitatively by expressions suc̀h as "rare" and "abundant."

In most situations, the investigator is interested in relative frequencies rather than in absolute ones. However, the total numbers of items must be known to estimate whether the sample distribution is representative of the universe. Thus, whenever possible, the total number of items involved should be stated.

Example 2.1 Distributions with semiquantitative or nonnumerical frequencies

Remains of some land mammals which are present in the Devil's Hole (Teufelslucken) near Eggenburg in Lower Austria (Adam, 1966):

Remains	Cave-hyena	Mammoth
Bones	1096	7
Teeth	955	31

This is a distribution with four classes (cave-hyena bones, etc.) and a total of 2089 items. By a rather complex analysis of the recovered material, the minimum number of animals that were dragged into or perished in the cave could be determined:

cave-hyena: at least 67 mammoth: at least 13.

This is a frequency distribution for two classes (a so-called *dichotomous distribution*). In the following example, frequencies of occurrence of mammoth remains are indicated by the signs + (a few items) and − (none):

Czechoslovakia, Magdalénien

Location	Remains of mammoth
Adlerova jeskyně	−
Balcarova skála	+
Hadi jeskyně	−
....	.....
Žitného jeskyně	−

This semiquantitative distribution has an interesting interpretation: At the end of the Ice Ages the mammoth had ceased to be a factor in hunting for man. The remains of mammoth at the Devil's Hole near Eggenburg are evidence of hunting at an earlier time.

Example 2.2 Distributions with nonnumerical classes

The left-hand distribution refers to the relative abundancy of minerals in a thin section that were estimated by comparison with a chart for visual percentage evaluation (cf. Müller, 1964, p. 143; Niggli, 1948, p. 149).