Developments in Atmospheric Science 4A

O. ESSENWANGER

Applied Statistics in Atmospheric Science

Part A. Frequencies and Curve Fitting

Developments in Atmospheric Science, 4A

Applied Statistics in Atmospheric Science

Part A. Frequencies and Curve Fitting

by

O. ESSENWANGER

U.S. Army Missile Command, Redstone Arsenal and University of Alabama, Huntsville, Ala., U.S.A.

ELSEVIER SCIENTIFIC PUBLISHING COMPANY
Amsterdam — Oxford — New York 1976

ELSEVIER SCIENTIFIC PUBLISHING COMPANY
335 Jan van Galenstraat
P.O. Box 211, Amsterdam, The Netherlands

AMERICAN ELSEVIER PUBLISHING COMPANY, INC.
52 Vanderbilt Avenue
New York, New York 10017

With 17 illustrations and 37 tables

ISBN: 0-444-41327-8

Printed in The Netherlands

Developments in Atmospheric Science, 4A

Applied Statistics in Atmospheric Science
Part A, Frequencies and Curve Fitting

Further titles in this series

1. F. VERNIANI (Editor)
Structure and Dynamics of the Upper Atmosphere

2. E.E. GOSSARD and W.H. HOOKE
Waves in the Atmosphere

3. L.P. SMITH
Methods in Agricultural Meteorology

5. G.W. PALTRIDGE and C.M.R. PLATT
Radiative Processes in Meteorology and Climatology

6. P. SCHWERDTFEGER
Physical Principles of Micrometeorological Measurements

TO MY WIFE KATE
and my daughters
MARIANNE AND ANGELIKA

FOREWORD

It was originally planned to supplement a text on elements of statistical analysis which will be included in the World Survey of Climatology (WSoC), Volume 3 (see Essenwanger, 1974a), by preparation of a manuscript on advanced topics. These two texts would have been published together as a book on statistical analysis for atmospheric science. The enormous material requiring treatment in statistical analysis of atmospheric data and the delay in the publication of Volume 3 (WSoC) made it advisable, however, to separate the elementary and advanced topics. Since it is intended to publish the chapter on statistical analysis from the WSoC as a separate text, the basic knowledge of these elementary topics must be assumed.

This text on advanced topics in statistical analysis builds upon the knowledge which is gained from the elementary text. Atmospheric scientists and statisticians who possess a basic knowledge in statistical analysis should be able to follow, with little difficulty, the course of this book on advanced topics without having the elementary text. To aid their reading, the nomenclature and general descriptions are repeated in a short introduction.

The field of statistical analysis of atmospheric data is very comprehensive and requires treatment of a wide variety of topics. The completion of a manuscript comprising all the subjects of interest would have further delayed the publication. The advanced topics have been split into two parts. While this first part deals with frequency distribution and what can be called "curve fitting procedures", the second part (in preparation) will treat smoothing and filtering, analysis of variance and advanced test methods. A section on special atmospheric topics will round off that part.

It is not the author's intention to add to the number of texts on statistical theory, and the reader may discover that theoretical points will most of the time find short treatise. The major emphasis is placed on problems arising from the application of statistical procedures in practical work. Thus, many examples illustrate difficulties encountered in data analysis rather than the cases of smooth compliance with theory. The applications of several tools to the same set of data should aid in interpretation and evaluation of results from statistical analysis. Statements on cost effectiveness by individual statistical procedures have been added, if appropriate.

It was considered essential to include in this text brief sections on numerical methods for solutions such as for calculation of eigenvalues and eigenvectors. These procedures belong to the topics on empirical polynomial representation and factor analysis. More sophisticated methods may exist than those presented in this text, and the theory had sometimes to be cut short.

Nevertheless, the reader may find it convenient that techniques for calculations of frequency distributions or eigenvalues, etc. are included. This should aid considerably in attaining a quick answer and enabling the fast evaluation of a particular technique for suitability in the analysis of atmospheric data.

Finally the author wishes to express his gratitude to all the persons who have given their support to the establishment of this text. It is not possible to list all of them, but some must be singled out. First of all, I want to thank Prof. Dr. E. Reiter (Colorado State University) for providing the opportunity to teach in a lecture series on statistical analysis at CSU. These lecture notes became the basic material for some of the topics in this text. His consistent encouragement to publish the notes and his keen interest during the writing of this text convinced me that a book on applied statistics in atmospheric science is necessary. My appreciation is extended to Prof. Dr. E. Wahl (University of Wisconsin) for his critical comments on some of the topics. Prof. Dr. H. Flohn (University of Bonn) and Prof. Dr. Landsberg (University of Maryland) deserve my thanks for inviting me to write the chapter on elements of statistical analysis for the WSoC which is expanded with this text.

My thanks go further to my colleagues Dr. J. Stettler, Dr. H. Meyer and Mrs. H. Boyd (Physical Sciences Directorate, Army Missile Command) who took the cumbersome job of editing and reviewing the text. Last but not least Mrs. C. Brooks deserves much credit for her patience in typing the manuscript. I also wish to thank the Army Missile Command for permission to include some of my unpublished work in atmospheric data analysis.

Huntsville, Al., November 1974 O.M. ESSENWANGER

NOMENCLATURE

1. $1/2B \equiv 1/(2B)$, both versions utilized, but always:
   $0.5B \equiv (1/2)B$
   $1/2Bx \equiv 1/(2Bx)$ in contrast to $(1/2B)x$
2. tanh, sinh, cosh are hyperbolic functions.
3. $\tan^{-1}$, $\cos^{-1}$, etc. are 1/tan, 1/cos, etc.
   arctan or arcsin is spelled out.
4. The square root sign without bar is valid to the end of the line, parenthesis, or equal sign.

   $(1 + \sqrt{1 + x}) = 1 + \sqrt{1 + x}$, etc.

   e.g.

   $$\sqrt{\beta_1/\{(c+2)\beta_1 + (c+1)\}} \equiv [\beta_1/\{(c+2)\beta_1 + (c+1)\}]^{1/2}$$

   $$\sqrt{2/\pi} \equiv (2/\pi)^{1/2} \quad \text{but} \quad \sqrt{2}/\pi \equiv (1/\pi)\sqrt{2}$$

5. matrix **M** (capitals)
6. vector x (small letters)
7. $\sim$ approximately.

CONTENTS

CHAPTER 1

INTRODUCTION

This text requires the knowledge of elementary topics treated by the author in a separate volume (1974a), but the reader who is acquainted with the basic principles in statistical analysis should be able to follow the advanced topics in this book independently of the cited reference. Comprehension of the basic frequencies such as the binomial, Poisson, gamma and Gaussian distribution is assumed. A further prerequisite is familiarity with regression analysis, power spectrum, and some basic test procedures such as the t, F, χ^2 and Kolmogorov-Smirnov tests. These basic topics can be found in the statistical literature (e.g. textbooks quoted in the list of references) although the author (1974a) has illustrated their special application to atmospheric science.

In the Introduction of this first of two books on applied statistics the reader will become familiar with the nomenclature. Chapter 2 treats frequency distributions and Chapter 3 can be called "curve fitting procedures" applied to atmospheric data. A final chapter comprises a brief description of required mathematical techniques. The second volume (Essenwanger, 1974b) will deal with the problems of filtering, analysis of variance, advanced tests and special meteorological topics.

It should be stressed that the author did not intend to prepare a theoretical treatise. Excellent textbooks which achieve this goal are in existence. This text is intended to serve the practitioner and to delineate the practical aspects of statistical analysis with examples from the area of atmospheric science.

1.1 *EXPECTANCY, PROBABILITY DENSITY, CUMULATIVE DISTRIBUTION AND CLASSES*

It is not the intention of this section to repeat the introduction to elements of statistical analysis (Essenwanger, 1974a). Some of the general definitions and formulae should be repeated, however, to enable the reader to utilize this book as a text of its own.

It was introduced that in general N will be the total number of data points in a sample, the n a partial sample, e.g. an observed frequency in a certain class with determined boundaries. Then:

$$n/N = f \to p = P(A) \text{ for } N \to \infty \qquad [1.1]$$

The p denotes the stabilized relative frequency of an event A. The $P(A)$ and p are expected probabilities for N being very large. We may also call $P(A)$ the

population probability. Then:

$$N_e = N \cdot p \qquad [1.2]$$

is the expected number of observations and n_i represents the empirical number of observations for the ith class interval. Relationships of probability will not be repeated here.

$$F(x) = P(A) = P(X \leqslant x) \qquad [1.3]$$

is called the cumulative distribution function (c.d.f.). It represents the probability of an event A with a variate value X being less than or equal to the threshold x.

If we can write:

$$P(X \in A) = \int_A f(x)\,dx \qquad [1.4]$$

then $f(x)$ is called the probability density function (p.d.f.) or, expressed differently:

$$F(a) = P(X \leqslant a) = \int_{-\infty}^{a} f(x)\,dx \qquad [1.5]$$

The class width is defined as:

$$w = x_u - x_l \qquad [1.6]$$

where x_u and x_l are the upper and lower boundaries, respectively. The central class value x_c is then:

$$x_l < x_c < x_u \qquad [1.6a]$$

In many cases:

$$x_c = (x_u - x_l)/2 \qquad [1.6b]$$

but this postulation is not fulfilled if the slopes to both sides of the central class value are not linearly related, e.g. for a logarithmic progression of the variate x. Then x_c must be redefined.

1.2 *MOMENTS AND CUMULANTS*

Primitive characteristics such as the maximum, minimum, range, or other characteristics based on c.d.f. (e.g. quantiles), or p.d.f. need not be redefined here. Mathematical characteristics can be defined as:

$$E(X^\nu) = \mu_\nu = \int_{-\infty}^{\infty} X^\nu\,dx \qquad [1.7a]$$

which are called the moments of the distribution, with the notation $E(\ )$ standing for the expectancy. The ν denotes the order of the moment. For discrete variables:

$$E(X^\nu) = \mu_\nu = \Sigma(X^\nu)/N \qquad [1.7b]$$

where it is immaterial whether the left or right expression of [1.7c] below is employed:

$$\Sigma[(X^\nu)/N] \equiv (\Sigma X^\nu)/N \qquad [1.7c]$$

If X can assume large numbers in practical analysis, the left side of the expression may be more accurate due to computer truncation of large numbers for ΣX^ν.

Note. If no boundaries for the summation are specified, it means automatically that the summation should be carried out over all available data.

Equations [1.7] can be modified for inclusion of the p.d.f., namely:

$$E(X^\nu) = \mu_\nu = \int_{-\infty}^{\infty} f(x)x^\nu \, dx \qquad [1.8a]$$

and

$$E(X^\nu) = \mu_\nu = \sum_1^n f(x)x^\nu \Delta x \qquad [1.8b]$$

Usually the Δx is assumed to be unity, and the n denotes the number of classes. As the reader will notice, the symbol μ stands for the general moments reference zero of the variate.

The central moments are based on the mean value.*

$$E[(X - \bar{X})^\nu] = \nu_\nu = \sum_1^n f(x_j)(x_j - \bar{x})^\nu = \sum_1^N (x_i - \bar{x})^\nu / N \qquad [1.9]$$

The following expressions are important:

$$\mu_0 = N \qquad [1.10a]$$

$$\mu_1 = \bar{x} = \bar{X} = \Sigma X_i/N = \sum_1^n f(x_j)x_j \text{ (mean)} \qquad [1.10b]$$

$$\mu_2 = \sigma^2 + \bar{x}^2 = \nu_2 + \bar{x}^2 \qquad [1.10c]$$

Furthermore

$$\nu_1 = 0 \qquad [1.11a]$$

$$\nu_2 = \sigma^2 \text{ (variance)} \qquad [1.11b]$$

$$\nu_3 = \mu_3 - 3\bar{X}\nu_2 - \bar{X}^3 \qquad [1.11c]$$

$$\nu_3 = \mu_3 - 3\bar{X}\mu_2 + 2\bar{X}^3 \qquad [1.11d]$$

$$\nu_4 = \mu_4 - 4\bar{X}\nu_3 - 6\bar{X}^2\nu_2 - \bar{X}^4 \qquad [1.11e]$$

$$\nu_4 = \mu_4 - 4X\mu_3 + 6\bar{X}^2\mu_2 - 3\bar{X}^4 \qquad [1.11f]$$

---

* The central moments are given for the summation only. The integral form can readily be deduced from a comparison between [1.8a] and [1.8b].

Some characteristics for frequency distributions are derived from the moments. Skewness (γ) and kurtosis (k_u) are defined as:

$$\gamma_1 = \nu_3/\sigma^3 \qquad [1.12a]$$

$$k_u = \gamma_2 = (\nu_4/\sigma^4) - 3 \qquad [1.12b]$$

Karl Pearson (1895) has defined parameters β, namely:

$$\beta_1 = (\gamma_1)^2 = \nu_3^2/(\sigma^2)^3 \qquad [1.13a]$$

$$\beta_2 = (\gamma_2 + 3) = \nu_4/(\sigma^4) \qquad [1.13b]$$

If moments are calculated from grouped data Sheppard's corrections are usually applied:

$$\nu_2' = \nu_2 - w^2/12 \qquad [1.14a]$$

and:

$$\nu_4' = \nu_4 - \nu_2 w^2/2 + (7/240)w^4 \qquad [1.14b]$$

The odd moments do not need a correction for grouping. (Note: $7/240 = 0.02917$.)

A mixed moment is defined as:

$$\nu_{xy} = \sum_1^N (X_i - \bar{X})(Y_i - \bar{Y})/N \qquad [1.15a]$$

$$= \sum_1^n f(xy)_j \cdot (x_j - \bar{x})(y_j - \bar{y})/N \qquad [1.15b]$$

The term ν_{xy} or $N\nu_{xy}$ is also called covariance or cross-product depending on the definition.

The integrals of [1.7a] or [1.8a] are very often difficult to solve. When all moments are finite, the integral exists and a "moment generating function $g(t)$" can be defined:

$$g(t) = E(e^{xt}) = \int_{-\infty}^{\infty} e^{xt} f(x)\,dx \qquad [1.16a]$$

By differentiation:

$$\frac{d^\nu g(t)}{dt^\nu} = \int_{-\infty}^{\infty} x^\nu e^{xt} f(x)\,dx \qquad [1.16b]$$

and for $t = 0$:

$$\frac{d^\nu g(0)}{dt^\nu} = E(x^\nu) = \mu_\nu \qquad [1.16c]$$

This procedure indicates that the νth moment follows from the νth derivative of $g(t)$ and substitution of $t = 0$.

Sometimes the "cumulant" generating function $k(t)$ is simpler:

$$k(t) = \ln g(t) = k_1 t + k_2 t^2/2 + \ldots k_\nu t^\nu/\nu! \quad [1.17]$$

It should be noted that:

$$g(t) \cdot \exp(-\mu_1 t) = g_1(t) = 1 + \nu_2 t^2/2! + \ldots \nu_4 t^\nu/\nu! \quad [1.18]$$

A useful relationship between moments and cumulants exists as follows:

$$\mu_1 = k_1 \quad [1.19a]$$

$$\mu_2 = k_2 + k_1^2 \quad [1.19b]$$

$$\nu_2 = k_2 = \mu_2 - k_1^2 \quad [1.19c]$$

$$\nu_3 = k_3 = \mu_3 - 3k_2k_1 - k_1^3 \quad [1.19d]$$

$$\nu_4 = k_4 + 3\nu_2 k_2 = \mu_4 - 4k_3k_1 - 6k_2k_1^2 - k_1^4 \quad [1.19e]$$

$$\nu_5 = k_5 + 10\nu_3 k_2 \quad [1.19f]$$

While the non-central moments are a function of the origin, the cumulants are invariants like the central moments.

One final version of the moments may be given. It is sometimes useful to determine the moments from a different reference point than either the mean $\bar{x}$ or zero, let us say for γ. Then:

$$\bar{x} = N^{-1} \Sigma (x - \gamma) + \gamma \quad [1.20a]$$

$$\sigma^2 = N^{-1} \Sigma (x - \gamma)^2 - (\gamma - \bar{x})^2 \quad [1.20b]$$

$$\nu_3 = N^{-1} \Sigma (x - \gamma)^3 + 3(\gamma - \bar{x})\sigma^2 + (\gamma - \bar{x})^3 \quad [1.20c]$$

$$\nu_4 = N^{-1} \Sigma (x - \gamma)^4 + 4(\gamma - \bar{x})\nu^3 - 6(\gamma - \bar{x})^2\sigma^2 - (\gamma - \bar{x})^4 \quad [1.20d]$$

For $\gamma = 0$ this set of formulae reduces to the version of [1.11]. These formulae are practical for calculation by hand when e.g. the γ can be substituted as a whole number close to the mean and $(x - \gamma)$ can be expressed in whole numbers. This procedure speeds up calculations of the moments considerably. In electronic computer data handling the γ-version is generally not necessary but may increase accuracy.

1.3 *HOMOGENEITY AND PERSISTENCE*

These topics have been treated extensively by Essenwanger (1974a). It may be repeated here that the homogeneity of meteorological data sampling is always a major concern in any data collection or analysis work. No formalistic concept exists to guarantee homogeneity. The only assurance is the homogeneity of the physical processes and instrumentation. It is, therefore, always appropriate to investigate the homogeneous background of atmospheric data sampling.