Statistical Methods in the Atmospheric Sciences

Academic Press, Elsevier
Second Edition 2006
by Daniel S. Wilks

The subject book is a comprehensive treatment of statistics as it relates to the atmospheric sciences, and in particular, meteorology. It is greatly expanded over the first edition, and because it is a second edition has the advantage of corrections and additions that previous readers have suggested. This book, which will serve both as a textbook and a reference work, makes for very interesting reading. Subjects are treated extensively in words and illustrations, even to the arithmetic of the solutions of posed problems. Not only has the number of pages been increased by almost 35 percent, but the page size is larger. The chapter in the first edition on multivariate statistics has been expended into six chapters.

In the introductory material on probability, the author sets the stage for dealing with the atmosphere and forecasting in statistical terms, and states, "In order to deal quantitatively with uncertainty it is necessary to employ the tools of probability, which is the mathematical language of uncertainty." He has adroitly avoided explaining what probability *is*, but rather concentrates on the *meaning* or *interpretation*, and both the frequency view and Bayesian views are treated; he states, "Both of these dominant interpretations of probability have been accepted and useful in the atmospheric sciences."

In many ways, the book is a comprehensive review of the pertinent literature, and the references to sources are voluminous, both distant past and recent. It is obvious the work is up to date with the inclusion of more than one 2005 reference, and even an in press 2006 reference. The use of a small data set for several examples adds to the coherence, and the single author makes for consistent terminology and notation, a very important aspect for a textbook and reference.

In the chapter on empirical distributions and exploratory data analysis, methods of fitting and smoothing are presented. Correlations between pairs, the correlation matrix involving multiple variables, and serial correlation, along with various ways of presenting such correlations are presented.

The chapter on parametric probability distributions treats both discrete and continuous distributions, the parameters associated with each specific one, and how to fit them. This chapter sets the stage for the next–hypothesis testing. This latter chapter includes the usual tests, both parametric and non-parametric. In the discussion of the very useful paired t-test, although not stated explicitly, the way it is presented leads easily to the conclusion that the data themselves don't have to be normally distributed for the test to hold, but only the differences. However, a definition of the calculation involving the correlation would also be informative. The Kolmogorov-Smirnov test is presented, as well as the Lilliefors "correction" when the parameters of the distribution have been fitted to the data; it is likely the latter is many times overlooked. Even so, one must be careful in using the test (Steinskog et al. 2007)

The chapter on statistical forecasting treats the most used statistical technique, linear regression, with predictor selection and stopping rules; however, discriminant analysis is left for a later chapter after multivariate methods have been introduced. Nonlinear techniques discussed include logistic regression and Poisson regression. The definitions of classical, "perfect prog," and Model Output Statistics (MOS) techniques are presented. Ensemble forecasting is treated, which is a statistical technique in the sense that multiple runs of a dynamic model are made with different initial conditions, resulting in the possibility of statistical analysis as a post-processing step. The difficulty of choosing initial conditions and the different approaches are discussed. Finally, the important consideration of "subjective" probability forecasts is treated, and methods of presenting them to users are given, such as credible interval forecasts for the difficult situation when the variables are quasi-continuous.

The chapter on forecast verification is excellent and covers the usual topics: probabilistic and non-probabilistic forecasts, discrete and continuous distributions, verification of fields and of ensemble forecasts, economic value as opposed to skill and or accuracy, and sampling and inferences that can be drawn from such measurers. Forecast verification is a many-faceted problem, and the definitions, uses, pitfalls, and relationships among measures are well addressed. I note that the somewhat modern terminology like hit rate is used and defined, but the older more well known terminology is also retained. The difference between what is usually called the Brier Score and the score Brier actually presented, which he called the "Score P," is noted. The discussion and diagrams concerning reliability diagrams are excellent. Concerning the Relative Operating Characteristic (ROC) method of "forecast verification display" (his words), he early hints at its main deficiency by stating that "...it does not include the full information contained in the joint distribution of forecasts and observations." Later, he explicitly states, "... the calculations behind the ROC diagrams are carried out without regard to the specific values for the probability labels... That is, the actual forecast probabilities are used only to sort the elements of the joint distribution into a sequence of 2 x 2 tables, but otherwise their actual numerical values are immaterial." His discussion on pp. 294-298 is especially good. It is interesting to note Wilks does not even present the now fairly common practice of drawing the ROC curve and finding the area under it by modeling the data to a Gausian distribution resulting in a continuous curve rather than one of line segments; it seems this is one place in the book where the diagram, *appearing* continuous, does not match the text where trapezoids, which result from line segments, are discussed.

This is a thoroughly enjoyable book. Every practicing meteorologist should have one within easy access.

The chapter on time series was very interesting to me, both time domain and frequency domain, and the relationships presented to other techniques.

Before the chapters on multivariate statistics, a thorough review of matrix algebra is presented, as use of matrices is necessary to deal with multivariate methods in a concise manner. Then, the all important chapter on the multivariate normal distribution follows, the single and bivariate distributions having been presented previously. Principal component analysis and canonical correlation analysis are extremely well-presented, as are later discussions of discrimination, classification, and cluster analysis. Wilks mentions the very important aspect of (unrotated) principal components, that, since they are constructed to be uncorrelated, "strong interpretations of this sort (about physical relationships of atmospheric modes) are often not justified" for Empirical Orthogonal Functions (EOFs) after the first.

Throughout the book, a very few and small data samples are used to illustrate graphically and very clearly the statistical methods presented and this brings coherence to the book. It is clear the author is interested in the reader's comprehension of the material. The explanations are almost carried to the extreme, and in some places the repetition and increased length could be distracting to some readers. But not so to me, the redundancy is just right. The connections among the current topic and previous ones are almost always brought out clearly with figures and examples.

The index is adequate, but could be enhanced. For instance, kurtosis does not appear, though it is briefly mentioned in the text (p. 441).

The references are abundant. Even so, I was disappointed that some pioneering work was not mentioned. A book in which much of the material is not new, but just presented in a new and perhaps more integrated and coherent way, should reference the seminal work if possible. For instance, the material on Cost/Loss ratio used in so many meteorological works dealing with economic value, including this book, does not reference Jack Thompson's (1962) pioneering work, along with that of Glenn Brier, who brought the Cost/Loss ratio into meteorological view. While, as Wilks states, Liljas and Murphy (1994) unearthed two earlier very perceptive publications of Anders Angstrom dealing with the subject, I believe Wilks' statement that the Cost/Loss ratio "...has been frequently used since that time" is not correct. Angstrom's publication did not actually use the Cost/Loss terminology (op. cit.), brought into the literature by Thompson, and did not influence decades of work dealing with the Cost/Loss ratio; Angstrom's work doesn't seem to have been used in the meteorological literature before the Liljas and Murphy reference to it in 1994. In fact, they say (in 1994) Angstrom's two papers were "rediscovered only recently."

I was also disappointed the pioneering work of Bob Miller (1964) was only mentioned in passing as having possibly been the first to use multiple discriminant analysis (MDA) in prediction. Actually, it is my opinion, he did much more for MDA than that. His AMS monograph, referenced by Wilks, laid out clearly the theory and application of MDA, along with the "screening" selection of predictors. While earlier publications, including books, dealt with MDA, Millers's was the one that we all followed in the early days of statistical forecasting. Even the terminology and notation used in many places harks back to the Miller treatment. Bob also used the distance concept in turning the probabilities generated by MDA into discrete forecasts through essentially a classification method. In this, he recognized and treated the unequal variances of the discriminant functions. To be completely fair, it is noted that Miller's mentor Joe Bryan developed the MDA concept and as he once said, "taught it to Bob."[1]

In the many references to the similarity and dissimilarity of the various techniques, I find no mention of the association of the discriminant function in the case of two groups to linear regression. The coefficients in a regression equation where the predictand is binary and those of the single discriminant function are proportional. The regression limits the variance of the result of the function whereas the discriminant function does not, but except for that, the two analyses yield the same result and any post processing to classify the result can be the same.

I also find no mention that the logit curve is symmetric. This can pose a problem if the two "tails," the one for the probability of the event and the other for the probability of the non-event, are not both fit well. In fact, it almost seems the diagrams presented do not recognize that feature. In conformance with most of the texts of today, a departure from 50 years ago, the sample standard deviation is defined with the division n–1 rather than n (p. 27):

$$s = \{ [1/(n-1)] \sum (x_i - \bar{x})^2 \}^{1/2} \qquad (1)$$

This definition makes the sample standard deviation synonymous with the unbiased estimate of the population standard deviation from which the sample (might have been) drawn. But what if the sample were the total population, what would be the sample standard deviation? Presumably, the division there would be by n. As a minority opinion, I believe this change in definition is unfortunate and leads to more difficulties with notation, etc., than it solves.

In defining the (Pearson product moment) correlation coefficient, by using n–1 in the denominator of the covariance, the number of cases cancels out. But the computed correlation of the sample is an overestimate for a population from which the sample is drawn, and no attempt is made in its definition to adjust for that biased estimate, so why single out the standard deviation for that honor? Even though Wilks defines the sample standard deviation with n–1, the statement in parenthesis, "The division by n-1 rather than n often is done in order to compensate for the fact the ..." seems to be almost an apology, and the "often is done" sort of leaves the definition open to choice. This would have been an excellent place to clearly define the difference between a statistic computed from a sample, with perhaps its deficiencies, and an unbiased estimate of the population parameter. A short discussion of n versus n–1 is found on p. 116. The presentation of the correlation coefficient as "(nearly) the average product of the variables after conversion to standardized anomalies" is very interesting. I believe the "nearly" is necessitated only because of the use of n–1 in the definition of sample covariance and variance.

It is surprising that the mathematical definition of kurtosis is not presented along with that of skewness early on, and then mention of it later will have a foundation.

Some other places where I would have liked to see the seminal work referenced are for the S1 Score (Teweles and Wobus 1954) and Miller(1964) for the terminology Regression Estimation of Event Probabilities (REEP) used by many today. The Karl Pearson distance is defined, but there is no reference.

Usually the material is factual; however in a few instances opinion finds its way in. For instance, in discussing the conversion of probabilistic to non probabilistic forecasts, Wilks states that: "This unfortunate procedure is practiced with distressing frequency, and advocated under the rationale that non probabilistic forecasts are easer to understand." There is no question that information is lost; however, if one is servicing a user, the user's requirements must be considered and if not met, the user may ignore that source of forecasts completely, and that would, presumably, be a loss of information to him also. I believe rather than condemning the practice, all of us should engage in educating the user groups in the use of probability forecasts; that is a long process, and progress has been excruciatingly slow.

While there must be some errors in a book of this extent, I noticed very few. On p. 78, the gamma function is stated as being on the left hand side of equation 4.6; it should say right side. On p. 119, equation 4.76c should have $\sigma 1$ instead of $\mu 2$. The data points in Fig. 6.11 are very faint.

This is a thoroughly enjoyable book. Every practicing meteorologist should have one within easy access.

Please see the next page for reviewer information and references.

[1] *Joe Brian's work exists as his 1950 Harvard University Ph.D. thesis: A method for the exact determination of the characteristic equation and latent vectors of a matrix with applications to the discriminant function for more than two groups.*

Reviewer

Bob Glahn holds a Ph.D. degree in meteorology with a minor in statistics from Pennsylvania State University. He has been working in the field of applied statistics as related to meteorology since joining the Weather Bureau (now National Weather Service) in 1958. He has authored well over 100 reports and journal articles dealing with that topic.

References

Liljas, E., and A. H. Murphy, 1992: Anders Angstrom and his early papers on probability forecasting and the use/value of weather forecasts. *Bul. Amer. Meteor.* Soc., 75, 1227-1236

Miller, R. G., 1962: Statistical prediction by discriminant analysis. *Meteorological Monographs*, 4, No. 25, American Meteorological Society, 53 pp.

_____, 1964: Regression estimation of event probabilities. Travelers Research Center, Hartford, CN, Technical Report No. 1, Contract CWB-10704, 153 pp.

Steinskog, D. J., D. B. Tjrstheim, and N. G. Kvamstr, 2007: A cautionary note on the use of the Kolmogorov-Smirnov test for normality. *Mon. Wea. Rev*.,135, 1151-1157.

Teweles, S., Jr., and H. B. Wobus, 1954: Verification of prognostic charts. *Bull. Amer. Meteor. soc.*, 35, 455-463.

Thompson, J. C., Economic gains from scientific advances and operational improvements in meteorological prediction. J. Appl. Meteor. 1, 13-17.