Teacher's Corner

How to Read the Statistical Methods Literature: A Guide for Students

James R. MURPHY

Statistical methods papers are densely written. The writers assume that the readers already have sophisticated knowledge of the topic. In addition, a standard statistical notation has not been developed. Students who learn a technique in one notation may be confused when reading articles written with a different notation. This paper contains suggestions for making the student's task easier and more productive.

KEY WORDS: Pedagogy; Reading statistical methods; Teaching statistics.

1. INTRODUCTION

Several guides tell the nonstatistician how to read and interpret applied statistical results. Huff (1954) gives five points for the skeptical reader to keep in mind. Sackett (1991) and Colton (1979) each provide check lists to determine whether statistical methods are appropriately used in medical articles. There are also guides for reading papers containing complex mathematics, including Cowen (1991), Phanstiel (1990), Parke (1958), and Pemberton (1969). Schechtman (1987) suggests teaching biostatistics through reading the medical literature. All of the references given above provide useful information for reading technical material. However, there are specific problems in reading the statistical methods literature that these articles do not address. In this paper I present the outline of a general method for organizing such reading.

The statistical literature presents several challenges. First, various skills are needed in reading statistical methods: basic language skills, knowledge of statistical notation, algebraic skills, and, increasingly, some recognition of how computers function. Second, technical articles are rarely models of expository style. They tend to rely heavily on technical jargon and on an interplay between the written word and written notation. A concept that is difficult to explain in language is sometimes easily explained in notation, but the resulting dissonance between the two can make the article difficult to understand. Third, advanced articles assume that basic concepts do not need to be explained. Fourth, we are most comfortable with notation we learned in class. A different notation can be confusing. Finally, many papers discuss both statistical theory and the computational techniques necessary to implement the theory. The theory and computation are not always clearly separated.

Every statistician has to come to grips with these problems in reading the literature, but our individual solutions do not get passed on to students in any formal way. This means that each new group of students has to develop a way of dealing with these problems. In this paper I outline strategies that I use to read statistical methods papers. These suggestions were developed from trial and error, discussion with colleagues, and suggestions from Polya (1945). The outline was written for the student, and is intended for any class that requires reading papers from the literature. The first time that I give this outline to a class I ask students to use the outline while reading two articles that I select from journals such as *Statistics in Medicine* or the *Journal of the American Statistical Association*. The articles selected deal with the topic of the class, and are intended to be slightly above the knowledge level of the average student. Students write a summary of each article, answering the questions given in the outline. I have made no formal evaluation of the outline, but informal discussions with students suggest that it does help them to read the articles.

2. HOW TO READ THE STATISTICAL METHODS LITERATURE

2.1 Right Attitude and Environment (It is a Long Process; Be Comfortable)

I applaud everyone who finds it easy to read statistical methods papers. For the rest of us it is best to start with the right attitude: "This is going to take some time." I would set aside 4 hours at a minimum for a simple paper, and considerably longer for more complicated work. This does not have to be in a single large block of time, but it gives you an idea of the total amount of time that it might take. It helps to have a comfortable environment in which to work. A comfortable chair, good light, pencil, paper, and possibly a computer are useful accessories.

2.2 Focus on Why You are Reading the Article

The adage that "You can't see the forest for the trees" often applies when you read a complicated article. Before you begin to read, you should determine why you are reading this article. Focus on that main point.

James R. Murphy is Professor of Biostatistics, Department of Preventive Medicine and Biostatistics, University of Colorado Health Sciences Center, Denver, CO 80262.

Table 1. Sample Entry for a Bibliographic Database

Reference: Andrews, D. F. (1971), "Sequentially Designed Experiments for Screening Out Bad Models with *F*-tests," *Biometrika*, 58(3), 427.
Statistical theories involved: Linear models, sequential designs, sequential *F* tests.
Computational techniques used: Generation of random normal deviations, setting up a design space and choosing sample points based upon accumulating data, simulating data.
Distributional assumptions: Normal distribution; requires replicate measures.
Other relevant assumptions: The models discussed here exist in a hierarchy of polynomial models, and you are trying to choose the best order for the polynomial.
Dataset used: Simulated data.
Relevant cross-references: None dealing with designing a space for the experiment.
Notes: Possible use in Phase I trials or selecting models for decline rates in repeated measures. Basic proposal is to select design spaces that will allow you to determine which of a set of possible models is invalid, and then run your experiment on the most valid model.

A statistician has three basic reasons to read an article: general interest, relevance to a particular application, or broader knowledge of a specific statistical method. Gleser's (1986) suggestions for a fourth purpose of refereeing a paper are consistent with the advice given in this paper.

These reasons are not mutually exclusive. However, it helps to focus on one reason for reading a particular article. If the paper is of general interest, you would focus on the introduction and background. If it has information about an application, you would focus on the results and data section. Reading to improve your knowledge about statistical methodology requires the most comprehensive examination of the paper.

Whatever your reason for reading the article, it helps to start with Huff's first point: "*Who said it and where?.*" Do the authors already have an established reputation in this methodology? Is the article published in a journal where these methods are likely to have had rigorous editorial scrutiny?

2.3 State the Problem in Your Terms

Read enough of the *abstract, introduction,* and *discussion* so that you can state the problem in a sentence or two. State the problem in notational terms with which you are familiar. Sketch a possible way to solve the problem (or several if they occur to you) in terms of your present knowledge. It may help to skim the article, reading only the topic sentences in each paragraph to make sure that you understand all aspects of the problem. If, after doing this you, cannot state the problem clearly in your terms, look at the references. Is there an earlier attempt to solve this problem? This earlier article may state the problem in more familiar terms or may be by someone you know to be a good researcher and writer. A general text covering the problem discussed in the article may give you related material for solving this problem. You may need several iterations to grasp the problem.

When you can state the problem in your terms, read the *introduction again along with the methods section.* Pay particular attention to the assumptions being made and the limitations that these place on the solution being offered. Compare the methods to the sketch of a solution that you made. How does it differ? What points did you miss that this method considers? What assumptions did you make compared to the ones made here? If you are reading for general interest, this may be as far as you want to go. I recommend making a brief outline of what you have just done for future reference (see Section 2.8).

2.4 Find a Similar Problem with which You are Familiar and Work Through the Technical Details of the New Problem by Relating it to the Familiar One

From this point on assume that you want to use the results in this paper either in an application or to understand and develop new theory. Simply reading an article does not give you a complete understanding of its contents. Using or teaching the methods in the article provides a more complete understanding. If you have an opportunity through a journal club or a class to teach someone else about the article, you should do so. Even if you cannot teach someone else, begin to use the methods in the article.

Start by relating this problem to one with which you are already familiar. Follow the arguments and manipulations of the familiar problem, and broaden them to include your new problem. For example, to understand a paper on estimating parameters for linear models with stochastic parameters, you could relate the problem to one with fixed parameters, and examine the differences in the matrix structures, the effect on the Gauss–Markov solutions, the variability of the estimates, etc. Starting with a familiar problem gives you a firm base for pushing into unknown territory. There may be several different ways to approach your new problem. Different starting points should get you to the same place. The solution to your new problem will fit with your expanding knowledge base, and can be used in other problems.

2.5 Apply the Problem to Data

This is similar to point 2.4, but emphasizes using your new knowledge in a concrete way. Work through the techniques using a dataset that you know. Think about what would happen if these data had a different distribution or structure. What happens if the assumptions are violated? Many applied papers supply data that demonstrate the use of the techniques discussed. Such a dataset may demonstrate the technique to best advantage.

Simple numerical examples may also be helpful. Try applying that new matrix manipulation on a 2×2 matrix, and see what happens. If appropriate, program the techniques and examine the statistics as they are generated.

2.6 Separate Theory from Technical Details of Execution

To understand and use a new technique with facility you will need to understand both its theory and method of execution. When you are starting to read, however, it is a good idea to separate theory from execution. An estimate derived from mixed model theory may require the EM algo-

rithm for calculations. When reading the article keep clear which part of the discussion concerns the EM algorithm and which concerns the theory. In a particularly complex paper you may want to go through the points in this outline once for the theory arguments and once for the execution of the theory.

2.7 Read the Article at Least Three Times Emphasizing Different Sections Each Time

You have now read the paper through once, examined all sections in some detail, and obtained a good general understanding of the paper. The second time through the paper examine the *internal consistency* of the arguments, concentrating on the methods and results sections. Are the assumptions necessary and sufficient? Are the logic and notation straightforward and understandable? Do you understand the statistics, the probability theory, and the mathematics? Could you explain and defend this technique to statisticians at your level of experience and understanding?

As you ask these questions also consider what the authors could have done to make the task easier for you. Everything you think of here should be a candidate for inclusion in your own papers. This is a good time to examine the references again, and possibly examine companion papers that will shed light on your remaining questions. Begin to talk to colleagues, and consider unsolved problems that await your solutions. You may want to present some of your thoughts and get feedback from a group at this point. You should feel comfortable doing this because you now have a firm grasp on parts of the problem and can explain what you are still confused about. Boen (1982) has good suggestions about making presentations and answering questions in front of an audience.

Finally, read the paper for *external consistency* or generalizibility. Scrutinize the introduction, results or examples, and discussion sections. Find out how to use this technique, what kind of data it is useful for, where it fits into a range of solutions for problems of this type, and whether tested, stable, well-supported computer programs are available. Look at other applications of this or similar techniques that are in the references.

2.8 Consider Setting Up an Annotated Database

An annotated list of references will make it easier to review a technique. You are not outlining the paper; you just need to put enough down to make the paper easier to read next time. This database could be as simple as notecards or as sophisticated as using a computerized reference manager. Table 1 gives an example from my database, but be creative. You do not want all of your effort in reading the paper the first time to be lost when you do not use the procedure for a period of time. Also, note the good writers, theorists, and applications people as you find them. Not all statisticians are equally good in all areas, but you can pick the best in each area to emulate. You might even consider adding notes on the best presenters at meetings and what makes them good.

3. FINAL COMMENTS

Mark your progress by what you have done, not by what there is to do. The amount of literature is increasing exponentially, and you will never be able to read it all. You may, however, be able to read all of the good articles on a given topic. If you keep track of articles with a database, you will be surprised at how much of the literature you do read.

[Received March 1995. Revised August 1996.]

REFERENCES

Boen, J. R., and Zahn, D. A. (1982), *The Human Side of Statistical Consulting*, Belmont, CA: Wadsworth.

Colton, T. (1979), *Statistics in Medicine*, Boston: Little, Brown.

Cowen, A. C. (1991), "Teaching and Testing Mathematical Reading," *American Mathematical Monthly*, 98, 50–53.

Gleser, L. J. (1986), "Some Notes on Refereeing," *The American Statistician*, 40, 310–315.

Huff, D. (1954), *How to Lie with Statistics*, New York: W. W. Norton.

Parke, N. (1958), *Guide to the Literature of Mathematics and Physics* (2nd ed.), New York: Dover.

Pemberton, J. (1969), *How to Find Out in Mathematics: A Guide to Sources of Information*, Oxford: Pergamon Press.

Phanstiel, O. (1990), "How to Read Chemistry," *Journal of Chemical Education*, 67, 57–59.

Polya, G. (1945), *How to Solve It: A New Aspect of Mathematical Method*, Princeton: Princeton University Press.

Sackett, D. L., Haynes, R., B., Guyatt, G. H., and Tugwell, P. (1991), *Clinical Epidemiology: A Basic Science for Clinical Medicine*, Boston: Little, Brown.

Schechtman, K. B., and Spitznagel, E. L. (1987), "Teaching Biostatistics with an Emphasis on Reading the Medical Literature," in *ASA Proceedings of Statistical Education*, pp. 111–115.