Practicing Statistics or What they forgot to say in the classroom[[1]](#footnote-1)

This report gives a brief summary of some of the things that are important for a practicing statistician to know, but which are seldom taught in statistics courses. Issues covered include: tips on keeping up with the literature, useful computer programs, the importance of good interpersonal communication in collaborative work, ideas on data gathering and analysis and suggestions for improving report writing.

## Keeping up

“Staticians, in their consulting capacity, are much like lawyers. They do not need to remember every technique exactly, but must know where to locate it when needed and be able to understand it once found.”- S. R. Searle, *Linear Models*.

 Good practicing statisticians need to know a lot and to continue to learn at a rapid pace. They need to have a good understanding of statistical methods, computing techniques, scientific methods and the subject matter of the field in which they are working. Even just keeping up with all developments in statistics is clearly impossible. Last year alone there were probably in excess of 7500 articles published on statistics. One must be quite judicious in choosing what is worth scanning, and even more selective in deciding what is worthy of study.

 One factor that compounds the problem is that statistics courses tend to be totally void of any techniques for solving non standars problems on a short time schedule. For exanmple, these courses almost never teach one how to deal with missing data. What we need to know as practitioners is how to locate quickly good, but not necessarily narrowly optimal, solutions to problems that have been “solved” and how to develop reasonably good approaches to problems that may never have even been considered. We know of no boks in statistics that have this flavor. The classic in this general are is a small book by the mathematician George Polya entitled *How to Solve It*. An occasional perusal of that book is certainly in order for any serious practicing statistician.

 Text on mathematical statistics, while necessary, are certainly insufficient since their emphasis is uniformly on problems that have neat analytic solutions while the problems one encounters in practice seldom fall into that category. An understanding of the basics of mathematical statistics, however, is extremely helpful. One book that is worthy of mention here is *Bayesian Inference in Statistical Analysis* by Box and Tiao. The importance of this book is in the emphasis placed on (Bayesian) techniques for solving some bothersome problems in data analysis. Examples include tests for homogenity od variance when data are possibly non-normal (which is to say, always), dealing with outliers and problems associated with negative estimates of variance component.

## *Some good books on statistical methods*

With limited resources a statistician needs to be careful in selecting his library. The following are listed roughly in order of priority for a general practicing statistician. Someone specializing in biostatistics, survey methods or some other area would surely need a list more tailored to that field.

 Snedecor and Cochran, *Statistical Methods*. This book is a treasure trove of good ideas and techniques. If one could afford only one book, this would probably by the best.(And it’s relatively inexpensive.)

 Box, Hunter and Hunter, *Statistics for Experimenters*. A very good book, 10 years in the writing. Full of useful “how to do it” information. Particularly good introduction to the design of experiments and to the key ideas in scientific inferences with much practical advice.

 Neter and Wasserman, *Applied Linear Statistical Models.* A very good introduction to the linear models approach to regression, analysis of variance and analysis of covariance. This is especially important since in practice most problems at one stage or another must be reduced to some from of a linear model.

 Mosteller and Tukey, *Data Analysis and Regression.* A mind opener; full of novel ideas for exploratory analysis of data with much emphasis on creative graphic techniques. Careful study is richly rewarded.

 Daniel and Wood, *Fitting Equations to Data.* A good book for showing some of the unexpected things one can find by careful analysis. Another book by Daniel, *Applications of Statistics to Industrial Experimentation*, has a similar quality.

*More specialized books*

It’s a bit harder to know what to include in this category, but here’s a brief list of some books we find especially useful.

 *Survey sampling techniques.* Two good books are Cochran, *Sampling Techniques* and Kish, *Survey Sampling.* Cochran provides a more unified view of the appropriate formulas while Kish gives more detail on some of the important practical issues. A nice basic book is Slonim, *Sampling.*

 *Time series analysis.* The bible here is, of course, Box and Jenkins, *Time Series Analysis: Forecasting and Control.*

*Categorical data.* Good books here include Bishop, Fienberg, and Holland, *Discrete Multivariate Analysis* and Haberman, *Analysis of Qualitative Data*.

 *Linear models.* A personal favorite is Winer, *Statistical Principles in Experimental Desing.* A good treatment of the Standard approach to analyzing balanced complex designs.

 *Discriminant analysis.* This is a more narrow subject but Lachenbruch, *Discriminant Analysis,* is so nice it deserves mentioning.

*Journals*

It’s hard to imagine a person who is serious about being a practicing statistician who doesn’t regularly scan at least two journals. For general purposes, the best one is probably the *Journal of the American Statistical Association.* A useful journal that is relatively unknown outside of the United Kingdom is *The Statistician.* It seems to have an abnormally large selection of good expository articles.

 There are numerous other journals, perhaps over 50 devoted primarily to statistics. One way to select journals for regular review is to see which ones tend to publish articles of value in your area of application.

*Finding things*

Few people begin to take proper advantage of the great wealth of information that is published each year. If you were concerned about factor analysis, chances are that some of the more than 40 articles published last year on that subject would be useful. The *Current Index to Statistics,* published annually since 1975 by the *American Statistical Association* and the *Institute of Mathematical Statistics*, provides compherensive coverage of articles and books on statistics. Last year (1978) over 7500 articles were indexed. The preface to the *Current Index to Statistics* also lists 15 other indexes or related items of interest to statisticians. In the back of each issue the Index also lists the names of journals denoted to statistics.

## 2. Statistical Computing

Proper analysis of data requires many calculations—and plots. High speed computers and modern software are making these easier every day. Familiarity with at least one or two statistical computing systems is a virtual necessity fort he modern data analyst. Here are brief descriptions of the major ones.

 **SAS**, *Statistical Analysis System*. In many ways the best system available. Good data base management, flexible and powerful statistical analyses and a rapidly growing user library. It is unfortunately available only on IBM computers or 100% compatable machines such as Amdahl and ITEL, and then only under selected operating systems. It is batch oriented (expects output to be in big chunks on line printers) but can be used interactively on some computers.

 **SPSS,** *Statistical Package for the Social Sciences*. This package is currently used at more institutions than any other. By background, it is oriented more toward the social sciences. It is used primarily for its table making and data base management, but it also has a moderately good library of more sophisticated capabilities. This package is also batch oriented.

 **SCSS**, the conversational cousin of SPSS. It is designed to be interactive (it responds at the terminal as appropriate commands are keyed). This package is still in the early stages of development. Its language is not compatible with SPSS.

 **BMDP**, *Biomedical Computer Programs*. This is a collection of powerful programs for statistical analysis, and is generally of very high quality. The package’s principal weakness is that it really isn’t a system, rather more a collection of independent programs. There are no data base management capabilities in BMDP and it is often difficult to save intermediate results from one program to use in another. The programs are batch oriented.

 **Minitab.** This package is effective for basic analyses in either interactive or batch mode. Its chief strengths are ease of use, convenient graphics at the terminal and data manipulation. It has good regression and table making capabilities but lacks more powerful techniques. Minitab is very convenient for data exploration.

 **APL.** Many statisticians have been enthusiastic at one time or another about APL. It’s not really a statistical system, but is very easily programmed to do statistical calculations. Data base management is a real problem on many computers.

 One useful mode of operation is often to use Minitab in the exploratory phases of an analysis then switch to a more sophisticated system once appropriate techniques have been settled upon.

## 3. T he practice of consulting

Virtually all practicing statisticians function in a consulting or collaborative role, and as such, much of the difference between success and failure rests on their ability to communicate effectively. Interpersonal skills are thus of great importance.

 It’s relatively easy to give advice on how to communicate well, but it’s often another thing to put it into effect. Nevertheless, we won’t resist the temptation to give some advice.

-Develop a helpful and resourceful attitude.

-Communicate a genuine interest in understanding and helping solve the real problem.

-Learn something about the subject under consideration. Library research is often extremely helpful - - ask for suggestions for background reading - - then read some of them.

-Don’t be reluctant to ask questions when something isn!t clear. We find repeatedly that simple questions about seemingly minor details often bring to light misunderstandings of important issues.

-Ask to see the apparatus, the laboratory or other “on-site” locations of importance to the data gathering. Seeing that the hill has a step slope or that the mice are in flor to ceiling cages or that the delivery truck parks with its motor running right next to the air sampler often makes a great deal of difference in the value of a statistical “solution.”

-Replay what you’ve heard in different words using phrases lile, “Like me see if I understand this; do you mean that…?” and “Could I just check; you didn’t mean that…?”

-Write memoranda which give your understanding of the problem; these often bring to light stil new avenues for improvement.

 There are a number of papers on consulting, many of which contain useful suggestions. The bibliography by Woodward ans Schucany is quite good.

**4. Data gathering**

The most important contribution a statistician can make in any Project is to help develop a clear specification of the goals of the project. Seek to gain agreement among the collaborators as to what the real goals of the project are and where the real problems lie. Point out some consequences of what may seem like minolr differences in opinion among participants but may in fact not be. Then help get priorities set. Most projects cannot accomplish all that would be desirable. Start from the top - - What is the single most important objective of this project? If you could only answer one question, what question would you most like to answer? Suppose you had all the data in hand, would you be able to answer all the key questions?

 Again the use of memoranda to summarize your understanding of the goals, the proposed approach and the expected consequences is most helpful. Once the goals of the project have been established and agreed upon by all involved, the next step is the gathering of data fort he analysis. Great care must be taken here, as an ill-defined data base makes for an ill-defined analysis.

 Here is a brief check list we find useful in making plans for data gathering:

-formulate goals precisely

-quantify goals

-specify variables precisely

-specify classes of candidate models

-describe blocking, realm of generalizability

-consider alternative means of exposing true uncertainty

- specify how you will seek to estimate

 -precision

 -bias

-describe plan for data gathering in detail

-describe randomization procedure in detail

-describe data logging procedures in detail

-if data will be computerized, describe process

-specify ways that data will be checked

-specify how data will tentatively be explored, summarized and otherwise analyzed

-communicate alternative, tentative data gathering plans.

**5. Data Analysis**

Good data analysis, like good design, starts from carefully phrased questions. A very common mistake is to start instead from a predecided method of analysis. It's very easy to fall into the trap of doing an analysis of variance merely because the data have a form amenable to an analysis of variance. In practice, there are no standard problems, only standard solutions .

Continual interaction with subject matter specialists to phrase questions in practical, not statistical terms, is quite important. Once key questions have been agreed upon and tentative analysis plans have been developed, another memorandum summarizing these is in order.

During the analysis an interactive process such as that shown in Exhibit 1 is usually fruitful. The importance of the core triangle of MODEL IDENTIFICATION, MODEL FITTING and DIAGNOSTIC CHECKING has been forcefully indicated by George Box and colleagues, especially in Box and Jenkins (1970). Additional key components illustrated in Exibit 1 that all too often get overlooked in statistical practice are the need to pay careful attention to the data - and its quality - and to the underlying theory or structure of the problem.

**MODEL IDENTIFICATION**

THEORY

**DATA CHECKING**

**DIAGNOSTIC CHECKING**

RAW

 DATA

EVEN BETTER DATA

THEORETICALLY PLAUSIBLE DEPARTURES FROM MODEL

BETTER

DATA

**MODEL FITTING**

**MORE DATA CHECKING**

**Exhibit 1** Key steps in the iterative process of data analysis and model building.

Implicit in Exhibit 1 is the all important distinction between "assuming" and "pretending". *The American Heritage Dictionary* includes the following definitions.

- Assume: "To take for granted." and

- Pretend: "To play like, to make believe".

There is a subtle, but very important distinction between these two words. When doing statistical analysis our life is complicated by the fact that we must continually shift back and forth between these two concepts, and in the past have tended to rely on only one word - assume - to describe both.

When we tentatively define a plausible model INCLUDING THE NATURE OF THE DISTURBANCES, we put on our mathematician's hat and ask "What is an optimal, or at least defensible, way of fit¬ting this model to data having the indicated error structure?" To seek such answers, we must ASSUME that we know the nature of the model and the error structure perfectly. The mathematics takes us literally and treats the model as if it could "take for granted" everything we have said.

Having thus developed a fitting procedure, we proceed to apply it to the data at hand, but in so doing we switch from acting as mathematicians, and turn to being scientists. As scientists we can only PRETEND that the data can be completely described in the fashion ASSUMED in the mathematics. Of course, nature cares not at all about our play acting. Our pretending does not change the underlying model, nor the error structure. It will be whatever nature has chosen, and invariably nature has chosen a more complex structure than that described by the mathematics we have been able to accomodate.

We then move around the bend in Exhibit 1 to the analysis loop and begin DIAGNOSTIC CHECKING. There our role is to ask if there are serious DETECTABLE differences between nature and what we PRETENDED was true. Even while doing DIAGNOSTIC CHECKING we need to switch back and forth asking questions like, IF the disturbances were uncorrelated, what are the chances of observing a first order autocorrelation this high or higher.

Perhaps a useful way to sum this up is to say that it seems to be helpful to use two different words for the two different roles. Let us ASSUME when we're doing mathematics and let us PRETEND when we're fitting models to data. Using these two different words may help us keep from taking too seriously what we ordinarily ASSUME, but in fact can only PRETEND.

## Uncertainties

Once we have obtained estimates of the quantities of interest, the real problem of deciding on the amount of uncer¬tainty in these estimates begins. Almost all data are correla¬ted in some important ways, many of which are not easily quantifiable. In addition, biases, systematic errors or the effects of "lurking" variables may exist.

Then too, we often are forced to try a multiplicity of analyses on a given data set before deciding on our "final model." What effect the selection of the model that fits best has on the uncertainties of our estimates is often difficult to assess. One good approach is the use of cross validation in which half the data set is locked away during the explanatory phase to be used only as a check against the final model. When using cross validation it is important not to lock away a ran¬dom half, but rather to set aside a chunk which might differ systematically from the rest in some apparently random way. Thus if one has data from 10 schools, the data from 5 of them should be set aside, rather than setting aside a random half from each school. Only then do we give the real uncertainty a chance to express itself.

**B. The Advantage of Simplicity**

Simple analyses are easier to explain than complex ones - and are often less likely to lead to serious blunders or over¬sights. The second edition of Daniel and Wood, *Fitting Equations to Data*, will present an example (called the 10 variable example) where serious flaws in the data negated all of the dozen and a half sophisticated analyses that had been performed on that set of data. Our personal experience is quite similar.

**C. The Role of Assumptions**

We have already mentioned the importance of understanding the role of assumptions in our analyses. This information should also be communicated to the subject matter specialist. Herman Rubin's commandments are as follows:

*Herman Rubin's Commandments*

*For Client*

- Thou shalt know that thou must make assumptions

- Thou shalt not believe thy assumptions

*For Consultant*

- Thou shalt not make thy client's assumptions for him

- Thou shalt inform thy client of the consequences of his assumptions

*For Person Who is Both Client and Consultant*

- Thou shalt keep thy roles distinct lest thou violate some of the other commandments

**6. Report writing**

We all know the importance of report writing, but it remains a difficult and often tedious task. It's usually much more enjoyable to start some new task than to explain clearly what has been done in a project we're already tired of. Nevertheless the good practicing statistician must see that the written documentation is completed. Results that aren't properly summarized in writing are soon forgotten and/or misinterpreted.

In Appendix A we've included a brief check list that Wisconsin students taking the required consulting course have found useful in learning to write better reports. Perhaps some of these suggestions will be valuable to others as well. In addition, students are urged to read the little book on writing by Strunk and White; it's chock full of good tips on how to improve writing skills.

An important, and often overlooked, advantage of writing a report on a design or the results of an analysis, is that it forces one to summarize what has been done. In the process, omissions or slips often cctne to light that can better be corrected before the matter goes any further.

**7. Other references**

Two papers that support the approach to statistical practice espoused here are Marquardt (1979) and Joiner (1981).

**References**

Box, George E. P., Hunter, William G. and Hunter, J. Stuart (1978). Statistics for Experimenters, John Wiley and Sons,New York.

Box, G. E. P. and Jenkins, G. M. (1970). Time Series Analysis, Holden-Day, Inc., San Francisco.

Daniel, Cuthbert (1976). Applications of Statistics to Indus¬trial Experimentation, John Wiley and Sons, New York.

Daniel, Cuthbert and Wood, F. S. (1971). Fitting Equations to Data, John Wiley and Sons, New York.

Joiner, Brian L. (1981). "Consulting, Statistical," Encyclopedia of Statistical Sciences, N. L. Johnson and Samuel Kotz (ed.), John Wiley and Sons, New York.

Marquardt, Donald W. (1979). "Statistical Consulting in Industry," The American Statistician, 33, 102-107.

Mosteller, F. and Tukey, J. W. (1977). Data Analysis and Regression, Addison-Wesley Publishing Co., Reading, Massachusetts.

Neter, John and Wasserman, William (1974). Applied Linear Statistical Models, Irwin-Dorsey Limited, Ontario.

Polya, George (1957). How to Solve It, Doubleday and Company, New York.

Snedecor, G. W. and Cochran, W. G. (1967). Statistical Methods, Iowa State University Press, Ames, Iowa.

Strunk, William and White, E. B. (1979). The Elements of Style, MacMillan Publishing Co., New York.

Tukey, John W. (1977). Exploratory Data Analysis, Addison- Wesley Publishing Co., Reading, Massachusetts.

Woodward, Wayne A. and Schucany, William R. (1977). "Bibio-graphy for Statistical Consulting," Biometrics, 33, 564¬565.

**APPENDIX A**

*Suggestions Made to Students for Writing Better Reports*

Don't write an odyssey: "An extended adventurous wandering."

Do write a report to a client. Explain what you now under¬stand about the problem, with very little emphasis on how you happened to get there.

Do give practical interpretations of results, in language the client will be very comfortable with. For example, "When flow rate increases by 10%, wasted material increases by about 5% (95% confidence interval is 4.1 to 6.3%)." "It is not possible to separate the effects of flow rate and temperature since these factors were not separately varied."

Do be as brief as you can while still including all important details. Key aspects of six regression outputs can often be better summarized in one table on half of a page. Many plots can be summarized in one sentence.

Do include summary tables of important results in the body of the report.

Do learn something about the field you're working in. Find out how an ammonia plant works, why stack loss is bad, what happens when you change one or more input characteristics , which are controllable variables and which are uncontrollable, what the purpose of the project was or might have been, and use this knowledge in your design or analysis. (Technical encyclopedias and libraries can be very useful here.)

Do label all figures and tables so well that each is under¬standable when viewed alone.

Do be careful with the word "assume". It's a very dangerous tool that very often works against proper analysis. If you make assumptions in analysis, always remember that these assumptions are virtually never true. The word "pretend" is closer to what we must really do in analysis.

1. Joiner, Brian L. - Pollack, Alison K.(1982) Practicing Statistics or What they forgot to say in the classroom, in Rustagi, J.S.- Wolfe, D.A. (eds.) (1982) *Teaching of Statistics and Statistical Consulting*, Elsevier Inc, Academic Press Inc.:327-342. [↑](#footnote-ref-1)