Good Writing in Mathematical Statistics

J. S. Marron*
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260
U.S.A.

November 18, 1992

Abstract

Good writing style for papers in mathematical statistics is important for two reasons. First it enhances a paper's chances to be accepted for publication. Second the literature has grown to the point that keeping up with it makes careful presentation imperative. This paper gives suggestions for improvement in two directions: presentation of mathematics and organization of papers.

1 Introduction

There is substantial dissatisfaction with many journals publishing papers on mathematical statistics today. In particular many journals are all too often criticized as being "unreadable". This is a matter of serious concern, especially for journals run by professional societies, which are under at least some obligation to be accessible to a wide range of readers.

This dissatisfaction can be traced to several possible reasons:

*This paper was supported by NSF Grant DMS-9203135. Helpful comments have been made by Ray Carroll, Peter Hall, Robert Lund and Lynn Seymour.

1. Mathematics is by nature difficult to read, requiring very substantial effort by the reader.

2. Some readers find it hard to generate the energy required to read mathematical papers.

3. Too many authors are sloppy in their presentation, writing only with their own needs in mind, ignoring those of the reader.

While nothing can be done about the first two reasons, there is much room for progress on the third. This paper attempts to start discussion concerning how this third point should be addressed. Specific suggestions are made here in two directions.

One direction is style of mathematical presentation. It is argued that writing style makes a very large difference in the understandability of mathematics. The main premise is:

- It is much easier to digest mathematics, especially new or unfamiliar notation, if the reader *first* understands the main idea at an intuitive level.

Section 4 shows, through a simple example, how reading can be very difficult when this principle is ignored. Then it is seen how much easier the reading becomes when thought is given with this principle in mind.

The other direction is organization of papers. A number of suggestions are made in section 3 on this topic. The main premise is:

- Reading will be done by a variety of readers reading at different levels, so the paper should be organized in a way which facilitates the process for as many as possible.

In both of these directions it appears that too many authors of papers in mathematical statistics (and in fact statistics in general), write papers with only themselves in mind. Sometimes I am left with a feeling that the author of a paper feels only those readers willing to slog through the mathematics (i.e. work hard to figure out the main ideas themselves) deserve to share in the main ideas of the paper. This is probably not true in most cases, since most who take the trouble to write a paper want it to be noticed as

widely as possible. But the fact remains that most authors write with only their own needs in mind, having little or no apparent regard for the reader. In particular there is a tendency to write things in only the first way that comes to mind, instead of choosing carefully from the several possible ways one could say the same thing. Many options are present both "locally" and "globally". "Local" concerns such issues as choice of phrasing (which is too personal for comment here) and presentation of mathematics, as discussed in section 4. "Global" options involve organizational issues, as discussed in section 3. It is not surprising that sloppy presentation typically gets worse when the subject matter becomes more mathematically sophisticated (and the author needs to concentrate more on that aspect).

Much of what is contained in this paper is common sense when one considers the viewpoint of the reader. Some of the suggestions made here will likely be considered controversial, since "style" is a highly personal matter. It seems likely there are some important ideas not brought up, although those deliberately not discussed are listed in section 5. There will doubtless be situations where good authors will perceive good reason for going against some of the specifics given here. However, it is hoped that there can be a general consensus on the underlying principles stated here and on the feeling that there is a need for improvement.

Section 2 discusses the specific and varied needs of the principal readers mathematical statistics: the researchers. It is interesting that careful thought about these needs will result in large benefits for two other groups as well: most other readers (including non statisticians) and also authors. Reaching the first group increases the chances that a given mathematical statistical paper can do what it is (or at least should be) intended for: providing useful information to end users of statistics.

Benefits for authors are less immediately obvious, since more work on the part of the author is required. However I submit that careful effort in these directions is well worthwhile. My experience as Associate Editor has made it quite clear that, especially for less well known authors, readability of a paper has a lot to do with its chances for being accepted for publication in a major journal. The reason for this is that most referees are very busy people, who do not enjoy the job. Most are not in a good mood when they open an envelope containing yet another request to serve as reviewer. Papers which are short, and make it clear as early on as possible what the main ideas are, and why they are important, have a good chance of changing

this bad attitude early on, and ending up with a friendly review. On the other hand, papers which are long, and require a lot of difficult digestion of mathematics *before* the main ideas become clear will deepen the reviewer's bad mood. Good referees will wade try to wade through the mathematics, to see if the ideas in such papers merit further consideration. But referees are only human, and do not have infinite time, so all too many will give up fairly early. Poorly written papers encourage referees to find some excuse for rejection of the paper, especially when they can use it to avoid trying to figure out the whole thing. Well written papers capture the referees' interest, and make them want to read on.

A quick summary of the main points is given in Section 6.

My personal motivation for writing this paper came from putting these ideas over and over into reviews of papers, in the roles of both Referee and Associate Editor. While most of these seem obvious, enough authors appear to be ignorant of them, that a published accounting seems worthwhile.

There are many books available on technical writing in general. Ms. Ros Boyce of the Australian National University, kindly did a computer literature search, using the subject "Technical Writing" revealing more than 100 such monographs. The recent ones are ubiquitous in most publisher's advertising pamphlets. However these do not seem to have had much effect on research papers in mathematical statistics.

2 Needs of the Reader

All modern researchers need to keep abreast of a very large, active, and growing literature. This has become an increasingly daunting task. Hence it is increasingly important for researchers to make the process as easy as possible for each other. Unfortunately many researchers curse the difficulty of keeping up with the literature, and then turn around and contribute to the problem (for others) with their next paper. A possible defense would be to point to the rather large number of important but impenetrable papers, but many of those were written in times when the literature was of a more manageable size.

Before suggestions for addressing the problem are presented, first consider how most researchers approach the literature. A common strategy is:

1. Scan a given issue of a journal for titles (and perhaps authors), picking

only a few of those for more careful study.

2. Of those chosen in 1, read the abstracts, and again select only a few of these for deeper study.

3. Of those selected in 2, read the introductions, and perhaps selected other sections, to get as much of the main ideas as possible in a short time, for example 10-20 minutes.

4. Only in cases where the reader is *directly* working in the area, and can thus justify spending a substantial amount of time, read a paper from 3 in detail (even then often ignoring the proofs).

A well written paper will attempt to simultaneously take the needs of each these readers into account. Specific suggestions for this are given in section 3.

3 Suggestions for good writing

This section makes some specific recommendations for enhancing accessibility to papers. I considered looking for, and referencing especially bad examples in each case, but decided against this. In addition to serving no useful purpose, I am far from being able to make a claim of perfection on these points myself.

3.1 Title

Ensure that the title maximizes usefulness to the reader at point 1 in section 2. There should be two goals here:

- Brevity
- Maximal information.

Brevity of a title is important because the reader obtains the needed information more rapidly if all the titles in a journal (most of which are pursued no further) are concise. The title should attempt to contain neither

all the information usually put in the abstract, nor microscopically precise information as to what the paper is about.

Of course maximal information is important, to aid in the decision problem going on. I suggest about one full line of type as being, in most cases, long enough to interest the desired audience, and yet not too burdensome on those who really want to read the other papers in the journal. Key words should be included which make the topic generally clear.

One way of pursuing these twin goals, is to first start with a title, and then ask: "would I look further at this paper if I saw this title?". If so, then ask, "which words could be cut out, without diminishing my desire to look further?" Even authors who are consistently good at the first step, can often improve their titles by doing this trimming at the second.

Note that properly addressing these twin goals often means authors will have to forsake "clever titles". For example plays on words can be fun for experts to read. However these often make it hard for those same experts to find the paper, so the joke can easily be wasted.

3.2 Abstract

Carefully choose material for inclusion for the abstract. Again attempt to balance the twin goals of brevity and maximal information content. Here there is room for more detail, but surely not enough room for all ideas covered in the paper. Be sure to include mention of the parts you are most proud of, and those that should interest the largest number of readers.

Any recommendations for length here must be more case dependent. Longer papers will usually need longer abstracts. However, I suggest something between 4 and 10 sentences in most cases.

Note that mathematical notation rarely serves any useful purpose in abstracts. In almost all cases, the ideas appropriate to this level of communication are better conveyed through words. Sometimes, when notation is introduced in an abstract, it is not used for anything else at all! Even when it is used for a real purpose, the point can almost always be conveyed more efficiently in words.

3.3 Introduction

With the goal of 3 from section 2 firmly in mind, an introduction should summarize all the main ideas. This point will surely be controversial, but I believe this should also be done with a minimum of mathematical notation. Certainly there are many instances where the presentation is helped by use of notation, but I suggest careful thought be given as to how much should be used. It should be kept in mind that each additional piece creates a burden on the reader, so a conscientious trade-off should be made by the author. When notation is introduced, the main principle in section 4 should be kept in mind.

3.4 Figures

Figures are very useful and important parts of some papers, however there is room for improvement in this area too.

A common problem is presentation of too many figures. Some authors seem to feel that every picture they generate in a project should be included in the paper.This wastes journal space and detracts from the main points of the paper.

After each picture is made, and the ideas behind it written up, the author should ask "Is the picture essential to the points being made?" Other important questions are "Can I use a simple verbal description, instead of including an entire picture?" and "Is this picture enough like a previous one that I need only give a verbal comment on how they compare?"

As to the content of figures, some papers have been written about what constitutes "good graphics", and what is "bad". Certainly the subject is well worth discussion and study. However I personally feel it is a mistake to suggest a set of rules. Such a step seems as doomed to failure as an attempt to determine what is "good art".

3.5 Conclusions

My recommendation here will surely be controversial. I know many people who believe a good paper is wound up with some conclusions which highlight a few of the most important lessons one obtains from the paper. There is a lot to be said for this viewpoint, and when done properly it gives an elegant

result. However, it is inconsistent with the way that most people read papers. In particular, the only people who see the conclusion are those who read the paper carefully (which is a small fraction of those who could get something useful from the ideas).

I suggest instead that all ideas which are saved for the conclusion are more appropriately placed in the introduction. It is not so elegant, since the conclusions are not properly backed up at that point. However, it will have the positive effect of leading those who have doubts to read further, and more carefully. There will be resistance to this suggestion by those who like to "pull the rabbit from the hat at the end" for maximal impact. However the impact is wasted on (and in fact the main point is missed by) the large group of readers who quit reading before the end. The overriding goal should be to get the main points to the reader most effectively.

4 Presentation of mathematics

When you do get into the meaty mathematics, try hard to remember the following basic principle of good mathematical writing, which is repeated from the introduction.

- It is much easier to digest mathematics, especially new or unfamiliar notation, if the reader *first* understands the main idea at an intuitive level.

Here is an example, first stating things poorly, using only mathematics, with no explanation:

> Given $x_1, ..., x_n$ with n odd, let $w_1 = \min\{x_i : i = 1, ..., n\}$, and then recursively define for $j = 1, ..., n$,
>
> $$w_j = \min\left(\{x_i : i = 1, ..., n\} \setminus \{w_1, ..., w_{j-1}\}\right),$$
>
> and use these to define for $k = 1, ..., n$, $u_k = \left|w_j - w_{\frac{n+1}{2}}\right|$, and then let $z_1 = \min\{u_k : k = 1, ..., n\}$, and then recursively define for $l = 1, ..., n$,
>
> $$z_l = \min\left(\{u_k : k = 1, ..., n\} \setminus \{z_1, ..., z_{l-1}\}\right),$$
>
> and then let $a = z_{\frac{n+1}{2}}$.

That was an intentionally clumsy introduction (featuring intentionally unclear and non-mnemonic notation) to the concept of the median absolute deviation from the median. However it illustrates something which happens frequently: an idea is presented in the first way that comes to mind, with no thought of accessibility for the reader.

Next the same idea is presented, using the same clumsy set of ideas, and also the same non-mnemonic notation, in a way that is much easier to digest simply because each bit of mathematics is preceded by a short indication *in words*, of what the author (and thus the reader) has in mind:

> The concept of "median absolute deviation" of a set of numbers $x_1, ..., x_n$, is based on taking the median of the "deviations", which are the set of distances from each point to the median. The median is the central "order statistic". Order statistics are a relabeling of $x_1, ..., x_n$, in increasing order, which can be defined as: $w_1 = \min \{x_i : i = 1, ..., n\}$, and then recursively for $j = 1, ..., n$,
>
> $$w_j = \min \left(\{x_i : i = 1, ..., n\} \setminus \{w_1, ..., w_{j-1}\} \right).$$
>
> When n is odd, the index of the central order statistic is $\frac{n+1}{2}$, so the median is given by $w_{\frac{n+1}{2}}$. Next define the deviations, for $k = 1, ..., n$, $u_k = \left| w_j - w_{\frac{n+1}{2}} \right|$. Now again apply the concept of median, but this time to the deviations, $u_1, ..., u_n$. The corresponding order statistics, defined in the same recursive fashion as above, are $z_1 = \min \{u_k : k = 1, ..., n\}$, and for $l = 1, ..., n$,
>
> $$z_l = \min \left(\{u_k : k = 1, ..., n\} \setminus \{z_1, ..., z_{l-1}\} \right).$$
>
> The median absolute deviation is then the median of these, given by $a = z_{\frac{n+1}{2}}$.

Note that the second is much easier to digest, because there is no need to individually figure out the idea behind each piece of notation (it is explained first in each case). Now the same concept is presented, using a better overall approach, and more mnemonic notation. It may be possible to improve this further, but the point here is how much easier it is to digest this than the first attempt.

An important basis for the concept of "median absolute deviation" of a set of numbers, is the notion of median. The median of $x_1, ..., x_n$, is the central "order statistic". Order statistics are a relabeling of $x_1, ..., x_n$, in increasing order, say

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}.$$

When n is odd, the index of the central order statistic is $\frac{n+1}{2}$, so the median is given by $m = x_{(\frac{n+1}{2})}$. The median absolute deviation of $x_1, ..., x_n$ involves two applications of this idea. First find the "center point" of the numbers, given by m above. Then to measure "spread", one studies the set of distances from each point to the center, called the "deviations". These are given by $d_i = |x_i - m|$, $i = 1, ..., n$. The median absolute deviation is then the median applied to this set of numbers, $MAD = d_{(\frac{n+1}{2})}$.

These points make a huge difference in how much time is required for reading two papers which have exactly the same content, but are written differently. There is some cost in terms of added length, but I submit that time on the part of researchers should have a higher priority than journal space.

5 Points deliberately omitted

Some points that might be considered in a discussion of "good writing style" have been deliberately omitted here. These include:

1. Length of papers. This must necessarily be very case wise, depending on the topic being addressed. However, it is important to realize that shorter papers have a much higher chance of acceptance, mostly because referees feel much more kindly disposed to them (think of your own feelings when asked to a review a paper with more than about 25 pages).

2. The proportion of text to mathematics. This also must be very case wise, and it seems impossible to set sensible general guidelines.

3. The "I" vs. "we" issue. I have heard this hotly debated, but my feeling is that everybody is right. In my opinion, choice of which to use is a personal matter. I believe authors should choose whichever they feel most comfortable with, as this stimulates creativity and allows more thought to be devoted to more important issues of good writing. However personal pronouns appearing too frequently can be distracting, and should be avoided.

6 Summary of recommendations

- Title: 1 full line
- Abstract: 4 - 10 sentences
- Introduction: words only, or at least an absolute minimum of mathematics
- Conclusions: not needed since main ideas should be in the introduction
- Presenting mathematics: explain ideas behind notation *first*