Comments on “Introduction to the practice of Statistics”
By David S. Moore and George P. McCabe
Muhammad Zafrullah

A good elementary level book should cover basic material in such a manner that the students can learn step by step. The organization of the material should be such that the material already covered is useful in understanding new concepts. The support material such as a CD or a manual for the use of certain software should be easily accessible and as free of errors as possible.

a. This book describes data in its preface “To the Student” and nowhere else. Then the book categorizes data as qualitative and quantitative and does not bother with the further subdivision of quantitative data as: discrete and continuous. This makes the understanding of discrete and continuous random variables somewhat difficult.
b. The book talks quite nicely about graphical description and histograms, but skips the details about how to subdivide a range into intervals and the authors do not seem to be bothered by the fact that the readers of the book will not be able to tell to which interval should a piece of data belong. They seem to want to relegate completely the task of deciding on the cut points to the software, but they never specify which software and how. There does appear to be an example of a sort of pre-assigned interval lengths (page 14, Ex 1.9), but the plan of how to assign a measurement to an interval is: x is in the interval if min of interval <x ≤ max of interval. The trouble with this plan is that it does not match any software or any calculator and of course it causes confusion and disbelief among the students.
c. Histograms become a useful graphical tool if you know how to read them and I see few exercises involving reading them. Just talking about symmetry is not enough. Also the students should know that although the identity, of each of the individual data pieces, is lost in a histogram but still we can get an estimate of the mean using the frequencies. This can later be put to use in the definition of the mean of a probability distribution. Then there is the somewhat interesting remark on page 16: About half of the states have less than 4% Hispanics (Example 1.10). How do the authors get that? Are they using areas in a discrete situation? Are they eyeballing it? Have they trained the students to do that?
d. While the description of quartiles and the comparison of mean and median is adequate, the procedure for finding percentiles is shoddy. Yet some elaborate questions are asked in the exercises. On the other hand the authors give an elaborate (standard) algorithm for finding the quartiles yet the answer to at least one problem (Problem 1.41) was given using “interpolating software”. Imagine the frustration of a student who “does everything right” only to find out that his/her answer does not match. “The interpolating software” explanation was given to me by one of the authors. (Of course he agreed that in the presence of an elaborate algorithm the answers must match.)

e. The coverage of standard deviation is standard and actually quite good, the formula is given, the importance of variance in describing the spread of data is mentioned, but unlike other books of a similar level no systematic method is given to compute it and of course the book does not at all talk about the gadgetry that would let the students compute and see for themselves. Similarly, at another point, the authors recommend, “In practice you should use software or a calculator that finds *r* from keyed in values of the variables *x* and *y*. Exercise 2.19 asks you to find the correlation step-by-step from the definition to solidify its meaning.” Well said but absent is the procedure that would let the student tabulate those computations. I do agree with the authors of the book and I wish they had said it more often and I wish they had brought in some standard procedures to consolidate the ideas.

f. While box-plots are employed, no reason is given for why they are so important. Definitions of some standard and important concepts are relegated to exercises such as the trimmed mean and only one exercise (1.76 page 65) is deemed sufficient. Similarly Systematic Random Sampling (SysRS) becomes important when you have to choose a small sample from a relatively large population and when you do not wish to miss out on any part. The example (relegated to an exercise (Ex. 346 p 257) talks about how to find a sample of five from a population of size 200 and does not talk about the case when the population happens to be 203. SysRS is significantly different from stratified random sampling in that in SysRS you make the strata yourself depending on the size of the sample and in the case of stratified random sampling you use the strata present in the population. It is sad to note that the somewhat important topic of stratified random sampling is “explained” with a real life example (Example 3.46, page 257) for which it impossible to give details and even if all the details are given the situation involves too many variables to understand the main idea behind. I think a simple, “artificial” example must precede this example.

g. The description of density curves is so elaborate that one wonders why the authors use z-scores and percentages at all. I believe the authors should bring in the notion of non-standard density curves and standardized density curves and then say that the area under a standardized density curve is one, or just stick to percentages. The analogy of a density curve to a lamina (page 67) is inappropriately presented in that if we are dealing with a lamina here, then the center of mass (or the point of balance) would be somewhere inside the lamina and the lamina can be balanced about any vertical line that passes through the center of mass and about which the moments balance out. What the authors have done happens usually with the people who talk lamina and think thin rods. In my opinion, if a lamina description is to work then the lamina has to be in a vertical xy-plane with the straight edge parallel to the x-axis. Then the mean of the data is the x-coordinate of the center of mass. I must note that when we draw a density curve on a histogram, we are assuming that the distribution represented by the histogram is the distribution of a sample from a population that is continuous with distribution approximately represented by the density curve. So, in my opinion, some notion of samples as sources of data, some idea of what a continuous random variable is has to precede the introduction of density curves.

h. The talk of normal distribution and use of standard normal tables is standard but misplaced. I wished that my students knew about continuous random variables and probability before doing normal distribution. Then a good deal of problems could be handled with more facility.

i. Normal quantile plots, which are akin to normality plots are described well but the procedure of constructing them does not match with standard software and TI-83. The result can be chaotic if the teacher does not explain the situation. Also there is the question of finding the z-scores for the hundredth percentile. (This seems to be a well-known flaw in the given procedure given on page 79 of the book. See e.g. Radford Neal's web page http://www.cs.toronto.edu/~radford/mm-errata/errata.html.) I think the problem can be solved by allotting the z-score 3 to the last entry, you do not need to be very exact doing statistics! Besides, I think that this "flaw" is tongue in the cheek sort of thing. The normal curve is supposed to extend to infinity on both sides, except for deeply theoretical computations no one wants to go all the way to infinity.

j. The treatment of regression is interesting, though I wish I did not have to come up with those simple examples of common response and confounding that make theory much more palatable. I also wish that instead of making long speeches the authors provide simple and understandable examples. I think a worked example like Exercise 2.19, with a table would be a good addition. (I usually tell my students: If you do not have a calculator that can find *r* for you then this table is your only way out. I also think that the authors should work on their definition of association between two variables. Their definition is very scholarly, covers every aspect but fails to register with the students, because it is too vague if you do not know what it is about. I had to write up my own piece on relationships. If the authors wish I can send it to them. Here it would be unfair to my students if I do not mention Ex. 2.5 (page 117). Without any warning and/ or training the students are given the table with 83, 70, 61 in front of 29 and are expected to punch the data correctly in. At such instances I have wondered if the authors are deliberately trying to confuse the students.

k. The chapter on producing data gives adequate coverage of observational study and designed experiment and things to do with block design yet the only example given, in the book, of stratified random sampling is extremely confusing (Example 3.17, page 251). Try reading it for a change. If this "real life" example had come after a rather simple artificial example, that I had to bring in, then probably better understanding of the concept could be expected. On the other hand the notions of variability of a statistic being linked with the size of the sample is hit home quite nicely. Section 3.4 and its exercises and examples are actually very good. One of the projects that I gave my class was to verify Example 3.22, page 262. Hopefully this verification has hit home the idea of proportion and the laws governing the distribution of proportions of samples. The other project is to do Exercises 3.72 and 3.73. Please note that these exercises presume the student being knowledgeable and proficient in the use of suitable software. (Another such example is Exercise 5.52, page 409.) So either you have a very elaborate computer lab with lab technicians knowledgeable in the use of various technical software, or, the training of your students remains incomplete. I think I will go for a book written by a lesser author, that has exercises and examples within the reach of the student, software-wise and otherwise. In this context, I must mention Ex 1.78 (page 84) where the authors say, "If you ask a computer to generate random numbers …" This raises questions in the students' minds, which must be answered by showing them how, and mind you a calculator cannot do it. While we are talking computers, I must say that the following approach exists at a lot of small schools: The students' grades are not very good, so let us buy some computers. But very few of the "regular faculty" have some idea of how to use computers, so often computers sit there gathering dust. They do occasionally permit adjuncts or visitors to use them but then only the preferred ones can use them. When this is the situation, it is cruel to mention computers.

l. In probability the "frequentist" point of view is stressed and the usual simple set theoretic approach is not given adequate coverage. The result is that even the very basic notions of probability have to be pushed into the starred section. I do not mind this as long as I do not see any problems with it. In their eagerness to get to the "practical" things the authors talk about disjoint events, and then about independent events. They seem to be very concerned about pressing home the idea that if two events are disjoint then their probabilities add and when they are independent, their probabilities multiply. Very true, but the effect, in the absence of any example of "non-independent non-disjoint event" is that wherever the student finds P(A and B) the student would have the urge to multiply P(A) and P(B). The classic occurs on page 294. In order to stress that the independent events are non-disjoint the authors go like this: "Suppose you toss a balanced coin twice. You are counting heads, so two events of interest are A = {first toss is a head}, B = {second toss is a head}. The events A and B are not disjoint." By this logic then A1 = {first toss is a head}, B1 = {second toss is not a head} are either not of interest or they are disjoint! If that were so what would happen to Bernoulli trials (which are mentioned in passing) and what would happen to binomial distributions! (I mean what would be the student's plight who gets confused by the above presentation?) I think a little bit of elaboration could have simplified the matter. That is if the authors had taken their time to actually bring the ordered pairs in the picture. (In this connection I am a follower of Professor M.F. Neuts [Probability, Allyn and Bacon, Inc. Boston 1973] who, after defining (Cartesian) product of two probability spaces, remarks on page 80 as follows," The product space (O,B,P) is the natural framework to discuss the combination of two unrelated (or independent) experiments. The appropriate families of independent events are included in the definition of product space itself." A lot of the independent events joined with *and* that show up actually seem to be Cartesian products of events from two sample spaces serving as sample spaces for two independent experiments. At least at the elementary level this seems to be the case. There is of course another way of avoiding confusion. In the not very distant past, people resolved problems to do with independence by first defining conditional probability: P(A given that B) using Venn diagrams and then saying that events A and B are independent if P(A given that B) = P(A) and from this it would follow that A and B are independent if and only if P(A and B) = P(A)P(B). (No need to admit that we cannot make pictures of independent events as is done in the book.) I myself have taught these concepts as part of a pre-calculus Mathematics course in the 80’s and I know that these concepts can be made quite palatable to students.

m. The chapter on sampling distribution was not too hard for my students, thanks possibly to the projects that they had done and I have not seen any problem with intervals of confidence, levels of significance etc. But I must talk about the uneasy feeling that perhaps the authors were trying to confuse the students. On page 426 (line –4, -3 etc.) the authors mention the situation in which the standard deviation of the population is not known and all we have is the mean and standard deviation of the sample. Then comes Exercise 6.2, where the mean of the sample is explicitly given and then, “Assume that the standard deviation is $80.” Standard deviation of what? Sample? Population. Only those would know that it does not matter who did not miss those two lines.

n. This much about statistics and I am not a professional statistician. In teaching and producing Mathematics I have learnt that a timely example, however simple it may be, is better than a thousand scholarly speeches. In this book you see long passages and a few examples and some of them are either too confusing or too advanced for the presentation.

o. The organization or rather disorganization of exercises is superb, if the purpose of the book is to confuse. The data is given with the exercise, which is a good idea but then perhaps to save the space the tables have to be moved around a bit. The result is that I have seen some students use Table 1.5 for Exercise 1.29 (pp. 31 and 32) and some who chose Table 1.4 (which is the right table for the problem) missing out part c of the Exercise. The CD also does its part in adding to the confusion. Data sets for exercises and for tables are given separately. Then the labeling of the data sets in the CD is their own and does not match the labeling in the book (If in the book the label of an exercise is 1.21, in the CD it goes under EX01_021 now keep looking for 1.21) I personally like the labeling in the CD and suggest that the labeling in the book be changed to match the CD. Then, as I have mentioned before, some exercises go with a numbered table and the tables are given separately. I think that by adding a column for COMMENT, the datasets for exercises should be given in their natural sequence in the CD. I also suggest that in the book, the labeling of the CD be adopted, and larger tables may be pushed to the end of the section (and not the end of the book!). I also suggest that the data in the CD be checked for correctness. One reason for this suggestion is the following: In Exercise 2.61, the data given in the CD has an extra pair which results in regression equations different from those given in the exercise. (The offending pair is the seventh entry in EX02_061 (102,107).)

p. I personally think that there should be no extra manuals for software use and that at this level the students should be given instructions in the book on the relevant procedure for popular software. Creating a book specific software could help in the understanding of the topics if it comes with the book. Though, of course, if you give book-specific software some students could face difficulties in applying their knowledge in the practical world.

q. Finally about the organization of the material. I have already said enough and frankly, it seems rather hard to convince the authors of a book that is now being prepared for its fifth edition that their organization is not efficient. Besides the organization that I would like would mean a major overhaul of the presentation and the authors may not want to take on all that much work, especially if the book is selling as it is. There are some successful books with that kind of organization. Just to complete what I started here is a suggestion for organization: 1. Sources and types of data, including a discussion of discrete and continuous variables. 2. Display of data, 3. Description of data from one variable, 3. Regression etc. 4. Probability, 5. Probability distributions 6. Continuous random variable 7. Sampling distribution, 8. Means and proportions 9. Hypothesis testing, 10 Inference, 11. ANOVA 12. Advanced topics.

r. A word about the advanced topics, either there should be enough so that two courses can be made out of the book or the sale of the book might suffer, that is if the authors are worried about this possibility.

This ends my comments on the book and it would be unfair if I do not tell my readers why I decided to make these comments public. In the following I narrate the events that led to the writing of the first draft of these comments.

Last Fall (2002) I found out that I would be teaching elementary statistics from this new book, “Introduction to the practice of statistics, by Moore and McCabe” I looked into the book and told the chairman, let us call him Professor X, that this book may not be suitable for our students in that it is a hard to read book, that its treatment of some topics was not very clear and that it gave exercises that required the use of computers for which it included no instructions. Professor X came back with a “philosophical” response that startled me. His response was, “I have chosen this book because I am concerned that too many students are passing this course." He also added that he had attended a Chairs' meeting where he had learned about the qualities of this book. This explanation infuriated me, but seeing no choice I had to decide to wait.

As I had expected, a month into the course and other instructors started complaining about the confusion, the lack of doable problems and the pace of the book. Some of them in fact told their students that the book would be changed soon. I confronted the chair again and reminded him of what he had said. Professor X refused to recall what he had said to me and said that he had entrusted the task of selecting the new book to an adjunct who it seems has some idea of statistics. When I asked Professor X as to why he had not consulted other instructors the gentleman came back with, "I only trust his judgment".

Then there was a meeting about the book, which I could not attend. But the report came and in it the chairman did the next best thing. It was said in the report that the book is "intelligently written" so it may be difficult to read. (This to me means that he did say to me what he refuses to recall.) About the lack of exercises doable without technology the gentleman was mum about the suggestion that we should settle for five homework problems from each section!

Other colleagues joined in the defense of the book. One gentleman came up with "Moore and McCabe" has set new standards and that all the other recent books are mere copies of Moore and McCabe. I had to come up with my, "Get real" stance and had to point out to the learned colleague that there are web sites pointing out errors in Moore and McCabe and some of them point to serious flaws in the book. In my search I discovered something that set me thinking. It seems that quite a few research articles reference Moore and McCabe as their source of definitions. This could be because of the authors' association with Moore and McCabe and it could be that some "secret society" is making an effort to promote Moore and McCabe; because of the way it is written. About secret societies I have only one thing to say, an organization decides to be secret because it knows it has an ugly face.

Now why should some secret organization be interested in supporting some books against others? The only reason that I can think of is that a lot of the universities are run as businesses and businesses need buyers. The buyers, in this case, happen to be young minds. If you teach them well they are lost as potential buyers in future. The thought is too ugly for me to dwell on, but there is a possibility.

I am not suggesting that the authors are involved or that really such an ugly thing is going on. What I am saying is that the insistence of some people on using a book that is not suitable for schools that do not have enough technological support may lead one to think that there may be something insincere going on. I trust that everyone thinks that the young students that we teach at various universities, high or low in standing, are our future.

My reason for making this review public: On seeing that my ordinarily sensible colleagues (or shall I say superiors) stonewalled against me on the issue of a book I suggested that perhaps there was a secret society pushing them to keep that book. The next day, literally the next day, I received a request to write a review of the book. Now this is one heck of a coincidence.

Finally, for those who have or who intend to use this book as a "standard reference": A statistician friend of mine has pointed out that a discrete random variable does not necessarily have values from a finite set as indicated in Moore and McCabe (page 306). I was willing to accept the explanation that this book is to do with "Practice of Statistics" and in practice you may not get more than a finite number of values of a random variable, but she was adamant that a definition is a definition is a definition. I have no choice but to agree with her.

Come to think of it, if discrete random variable has only a finite set of values then where would you put the following model? (Here *N* is the set of natural numbers {1, 2, 3,})

X	$n \in N$
P(X)	$\frac{1}{2^n}$

Indeed every convergent geometric series can be used to represent such an infinite model, which is discrete. Here is a simple situation in which this model is applicable: For each $n \in N, P(X = n)$ is the probability of all heads in simultaneously flipping *n* fair coins. Of course in practice you can flip but only a finite number of coins simultaneously, yet no one can put a bound on the *n* someone can get to. In other words the model is legitimate. Indeed it would be instructive to provide physical interpretations to other sequences of positive numbers with convergent sums.

Finally, here is a thought that might occur to someone. There is so much data being bombarded at us and most of it is badly presented so why not put a book in the market that prepares our students for badly prepared articles. Fine with me, but first the book should give them the basics in a very simple and clear language and indicate the possibility that someone could try to cheat by giving a misleading histogram for instance (something that authors forgot to include).