What statistical methods do journal readers need to understand?
Do 75% of radiologists understand fewer statistical articles than the "average" radiologist? .

HANLEY JA: *American Journal of Roentgenology* , 163(3):716-718, 1994.

Introduction

The article by Elster[1] in this issue of the journal reports on a detailed survey of the statistical techniques that were used in the major articles in two mainstream radiology journals in 1993. The main findings, which are similar to those of a survey of articles in the New England Journal of Medicine some 15 years ago[2], are that just under half of the articles used no statistical methods or descriptive statistics only, and that a reader who was familiar with the topics usually covered in a basic introductory course in statistics would have statistical "access" to more than 80% of articles.

This latter figure of 80% should be especially helpful to those who have not taken a formal course in statistics, or took it at a time in their career when it did not seem to be relevant, and who would like to know how much time and effort it would now take to catch up. At the start of a 40-hr graduate biostatistics course I teach each year to a mix of health sciences graduate students, physicians and other health professionals in our 8-week summer school, I use the data from Emerson and Colditz[2], very similar to those in Elster's Table 3. I do so to try to convince local medical residents and staff to stick with the course, in spite of the competing pressures on them and the fact that many of them have had time freed up for only 4 of the 8 weeks. (Incidentally, I have been asked many times to compress the course into 4 weeks but have always refused, arguing that there are too many new, and somewhat unnatural, concepts, to be absorbed in any less time.) Interested readers who are unable to attend formal classroom courses should be helped by the fact that in the last decade or so there has been a welcome increase in the number of good biostatistics texts. Some recent favorites of mine are [3,4,5,6,7] There is also a very interesting video series (Against All Odds, a 26-program telecourse on statistics and its applications, developed by David S Moore, sponsored by The Annenberg Corporation for Public Broadcasting Project. For more information, call 1-800-LEARNER.). I have not used it, but colleagues of mine speak highly of an interactive computer program, "Understanding Biostatistics", aimed at undergraduate medical students (available from Formal Systems, Princeton, NJ). See also Hanley[8] for an earlier annotated list of textbooks and other sources. Some of these are specially designed for self-study.

In this short space, I will comment briefly on one of Elster's findings and offer some personal observations. The general intent of my remarks is to urge those who prepare research reports to use quantitative methods with their readers in mind and to use statistical techniques to enlighten rather than obfuscate. My more specific aim is to emphasize the value of good data summaries, or descriptive statistics, presented in such a way that readers can visualize the raw data behind the reports.

"Descriptive statistics only"

The finding that stuck me most, both in Elster's article, and in the articles reporting usage in other specialty, as well as more general medical, journals was the high percentage of articles that are reported to have "used no statistical methods or *descriptive statistics only.*" Descriptive statistics are concerned with the presentation, summarization and presentation of data, whereas inferential statistics allow us to generalize from our sample of data to a larger "universe". I devoted most of a 1989 commentary[8] to the use, and misuse, of techniques of statistical inference, especially tests of significance. My plea for greater use of the "more natural" confidence intervals has been expanded on by Metz[9]. Therefore, with the few pages allotted to me, I will limit my comments here to the more mundane but I think somewhat neglected, and even disparaged, topic of descriptive, or should I say, "nondescriptive", statistics.

A beginning course in statistics usually devotes less than 10% of the time at the start of the course to the topic of descriptive statistics. The remainder is devoted to such items as standard errors, intervals formed from a mean ± some multiple of the standard error of the mean, or test statistics obtained by dividing an estimate by its standard error, how to look up (or obtain by computer) the multiple in z- or t-tables, and when such reference distributions are and are not indicated. I believe that this imbalance and lack of emphasis are reflected in the non-informative-ness and potential for misinterpretation of many of the descriptive statistics we find in the medical literature, and indeed, in many reports that use quantitative methods.

Purpose of descriptive statistics

The poor choice of summary statistics, and indeed the lack of appreciation of their purpose, is often found in what is Table I in most research reports. In this table, we are given "characteristics of the patients studied," presumably as a way to describe the patients, and to allow us to get to know what they are like, just as if we were being taken to see them on rounds or in clinic. In reports of randomized controlled trials, or indeed of comparisons formed nonexperimentally, there is an additional purpose of showing how similar the two groups of patients, in whom outcomes are to be compared, were with respect to these characteristics at the outset. Many authors still make the mistake of using formal statistical tests and reporting pvalues in these comparisons of baseline variables; I prefer to ask whether the groups are "embarrassingly different" rather than "(statistically) significantly different". In any event, whether describing one group or several, the main intent should be to describe the type of patients studied so that readers can decide whether they are close enough to their own patients that they should be interested in the results. If you were describing some quantitative characteristic of the people you went to high school with, or shared a cruise with, or whatever, you would not always use the mean and you certainly would not use the standard error of the mean. Unfortunately, authors often use the standard error of the mean, which describes the uncertainty in the mean, when the standard deviation, which describes the variation of individuals, is more appropriate, and vice versa[10]. Fortunately, if one is needed but the other was reported, the reader can, with a little knowledge of how they are related, derive one from the other.

Does the average conceal more than it reveals?

Among descriptive statistics, the venerable average or mean must surely be the most overused descriptive statistic. Reading Elster's use of the term "an average physician" I am reminded of what Galton said about averages: "It is difficult to understand why statisticians commonly limit their enquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull as to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that 'If its mountains could be thrown into its lakes, two nuisances would be got rid of at once'." I am also reminded of a statistic that a urologist recently related to me: that the average human has one testicle and one ovary (I am not sure what the statistics are for radiologists). Here are situations where neither the mean nor the median gives us the correct picture. In these cases, even the only other measure of central tendency, the mode, will not suffice; we need to use two modes. Fortunately, in this situation we already have sufficient independent data on variation of landscapes and humans that the poor summaries did not mislead us. But what if we were summarizing information on some newly studied characteristic?

Which do you mean by "average"?

Part of the difficulty in the use of descriptive statistics is the mixing of the lay and scientific use of terms like average, typical and middlemost. These lead to the kind of apparent paradox in the subtitle of my commentary, which I adapted from the following quote from a British newspaper "The usually wonderful television commentator, introducing a Newsnight discussion last Friday on the teaching of reading skills, expressed dismay that 'a third of our primary schoolchildren have below-average reading ability'. Had he paid more attention in his 'rithmetic lessons, perhaps he would have realised that half our schoolchildren are below average in everything. As, indeed, are half our Newsnight presenters." We would all be less confused if data were reported in quantiles, or displayed graphically, using such techniques as "dot diagrams" and boxplots[11].

Although it belongs under inferential rather than descriptive statistics, the word significant creates similar problems of semantics. It is interesting that if a 100% survey that compared average (or median or whatever measure of central tendency you prefer) annual incomes of radiologists and physicians in another medical specialty found a $1000 difference, it would be dismissed as trivial. Yet if the same comparison were made in large random samples from these two specialties, the same difference could easily be labeled "statistically significant". Readers should appreciate that a trivial difference would be termed "statistically significant" if the sample sizes on which it is based were very large. It might help if readers replaced the words "is statistically significant" with "provides evidence that the difference, in the universe these samples came from, is non-zero"

What is "standard" about the standard deviation?

If the mean is an overly used descriptive statistic, then the standard deviation (SD) must be a close second. The rationale is probably that if it is difficult and costly to calculate, it must be worthwhile. But how much of a picture do you have of the variation in the length of a procedure or hospital stay if it is reported that the mean ± SD is 10 ± 15? Not much, unless you are quite adept at fitting distributions to means (and even then there are a lot of shapes of distributions that yield the same mean and SD) and unless you already know quite a bit about the pattern of variation of the characteristic in question.

And if someone calculates a SD, can the "Normal" distribution be far behind?

After all, as the mathematician Poincaré said about it, "everyone believes in it [the gaussian or "bell curve" law of variation] however, for the experimenters fancy that it is a theorem in mathematics and the mathematicians that it is an experimental fact." Incidentally, whenever I can, I use the word gaussian rather than normal, in order to avoid the many other meanings of this latter term: in medicine, we speak of normal and abnormal, upper limit of normal, and so on; meteorologists tell us that the day's temperature will be so many degrees above or below normal. Variability of most characteristics of individuals is far from gaussian and so the SD is of limited use in reconstructing the pattern of individual variation. To go back to our example of "normal" anatomy, it is also (approximately) true that in humans the SD is one testicle and one ovary, but I hope we would not conclude, by adding two SDs to the mean, that some 2.5% of persons have more than the 3 of each. Worse still, what about somebody whose values are in the lowest 2.5% of the population? Moreover, contrary to a widely held view, the interindividual variation does not get any more gaussian and the SD does not get decidedly bigger (or smaller), if we observe a bigger sample of humans.

Why these misunderstandings about some basic statistical ideas?

They stem in part from the large portion of an introductory biostatistics course that is devoted to inferential statistics. When we are taught about the sampling distributions (t, F, chi-square, etc.) for the behavior of means, proportions, slopes, correlations and so on, considerable emphasis is often placed on checking that the assumptions, such as gaussian-ness, under which these reference distributions were derived, are fulfilled in our data. However, it is somewhat paradoxical that in the situations where the assumptions matter most (i.e., in small samples), we do not have enough data to assess whether they could have reasonably come from a gaussian distribution; when we do have enough data to be able to assess whether the gaussian distribution holds, we actually do not need to worry about whether it does or not. Many authors forget that if sample sizes are substantial and if the skewness or other non-gaussian-ness in observations on individuals is not overwhelming, the Central Limit Theorem is a strong "gaussianizer" of the sampling variability of statistics such as means, proportions, slopes, and the like. Instead, even with sample sizes in the hundreds, we see statements that "because the data did not show a gaussian distribution, we used nonparametric statistics." although the decision to use such alternatives may have good reasons to commend it, the absence of Gaussian-ness in and of itself should not be taken as an indication that inferences based on the t or z distribution would be inaccurate. Unfortunately, this pre-occupation with gaussian-ness as a prerequisite for validity of statistical procedures carries over into regression methods. Paradoxically, authors will go to great lengths to get around the fact that a measured 'x' variable in a regression does not have a gaussian distribution, while they are quite willing to directly include a two-point (binary) 'x' variable, such as those we discussed earlier, as is. Likewise, those new to the use of regression methods to describe, say, the relationship between some anthropometric variable and age and sex will painstakingly check that the distribution of this variable is gaussian in the aggregated ages and sexes, and they will be worried when they find that it is not. In fact, what the usual regression techniques require for accurate inference is that the variation within each age-sex category be gaussian, and even this is not critical for certain inferences in certain situations.

Concluding remarks:

Elster has reassured those who are eager to catch up on their knowledge of statistical methods that the essential curriculum is not too large. However, I suggest that they spend some additional time, possibly after the course is completed, to go back again to those chapters on descriptive statistics that were covered so quickly. Also, now that statistical packages with easy–to–use graphical data analysis methods are common, readers should expect informative descriptive displays and summaries. With these, and with the help of the increasing learning sources available, they should be in a better position to visualize the data being reported.

REFERENCES

[1] Elster AD. Use of statistical analysis in the AJR and Radiology: Frequency, methods, and subspecialty differences. AJR 1994, 163:

[2] Emerson JD, Colditz GA. Use of statistical analysis in the New England Journal of Medicine. N Engl J Med 1983;309:709-713

[3] Dawson-Saunders B, Trapp RG. Basic and Clinical Biostatistics. 2nd Edition. Appleton and Lange, Norwalk Conn. 1994

[4] Freedman D, Pisani R, Purves R, Adhikari A. Statistics, 2nd Edition. Norton, New York, 1991

[5] Knapp RG, Miller MC. Clinical epidemiology and biostatistics. NMS, Wilkins and Wilkins, Baltimore 1992.

[6] Moore DS and McCabe GP. Introduction to the practice of statistics. Freeman, New York, 1993

[7] Norman GR and Streiner DL. Biostatistics: the bare essentials. Mosby, St. Louis, 1994

[8] Hanley JA. The place of statistical methods in radiology (and in the bigger picture). Invest Radiol 1989; 24:12-16.

[9] Metz CE. Quantification of failure to demonstrate statistical significance. The usefulness of confidence intervals. Investigative Radiology 1993;28(1):59-63

[10] Brown GW. Standard deviation, standard error: which standard should we use? Am J Dis Child 1982; 136:937-941

[11] Moses LE. Graphical methods in statistical analysis. Annual Review of Public Health 1987; 8:309-353