On Teaching the Undergraduate Statistics

Gregory L. Light

Department of Management, Providence College
Providence, Rhode Island 02918, USA
glight@providence.edu

Copyright © 2013 Gregory L. Light. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

This paper resulted from a conference presentation by the author on the pedagogy for a first course in undergraduate statistics. We identified the major obstacle for students to learn this subject to have been being the perpetually cumbersome and incoherent presentations of the course materials permeating in just about all textbooks. Here we streamline the logical structure by clearly separating population/probability descriptive parameters from sampling statistics; in addition, we use binary variables to eliminate the need of studying inferences about population proportions, testing between two population means, and ANOVA. Otherwise we present some of our insight into the subject matter, such as the geometric invariances of the mean, the standard deviation, and the standardized z scores, and we also share our classroom experiences, such as the necessity of doing drills on certain foundational concepts.

Mathematics Subject Classification: 97K10, 97K40, 62-01

Keywords: Statistics pedagogy, standardized z, statistical geometry, binary dummy variables, anova

1 Introduction

The subject of Statistics has a reputation of being difficult to learn and thus to teach. People in the academic profession through their years of successful practices of statistics in their own fields tend to forget how little they had learned in their initial encounter of this subject as undergraduate students,

and as a result mistakenly attribute their later acquired mastery of the subject nevertheless to the flawed pedagogical structure that has been perpetually permeating in introductory statistics textbooks. This paper presents our view on how a first course in Statistics ought to be presented, as encouraged by the favorable responses received in a conference presentation ([3]).

The overwhelming backdrop of today's educational environment is information technology (*IT*), offering an unprecedented supply of data and readily available computational packages (cf. e.g., [4], for this prevailing trend and its impact on pedagogies). The fundamental drawback of such an orientation is that it tends to deprive people of thinking on their own. As an illustration, presently when students see an equation like $y = mx + b$, their impulses are to plug values into m , x , and b to compute y . Such a tendency easily deters their study of the topic of regression, wherein the focus is on estimating the values of m and b . That is, students today have mostly stopped appreciating an *equation as a whole*, thus the notion of cause and effect. Worse, a question like whether the meaning of $mx + b$ is $m(x + b)$ may exist in some students' minds. The foundation of Statistics is inductive thinking, to infer about certain population parameters from a small sample, thus the t-distribution, among others. Flooding data to students is as a matter of fact contrary to the intent of statistical analysis.

Secondly, while people take it for granted the importance and necessity of doing drill exercises in, say, learning a sport, music instrument, driving a vehicle, ..., they do not see the need for doing drills in learning Statistics. The "task-directed or contents-based" approach of enticing students to learn by tailoring the course materials into some "real-world" situations (see, e.g., [1]) fundamentally undermines the need for basic computation drills, e.g., that of the standard deviation; furthermore, the true real-world situations are inevitably individually varied and innumerable. Most students in their lifetimes take Statistics just once; to teach them using statistical *techniques* to solve problems on a task-by-task basis in the name of preparing them for their future workplaces is at best unrealistic. In this regard, there has been this great mystery that when the topic comes to teaching arguably one of the most technical subjects in undergraduate studies, *Statistics*, just about everyone knows how, by basically advocating a compression of his/her own post-graduate multi-year practical applications of the subject into a one-semester college introductory course, and in this connection some textbooks actually contain blatant serious mistakes, such as presenting the distribution of sample-means with the normal distribution centered at $\bar{X}$, or marking $z = 1$ under the normal distribution without regard to $P(z \in [0, 1]) = 0.3413$.

In the following, we will present our treatments of the subject matter in Section 2, and make a concluding summary in Section 3.

2 Contents and Treatments of the Undergraduate Statistics

Foremost in our view, the general governing themes in Statistics are two: (1) "Things" (intentionally broadly expressed so as to cover the population values of a variable x , $\bar{x}$, or an estimator of a coefficient in an equation) vary around a center. (2) They, i.e., these "things," are *Easily Off* the above center by a certain quantity (*standard deviation* σ , or standard errors correspondingly). They vary by pure random chances and/or cause(s) (regression). Here it can never be overemphasized that to undergraduates just emerging from their previous deterministic learning environment such as from algebra and calculus, the idea of standard deviation is actually *new*. We the instructors might want to recall how we ourselves once felt when we were learning this concept for the first time. An abundant amount of drills should be assigned for the calculation of σ . Only through a thorough grasp of σ can a student be comfortable with an expression like "how many standard deviations is x from the mean." In this regard, the fundamental idea of the standardized z scores should not be exclusively cast in the context of the normal distribution. Consider a population containing only two distinct numbers; then the larger one must have $z = 1$, with probability of occurrence 0.5. In this connection, for consistency and uniformity the Chebyshev's inequality, instead of using " k " as the common textbook symbol, should simply be presented as

$$P(|z| \leq z_0) \geq 1 - \frac{1}{z_0^2}. \quad (1)$$

Secondly textbooks ought to have a clear dichotomy between "descriptive statistics" and "inferential statistics." In the former topic, one describes a population, not a sample; that is, all the references about sample statistics should be removed entirely from those beginning chapters of population parameters. A sample is intended only to infer something about the population; one is never interested in the sample *per se*, for changing one sample to another would bring about a whole new set of statistics. To be absolutely clear, if one is really interested in the very contents of a so-called sample, then by definition that "sample" becomes his/her momentary population. Also, population and probability studies should be merged as one. Conceptually, drawing an x from a population randomly makes x a random variable. Computationally, any mean (including variance) can be calculated via multiplication by weights ($\equiv$ relative frequencies $\equiv$ probabilities). In this way, μ and σ of a population or random variable can be covered in an integrated manner. In the latter topic, sample statistics ought to be presented as estimation schemes. For example, $\bar{X}$ is a random variable, a formula, a way to estimate the population mean

μ , verbalized as "given a random sample of values, we take the average of these values," in total analogy with "given a set of scattered points, we apply the least-squares procedure to fit them." In short, an estimator is not a number, even though when applied to a sample the estimator does produce a number. As such, while it is "OK" to express $\bar{X}$ as "the mean of a sample," it would definitely not be advisable to express s as "the standard deviation of a sample." In our pedagogy, we use hyphenated "sample-mean" for $\bar{x}$ and "sample-standard-deviation" for s , to underscore the critical distinction between population (Greek) parameters and sample (Roman) statistics. By the above perspective, it would not make too much sense to calculate the z score of a value in a sample, as in $z_1 = (x_1 - \bar{x})/s$, for the supposed meaning would then be: "This value x_1 is *estimated* to be $z_1 s$ away from μ in the population." In short, we strongly disagree at the "side-by-side" presentation of population parameters with sample statistics in beginning chapters. To express our view cogently one more time, consider "range;" then one would have population range and sample range, but sample range apparently cannot be an estimator of population range.

Next, to students majoring in mathematics and science, the geometric nature of μ and σ ought to be introduced. Consider two distinct values $A < B$ separated by $(m + n)$ units, $m, n \in \mathbb{N}$; then by setting the frequencies of A and B respectively equal to n and m we achieve $\mu = A + m \equiv B - n$, with the familiar interpretation of μ being the center of mass, or an economic interpretation of cross-subsidization: "those richer than μ subsidize mn to those poorer than μ and all the $(m + n)$ people land μ , or $\sum_i (x_i - \mu) = 0$. To continue, if we enlarge the population by adding $nm^2 + mn^2 - (m + n)$ observations of value μ , then we achieve

$$\sigma = \sqrt{\frac{nm^2 + mn^2}{(m + n) + nm^2 + mn^2 - (m + n)}} = 1. \quad (2)$$

As an illustration, consider a random variable X with probability distribution P :

X	P	
10	1/6	
30	3/6	
40	2/6	(3)

Then $A = 10$, $B = 40$, $m = 2$, $n = 1$, $\mu = 10 + 2 \times 10 = 40 - 1 \times 10 = 30$, $nm^2 + mn^2 - (m + n) = (mn - 1)(m + n) = 3$, and $\sigma = 1 \times 10 = 10$, so that $z = 1$ for B and $z = -2$ for A . I.e., μ , σ , and z are geometric invariants. In particular, z is translation and scale invariant, which may be expressed as

"uniform pay raises" and/or "changing currencies" do not affect one's z number of pay. As another demonstration, we show how an instructor can manufacture a standard error $(s/\sqrt{n}) \in \mathbb{N}$ for classroom teaching conveniences. Consider a sample of size $n \geq 2$, containing value A of frequency $= 1$ and B of frequency $= (n - 1)$, with $B - A = nl$, $l \in \mathbb{N}$; then $\bar{x} = A + (n - 1)l$ and

$$\frac{s}{\sqrt{n}} = \sqrt{\frac{(n-1)^2 l^2 + (n-1)l^2}{n(n-1)}} = l. \quad (4)$$

In general, students are best served when they can appreciate the significance of unit-free or dimensionless values, such as z , t , etc. In this connection, while variance σ^2 serves some theoretical purposes, $cov(x, y)$, with obscure product units, ought to be replaced by correlation coefficient.

Lastly we collect some miscellaneous points of remark: (1) Continuous distributions ought to be interpreted as smoothed-out histograms to emphasize that the meaning of an area under a curve is nothing but a "head count" except expressed in terms of a percentage of the total population. (2) The expression "sampling distribution of the mean" ought to be replaced with "the probability distribution of the sample-means" for simplicity and transparency. (3) The use of binary (dummy) variable $x = 0$ or 1 ought to replace inferences about population proportions. True, while the two approaches agree at $\mu = p$ and $\bar{x} = \hat{p}$, the estimated standard error by our suggested binary-variable approach is greater than that by the conventional treatment, i.e.,

$$\begin{aligned} \frac{s}{\sqrt{n}} &= \sqrt{\frac{(1-\bar{x})^2 n\bar{x} + (0-\bar{x})^2 n(1-\bar{x})}{n-1}} \\ &= \sqrt{\frac{(1-\hat{p})\hat{p}}{n-1}} > \sqrt{\frac{(1-\hat{p})\hat{p}}{n}}, \end{aligned} \quad (5)$$

but it does not make sense to devote a whole chapter as in many textbooks just to obtaining a smaller standard error when the difference between the two approaches is as small as indicated in the above equation (5) and when the context is statistical inference. Similarly, the topic of testing two population means ought to be dealt with by a regression via a binary variable with a possible adjustment for homoskedasticity; in the same vein, ANOVA ought to be dealt with by a regression against some binary variable(s) (see [2], pp. 413-419, for a demonstration of the superiority of the binary-variable approach over ANOVA). (4) In hypothesis testing, a parameter value as presented in the null hypothesis ought to be clearly distinguished from the true population value. Here in class we emphasize the existence of "3 identities" in a hypothesis testing, that is, the true value, the default value, and the estimated value. (5)

It is always regrettable that in a one-semester course the time left for the topic of regression is so minimal when the statistical inferences are about coefficients in equations, of much more practical importance. Here as we alluded to in the beginning, many students at present have become so number-oriented that they have lost their appreciation of the concept of an equation, when supposedly higher education is to cultivate their abilities to pursue causal relationships.

3 Conclusion

There is a fine line between motivating/engaging students to learn and enabling/facilitating their declining attention spans due to the ever-present digital distractions. Basic drills such as writing out the squared-deviations of individual values from μ , averaging them to get σ^2 and taking the square-root to arrive at σ are indispensable. There does exist a linear chain of reasoning in Statistics: Only by acquiring σ as an instant vocabulary can a student make use of the z values and then later the even more ubiquitous t values in statistical inferences. Replacing a clear conceptual and structural understanding of the subject with a plethora of data, terminologies, and formulas in an ostensibly applied setting is to cause a student in the end unable to verbalize something like $t = -1.3$ into plain words: "The estimate in question is 1.3 standard errors below the null-hypothesis claimed value." We need keep in mind that most of our students take Statistics just once in their lifetimes. For their life-long self-learning precisely due to the ever changing environment of *IT*, a good basic understanding of the subject is much more valuable than a bag of myriad isolated tools. Depriving young students of the opportunity to expand their reasoning capacities undermines future democracy. We thus present this paper with the hope that our simplified structure for a first course in undergraduate statistics will help students to achieve a genuine understanding and appreciation of the subject.

References

- [1] J. Garfield and D. Ben-Zvi, How students learn statistics revisited: a current review of research on teaching and learning statistics, *International Statistical Review*, 75 No. 3 (2007), 372-396.
- [2] J. Kmenta, *Elements of Econometrics*, Macmillan, New York, 1971.
- [3] G.L. Light, On teaching the undergraduate statistics, Hawaii University International Conferences, June 10-12, 2013.

[4] S. Tishkovskaya and G.A. Lancaster, Statistical education in the 21st century: a review of challenges, teaching innovations and strategies for reform, *Journal of Statistics Education*, 20 No. 2 (2012), 1-56.

Received: July 19, 2013