2020

Statistics: A Cautionary Tale

Len Zane

Statistics: A Cautionary Tale

Len Zane
University of Las Vegas, Nevada

Abstract: Many of the numbers used to assess students are statistical in nature. The theoretical context underlying the production of a typical number or statistic used in student assessment is presented. The author urges readers to recognize objective data as subjective information and to carefully consider the numbers that often determine admission, retention, and scholarship distribution in honors.

Keywords: data-based decision making in education; student assessment; standard deviations; probability distribution; central limits theorem

PROLOGUE

For many years, I would go through the following series of steps as part of my responsibility as a professor of physics. Administer an exam to a class of approximately fifty students, grade the exam, calculate the average and standard deviation, plot the grades in a histogram, and label approximate letter grades for that exam on the histogram.

Periodically, I would mull over what the test statistics—the average and standard deviation—meant in the context of the test-taking scenario. The mulling was caused by an uneasiness about the implied connection between these statistics and the histogram of the test scores. I never managed to clearly identify the root cause of my uneasiness, but the sense of discomfort remained.

In the fall of 1985, I became the founding director of the honors program at the University of Nevada, Las Vegas (UNLV). The importance of numbers and their averages became more central in this new role: first, the high school GPAs and class standings and SAT and ACT scores of students

applying for admission; then, once students were enrolled in honors, the GPA requirements for students to remain in good standing or to receive and retain academic scholarships. All these numbers were fraught with uncertainties that I could not put my finger on but that made me nervous nonetheless.

Since becoming an emeritus professor, I have had time to think more seriously about statistics and how they often get misapplied in analyzing sets of numbers, each of which has a range of possible values that are often ignored and instead replaced by a single, inviolate number. "Misapplied" is probably too strong an indictment; more accurately, there is a strong tendency to give too much credence to the various numbers earned by students and less thought to the range of values surrounding each of these numbers.

At this point, I should explain what this essay is and what it is not. It is not a typical research article. It contains no data collected from students at my university. It does not refer to other honors publications that have explored similar questions. Rather than any of those usual essays, it is an attempt to look at the underbelly of statistics in order to make readers less confident and more skeptical of many of the numbers used to evaluate students enrolled in honors programs and colleges.

The first section defines the elements of statistics that will be central to the later arguments. The primary reason for inserting these definitions at the beginning is to level the playing field, as much as possible, for readers with different backgrounds in statistics. Also, readers with a background in statistics may find some of the definitions idiosyncratic. Therefore, please do not skip this section.

As an example of the "cloudiness" of a number, the main part of this essay—sections II, III, and IV—looks at the results of a classroom exam and how the preciseness of the scores invites more analysis than is warranted. Other numbers that abound in higher education could easily have served as the example. Sections V and VI discuss the size of a representative sample and the meaning of an individual SAT score.

I. BASIC DEFINITIONS

The following terms that appear in the body of the article are defined as nontechnically as possible.

Average

Take a bunch of values, say 25 of them. Add them together and divide the sum by the number of values, in this case, 25. This is the average; it tells

you something, but less than you may think, about the group of numbers. For example, if all 25 numbers are 10, the average is obviously 10. On the other hand, if 10 of the numbers are 0, 5 of the numbers are 10, and 10 of the 25 numbers are 20, the average of the 25 numbers is also 10. Clearly these two sets of numbers are different, but their averages are the same.

The two examples show how the average says nothing about the spread of the original bunch of numbers that were used to form the average. The variance, defined next, is a number that tells you something about that spread—at least how the numbers spread around the average.

Variance

Subtract the average from each of the 25 numbers, square each difference, add them, and divide by 25. This is the variance, which is a useful measure of the spread of the numbers that were averaged. In my first example above of an "average," the variance was zero since all 25 numbers were equal to the average, 10. The second example also had an average of 10, but the variance was 80. This relatively large value alerts you to a lot of "scatter" in the second set of numbers.

Standard Deviation (SD)

The standard deviation is just the square root of the variance. For the two examples with an average of 10, the SDs are 0 and 8.94, respectively.

Probability Distribution (PD)

A statistical event, for example tossing a die, has an underlying theoretical set of possible outcomes, each with a well-defined probability. For the die, each of the outcomes of a single toss has an equal probability. An analog that will be useful later is to replace the die with a spinner centered on a circular pie graph divided into six equal sectors. Since a circle has 360°, each of the sectors representing one side of the die encompasses 60°. A toss of the die is statistically equivalent to a single twirl of the spinner. The PD represented by the spinner and the associated pie-shaped graph can be generalized much more easily than the die.

The PD for a die has six possible outcomes, 1 through 6, each with an equal probability. The mean value for this unbiased die is just the average of these six values, 3.5. Reserving the word "mean" for the PD average will distinguish PD averages from statistical averages. The variance for the PD in the die example is 2.92.

In the case of the die, we can comfortably assume that each outcome (each side) has an equal probability, but imagine that some nefarious character has replaced our unbiased die with a biased (loaded) die that will land disproportionately on one of the six possible values. This bias could only be determined by tossing the die many times and keeping track of the outcomes. The point here is that even with something as simple as a die, we cannot know, *a priori*, the probabilities that ought to be associated with each face of the die.

For a better example, suppose six horses are entered in a race. The probability that a particular horse will win the race is represented by an appropriately sized sector in a pie graph. The angular spread of the six sectors has to add up to 360° since one of the six horses has to win the race. Twirl the spinner, and the winning horse is selected by the sector the spinner lands on, saving a lot of wear and tear on the horses!

Now it is time to introduce the big CAVEAT. In most cases of interest, the underlying PD, which is to say the sizes of the different sectors (e.g., the probability that one horse will win), is unknown and is in principle unknowable; it is, in fact, what we are trying to unmask by collecting data.

Statistical Distribution

Data are collected by sampling the PD. In the example of the horse race—remember that the actual sizes of the sectors in the pie-shaped graph are unknown—imagine twirling the spinner 100 times; this is equivalent to having the horses rerun the same race under exactly the same conditions 100 times—an impossibility!

The data collected by sampling the PD are used to construct the statistical distribution. Horse A won 5 times, horse B 13 times, etc. The outcome of the sampling is that horse A has an approximately 5% chance of winning while horse B has a 13% chance, etc. (Of course, the percentages have to add to 100%.) Even these percentages are approximate: if we had the same horses run the 100 races again, the outcome would likely change. Maybe horse A would have a 7% chance of winning the second time around.

The game of statistics uses the data collected by sampling the PD, twirling the spinner, to learn as much as possible about the unknown underlying probability distribution. Intuitively, it makes sense to think that the more times the spinner is twirled, the better chance we have of getting a truer estimate for the PD, but regardless of the number of samples taken, the result will always be an *estimate* of the PD.

Normal Distribution (ND)

When people picture a "bell-shaped" distribution, they are picturing a normal distribution. If data conform to a normal distribution, the following quantitative facts are true: the peak of ND is at the average value of the data, and the spread of the data is determined by the SD; in selecting a single sample from a statistic described by the ND (twirling the spinner once.), 68% of the time the selected value will lie within one SD of the average, 95% of the time it will lie within two SDs, and 99.7% of the time it will lie within three SDs of the average.

For example, if the average is 10 and the SD is 2, then there is a 68% chance that the twirl of the spinner will land on a value between 8 and 12 and a 95% chance that the spinner will land on a value between 6 and 14. The chance of the spinner pointing to a value outside the range 4 to 16 is just 0.3%.

Central Limits Theorem (CLT)

This theorem suggests that almost any statistical set of data can be approximated by a normal distribution. Suppose 25 samples are taken from 25 PDs, one from each PD. The 25 PDs could be the same, similar, or different. The sample consists of one twirl each of the 25 spinners representing the 25 PDs. Note that each of these 25 PDs will have a mean and a variance, values that are typically unknown.

The theorem says that the average of the 25 values determined by the spinners will have an approximate ND, centered on the average of these 25 PD means. The variance of that ND equals the average of the 25 PD variances divided by 25. Remember, usually these 25 means and variances are unknown, and it is impossible to find the averages of 25 unknown means and variances.

The phrase "divided by 25" in the previous paragraph is the quantitative statement of the fact that our approximate ND gets better as the sample size is increased. The width of the ND is determined by the SD, the square root of the variance, which is inversely proportional to the sample size. Therefore, the bell-shaped curve gets narrower and narrower as the size of the sample increases.

The CLT turns out to be true for a surprisingly wide range of different, very un-bell-like underlying and unknown PDs used in the above example; in a real sense, it is a statement about the power of averaging. The CLT speaks to the outcome of the average of the 25 twirls. Each of the 25 spinners is twirled

only once. With only a single sample from each of the 25 PDs, nothing useful can be said about the mean or variance of the 25 possibly different PDs based on a single twirl of each. In point of fact, the variance is undefined for a data set consisting of a single value.

At the risk of being overly repetitive, I stress that the CLT predicts that the average of a single spin of each of the 25 different spinners ought to be statistically distributed as an approximate ND. But, and this is a big but, the center and variance of that predicted ND are given by the average and variance of the 25 means and variances of the unknown PDs.

The utility of the 25 values collected depends on the relationship or lack thereof of the 25 PDs to one another. For example, suppose the PDs represented the probability for the outcome of the six-horse race described earlier. In this case, the 25 twirls refer to the same PD, so the outcome of running the race 25 times could supply useful information about the approximate sizes of the six sectors in the PD describing the outcome of the horse race.

On the other hand, if each of the 25 PDs has a unique distribution of sectors, it is impossible to use these 25 values to estimate the mean and variance of the 25 individual PDs that determine the peak and width of the ND predicted by the CLT.

II. THE TEST SCENARIO

Since grades, both in high school and college, play such a central role in honors and student self-esteem, the following is a detailed look at the data collected when an exam is administered to a group of students. This analysis depends heavily on the definitions presented in the previous section. Keep in mind that the same careful deconstruction could be applied to numbers that often determine admission, retention, and scholarship distribution in honors.

Professor Q gives a test to a class of 25 students. The students are the usual heterogeneous group with an array of study habits, different levels of interest and aptitude for the material, a bewildering range of living situations, and so on. The test design and questions that Professor Q creates will affect different students in different ways. Consequently, on the day of the test, each student enters the room with a different probability distribution. The sectors on each of their 25 pie-shaped graphs (with one sector representing an A, one a B, etc.) have individualized probabilities for the range of possible outcomes on Professor Q's test. Keep in mind that neither Professor Q nor the students themselves (nor anyone else for that matter) know how the sectors are divided on each student's spinner.

Furthermore, there is no reason to believe that all these PDs are approximately bell-shaped with different means and variances. For example, imagine a highly motivated student who suffers from test anxiety: sometimes this student gets through a test with no problem and does well, but at other times the student becomes anxious and incapable of answering the simplest questions. The PD for this student would likely have two peaks: one centered on a high score and the other centered at a much lower score that reflects the impact of an anxiety attack.

In the scenario of Professor Q's exam, each student's score represents a single twirl of his or her spinner: twenty-five spins, 25 test scores.

III. THE ANALYSIS OF THE TEST

The resulting data are 25 test scores, substantial-looking numbers that can be used to calculate an average, variance, and standard deviation. Professor Q cannot resist this temptation. The 25 numbers ask to be averaged; the SD is just waiting to be calculated. Once the numbers are calculated, Professor Q feels an obligation to go further and use these numbers, which can be calculated to hundredths of a percentage, to analyze the result of the exam. Having found the average and SD, the slippery slope of statistical sloppiness lies a short step ahead. Professor Q has a vague memory of a theorem from statistics that essentially says that most data can be explained by the ubiquitous bell-shaped curve, namely the normal distribution.

Professor Q uses the average and standard deviation calculated from the 25 test scores to create an ND based on these values and then superimposes the ND on the histogram of actual test scores. The comparison of the real test results to the ND implies that the test results ought to look bell-ish, but almost invariably the histogram and the ND are embarrassingly dissimilar. This discrepancy ought not be surprising since the bell-shaped curve defined by the ND based on the average and SD of the 25 test scores implies nothing about the distribution of scores shown by the histogram.

Professor Q has endowed the 25 test scores with significantly more meaning than statistics warrants.

IV. THE FOLLY OF THE ANALYSIS

When considered carefully, none of Professor Q's analysis, though reasonable sounding, makes the slightest bit of sense. Obviously, given 25 test scores, the average, variance, and SD can be calculated. Professor Q ought

to have stopped there. When Professor Q posted the histogram of the test results, the average and SD could have been included as fodder for the students to mull over.

If a student was lucky on the day of the test, his or her spinner would have landed on a score higher than what was typical for that student (higher than the mean for their particular PD). Conversely, maybe another student was coming down with the flu on the day of Professor Q's test, so this student's pie graph would look different that day. Lower scores would have a higher probability than they normally would have had for that student, so the spinner would have had more chance to land on a score that was lower than the mean for that student *sans* the flu. Clearly, this single test score says almost nothing about the range of scores available to that student or for any of the other students for that matter.

The prime folly committed by Professor Q, probably without even realizing it, is assuming that each student twirls a spinner with an identical, or at the very least, a similar range of possible outcomes. This scenario pictures 25 pie-shaped graphs with approximately the same outcome profile for each of the 25 students. The result of the test according to this scenario was 25 samples of these similar probability distributions.

This fictitious view gives statistical meaning to the average, variance, and SD of the 25 scores. These values now represent estimates of the mean and variance of the single PD from which all of the 25 samples (test scores) were drawn. Keep in mind that under the erroneous assumption that there is a single underlying PD, the average and variance are estimates for the mean and variance of that PD, but there is no reason to add the further assumption that the underlying PD ought to be approximated by a normal distribution.

Professor Q compounds his folly by assuming that the average and SD of the 25 test scores define an ND that approximates each of the 25 individual PDs that defined the possible outcomes for the 25 students. This erroneous interpretation allows Professor Q to deduce that an individual test score ought to have a 68% chance of being within one SD of the average score. Besides the terrible mistake of conflating the 25 different PDs into some mysterious average PD, Professor Q is also completely misinterpreting the Central Limits Theorem.

Remember that the CLT says nothing about how a single test score ought to be distributed. Instead it says that if you knew the mean and variance of the 25 individual PDs, these numbers would define an ND that would be the approximate PD for the average value of the 25 tests.

Imagine 1000 identical universes where Professor Q gives identical exams to a class consisting of the same 25 students. The result will be 1000 values for the average of 25 test scores. The shape of the histogram of these 1000 scores is the thing predicted by the Central Limits Theorem.

If we knew the mean and variance of the 25 individual PDs, which we don't and can't, we could draw the normal curve predicted by the CLT. That curve gives the probability for the different possible average scores for that class of 25 students. The CLT says that the average score on Professor Q's test has a 68% chance of being within one SD of the average of the 25 means and a 95% chance of being within two SDs of that average. Even in this imagined but impossible best-case scenario, the test data say little about the meanings of the 25 individual scores. Professor Q and the students can compare their scores to the average, but it is not possible to say whether these differences are statistically significant.

Assessing an individual student on the outcome of a single exam is at best iffy. Luckily, in a college class we rarely have to base decisions on a single score. A student takes many different classes, graded in different ways. We can take some comfort in imagining that each student is twirling a spinner with a similar set of outcomes for each of these graded activities. Over time, these various outcomes can be used more confidently to assess the quality of the student's education. We do not have the same luxury in assessing ACT or GRE scores when we use them as criteria for admission to an honors program because these scores are often one-time events.

V. CATS AND DOGS AND THE SAT

Clearly, though, there are times when test statistics are meaningful. In cases where the statistics make sense, the argument is that the group of students taking the test is a representative sample.

Here is a thought experiment to shine light on the notion of a "representative sample." Suppose you and a team of helpers weighed groups of ten dogs and ten cats, and the average weight of each group of ten was plotted on a graph. The results will most likely bear out the fact that the weight of pet cats has a much smaller range than that of pet dogs.[1] In fact, any single average weight of ten cats may not be a bad representation for the average weight of any random sample of ten cats. On the other hand, it is extremely unlikely that the average weight of ten random dogs could act as a predictor for the average weight of ten different random dogs.

It seems intuitively obvious that if we increase the number of cats and dogs in each group, to a 100 for example, the average weight of a single group of 100 cats or dogs is more likely to be representative of any other random group of 100. At some point, there are enough cats and dogs in each group that our confidence in the average weight of the random cats or dogs as representing the average weight of any other equal-sized random group is high. The point is that the random nature of individual weight ranges for each cat or dog in the group being averaged can be smoothed out by including enough cats or dogs. Figuring out the actual number of cats or dogs it takes to be "enough" is not so obvious, even if it is obvious that such a number, the size of a representative sample, exists.[2]

The same is true for students taking a test. In a given year, between 5 and 10 million students take the SAT. It would be surprising if the results in 2017 looked decidedly different from the results in 2018. Five to 10 million students are a representative sample.

Back to Professor Q: Can Professor Q argue that his class of 25 students is a representative sample? Ignoring the fact that anyone can argue anything, 25 seems much too few to be representative, so when does the number of students taking a test become a representative group? That is a question for someone cleverer than me to answer.[3] I am confident that 25 is too few and that a million is more than enough.

VI. AN EXAMPLE OF TWO SAT SCORES

The histogram representing the millions of scores on the EBRW (Evidenced-Based Reading and Writing) or mathematics part of the SAT will be well approximated by an ND centered at 500 with an SD of 100. The scores, in increments of 10 from a low of 200 to a maximum of 800, are scaled to fit that normal distribution. If asked to guess the score on the EBRW or mathematics part of the SAT for a random one of these millions of students, the statistically best guess is 500. In all my years in higher education, no one has ever asked me to make such a guess!

On the other hand, important decisions are often based on a comparison of an SAT score of 630 for student A versus 650 for student B. Two statements can be made about these SAT scores. First, 650 is 20 points higher than 630. Second, the fact that the average score and SD for the millions of SAT test takers are 500 and 100, respectively, has nothing to do with the relative value of the scores achieved by students A and B.

As discussed in detail earlier, students A and B entered the room prepared to take the SAT with individualized probability distributions. The sectors on their pie-shaped graphs were determined by how much studying they did, the amount of rest they got the previous night, the quality of their high school education, the socioeconomic background of their respective families, and a multitude of other factors. These factors affected their individual PD in unknowable ways.

To help illustrate how knowing the unknowable can alter our view of the two scores, imagine that each student entered the test with a probability distribution of scores that was essentially normal, i.e., with a mean and SD that accurately predicts the histogram of scores achieved if the test were taken and retaken many times by each student. Remember, this is impossible to know.

Suppose student A who scored 630 had a PD that was essentially normal, centered on 700 with an SD of 50, so student A had a 68% chance of scoring between 650 and 750. Student A's actual score was disappointing.

On the other hand, imagine that student B, with a score of 650, had a PD that was essentially normal, centered on 600 with an SD of 25. B had a 95% chance of scoring above 550 and below 650. Student B's score clearly beat the odds.

Based on the information that is unknowable, student A would produce an average score of 700 compared to the average of 600 for student B. Of course, these averages are based on taking the exam many times under identical conditions. But if these PDs are known, they do not lie. Student A's true test score is closer to 700 than 630. Analogously, a more representative score for student B is 600. The actual test scores of 630 and 650 were obtained by a single twirl of the spinner, one with the largest sector centered around 700 and the other around 600.

Based on the actual scores earned by students A and B, 630 is still lower than 650, but the possible origin of these numbers should make the difference look less substantial.

VII. CONCLUSION

In academia, numbers are used extensively for assessment, and they typically play a crucial role in honors admissions, retention, and scholarship-award policies; they might also play a role—though much less frequently than outside of honors—in grading policies. When considering the implicit as opposed to explicit value of a number, ask yourself about the origin of the

number. Remember that many of these relevant numbers were produced by a single twirl of a spinner sitting atop a pie-shaped graph with sectors of various but unknown sizes. With that thought in mind, I hope that you will recognize objective data as subjective information and give such data the importance they deserve by becoming a data skeptic.

ENDNOTES

[1] The average weight of different cat breeds ranged from 5 to 20 lbs. For dogs, the range was 4 to 200 lbs.

[2] I estimated the number of cats or dogs needed in a group to have 95% confidence that the average weight of that group would be within + or - 5% of the "actual" average weight for cats or dogs. The number for cats was 140 and for dogs it was 1500.

[3] "Power Analysis" is a method used in statistics to estimate appropriate sample size. I used that to estimate the number of cats and dogs listed above in endnote 2.

The author may be contacted at

Len.Zane@unlv.edu.