Teaching Bits: "Random Thoughts on Teaching"

Deborah J. Rumsey
The Ohio State University

Journal of Statistics Education Volume 17, Number 1 (2009),
www.amstat.org/publications/jse/v17n1/rumsey.html

"Watching our Language When We Teach Statistics"

One of the things I've struggled with over my years of teaching is the language that we use to teach certain statistical concepts and how it affects my ability to get ideas across. In any situation, using language and terms that are more complicated than needed lowers our ability to process information and to see it in a broader context. I believe the same is true for statistics.

For example, I always struggled with the word "spread," as in "shape, center, and spread." Since this term appears in many textbooks, I suspect many other teachers feel the same. Students would ask me what I meant by spread. I'd say it's the amount of variation in the data - then they would ask what <u>that</u> means. We would talk about differences and distances around a central point, and I would end up practically spelling out the idea of standard deviation before I wanted to. There had to be a better way.

Now instead of using the term "spread," I describe variation as "diversity" in the data using various contexts. For example, suppose you have two ponds of fish; in one pond the fish are all the same age and of the same species. In the second pond the fish are of different ages and species. You collect a random sample of fish from each pond and measure their lengths. Which data set has more diversity? The second one does. So the amount of variability is higher in the second pond than the first. Using the term "diversity" helps my students make an intuitive connection regarding the critical concept of variability.

The word "spread" is just one example where having to explain a particular term gets in the way of my teaching the underlying concept. Over the years I've collected many such examples. Here is my top ten list of phrases and terms whose names and usage I believe impede the teaching and learning of statistical concepts. I certainly don't expect you to agree with all of the points on this list (or with any of them for that matter); in fact, some of you might become downright outraged by my opinions herein. My goal, however, is to get us thinking about the impact of the terms and language we use in statistics, and to be bold enough to consider changes where needed.

1. "**Sampling distribution of the sample mean**." We all agree that this concept is hard no matter how you slice it, but the language we have chosen to use here doesn't exactly help matters. There has to be a simpler way to describe the result of taking all possible samples of size n and plotting their means. Now add the fact that we're supposed to talk about the mean of the sampling distribution of the sample mean! I won't pretend to have an easy answer for this one, but for starters we can eliminate the word "sampling" because it seems redundant to say the SAMPLING distribution of the SAMPLE mean.

2. "**Conditional**" probability. The common definition of P(B|A) is the probability of B occurring given that A has occurred; the term "conditional" doesn't come up. In real-world terms the word "conditional" is used in an "if then" context. For example "You can have a cookie on one condition - that you finish your homework." It's not a direct analogy because in statistics we are talking about the probability of B depending on A, not the occurrence or non-occurrence of B depending on A (which is a special case.)

3. "**Given**." You see this word and you already know what I'm talking about – our use of the word "given" in the context of conditional probability problems. How did this term come about? My opinion is that this word was coined by teachers who were frustrated that students couldn't answer questions about conditional probability. Using the word "given" easily cues the students. But then when the teacher changes the wording even slightly, students can't do the problem and accuse the teacher of being tricky because the word "given" is not there. By reducing the process to looking for a certain word, we've lost the opportunity to explore a very commonly used practice of breaking down data. The media doesn't say "Given a voter was a democrat, the probability that he/she voted for Obama is …" They say "For democrats, the breakdown of their voting pattern looks like this…" Big difference.

4. "**A or B**." When talking about the probability of A union B, we use the phrase "A or B." In common language, "A or B" means you either have A or B but not both. (For example, you ask someone if they want a cookie or a piece of cake - this does not imply that they can have both.) However in statistics, when we write P(A union B) or P(A or B) we really mean the probability of A or B or both. Moreover, the formula for the probability of A union B, stated as P(A or B) = P(A) + P(B) – P(A and B), can be deceiving; students often look at it and think we are doing the subtraction to eliminate the chance for both A and B to occur, which is incorrect. We can fix this problem by simply saying that the probability of A union B is the probability of A or B or both.

5. "**Histogram**." A student once asked me why we use the term "histogram" for this particular type of graph, and I had no idea. I did a little research on the history of statistics and this is what I found:

   > Karl Pearson was the first known user of the term "histogram" in a statistical context. The root could be from the word "history" since a histogram provides a record. The Greek root of history is from *histor*, which means "a learned man." The implication is that a learned man is aware of history.

   With all due respect to Karl Pearson, that was then (the 1880s) and this is now; perhaps it's time to revisit the word "histogram" and come up with a more meaningful term. One possibility is to call it a "quantitative bar graph" as compared to a "qualitative bar graph" (our current bar graph.) This broadens the well-known term "bar graph" and helps students to see that histograms and bar graphs are not complete apples and oranges.

6. "**Variance**" of a data set. Teachers and textbooks alike typically include the variance of a data set as one measure of spread (there goes that word again!) Variance is defined and calculated; we get a number but we don't talk about what it means. "Find the variance of this data set of students' ages… okay, the variance is 5; on to the next question." What the students don't realize is that the variance is not in terms of the original units, but rather in original units <u>squared</u>. (For example, for the data set of students' ages, the variance is 5. This means 5 years squared, which makes no sense.) Why discuss a statistic that we can't even interpret? It's a part of the well-worn path that for years and years has led us to calculating the standard deviation. Why not just find the standard deviation in the first place and not even talk about the variance – are we that afraid of a little old square root?

7. Margin of "**Error**." The word "error" means mistake – it's as simple as that. That said, why do we take a common term and change its meaning in a statistical context? What teacher wants to explain that in statistics "margin of error" doesn't actually measure the chance of making an error? Maybe we could use the phrase "margin of variability" instead.

8. "**Correlation**." To those outside the statistical community, correlation simply means two things are related, as in "There appears to be a correlation between gender and political affiliation." Correlation is likened to the word "pattern." But in statistics, the word correlation has a very specific meaning and we fight to make sure everyone knows it. We define correlation as a number that measures the strength and direction of a linear relationship between X and Y – so in our book it's wrong for anyone to say "political affiliation and gender are correlated." Instead of fighting the street use of correlation, we could present our version as a special case under the correlation umbrella. How about using the term "linear correlation," or better yet, "statistical correlation" to get at what we are talking about?

9. "**Regression**." Webster's dictionary defines the word "regression" as "to move backward," as in "He is regressing back to the way he was in high school." How does this term help teachers explain that we are creating a model for a certain type of relationship? It's important to remember that if our goal is to communicate with our students, we should use the language that is most effective. The word "regression" to me is not effective, at least not for introductory statistics students.

10. "**Inference**." In general terms, the word "infer" means to generalize conclusions to a larger entity, and it means the same thing in statistics. However, I don't think the word "infer" is used enough in the real world to help students make the intended connection. To clearly make the point that we are moving from a sample to the entire population it might be better to use the phrase "generalizing to the population;" or we can talk about the process of "drawing conclusions" rather than the process of "making inferences." I don't think the word "inference" is helpful or needed.

Those are my random thoughts on teaching for this time around. Now what do YOU think?