HIDDEN JARGON: EVERYDAY WORDS WITH MEANINGS SPECIFIC TO STATISTICS

Christine M. Anderson-Cook
Statistical Sciences Group, Los Alamos National Laboratory, United States of America
candcook@lanl.gov

If students of statistics or our collaborators from other disciplines do not immediately understand the terms "probit regression" or "kriging", we are not surprised and are happy to carefully explain these advanced statistical terms and concepts. A different class of words has one or more distinct statistical meanings in addition to their standard English definitions and is ripe for confusion among students. More subtly, some of the key words of statistics, such as design, parameters, model and analysis, are also the key words of other disciplines. Each discipline has assigned subtle associations to these words beyond their standard dictionary definitions with hidden connotations and implied connections to concepts. We consider some multi-discipline keywords, discuss how to build awareness about the potential for confusion by interdisciplinary collaborators or out-of-major students, and how to resolve the communication gap.

INTRODUCTION

As instructors and communicators of statistics, we have all likely encountered difficulties with language. If language obstacles exist in the statistics learning environment, then deeper conceptual understanding cannot even be undertaken. Statistics has some specific vocabulary which has the potential to confuse. We consider three types of language issues which might present themselves during discussions of statistics between instructor (expert) and student (novice): 1. specialized statistics words, 2. distinct definition words, and 3. connotation laden words. Note that we consider this topic broader than just one focused in the statistics classroom, with novices are students in introductory statistics classes or research collaborators from other disciplines.

Specialized statistics words are those which exist and have meaning only in the context of statistics. The examples of "kriging" (a group of advanced techniques for interpolating spatially connected observations) or "probit regression" (a technique for predicting a binary response using a generalized linear model with a link based on the cumulative distribution function of the normal distribution) are typical, in that the words are not recognizable to novices, and hence are an immediate trigger for lack of understanding.

Distinct definition words are statistics words which exist in the broader English language, but are used in different ways in different situations. An example is the word "random", which in general English usage typically is interpreted as "haphazard", but in statistical terminology can mean an outcome is not predictable (e.g., random event), or can be associated with numeric values assigned for different outcomes (e.g., random variable), or can relate to the design of experiments phase where treatment combinations are assigned to experimental units (e.g., randomize). Novices are likely to recognize the word, but find it difficult to discern which interpretation of the word should be appropriately associated with a particular usage.

Connotation laden words exist in the broader English language, but have been adopted by multiple disciplines to include additional connotations specific to that discipline. The core essence of the word has typically been preserved, but possibly different existing English definitions for the same word have emerged as the dominant interpretation within the discipline. For example, the word "design" can mean "prepare the preliminary sketch (verb)", "intend for a definite purpose (verb)" or "the combination of details or features of a picture (noun)". The primary statistical definition is likely tied to "design of experiments" where data are collected for a specific purpose. As we will see, other disciplines have selected different primary associations. Novices with expertise in other disciplines may find this category of word particularly vexing since they bring the connotations with them, only to discover that these associations are not relevant in statistics.

In the remainder of the paper we consider each category of potentially problematic words, and then provide some suggestions for how a statistical instructor (expert) can help students of statistics overcome their confusion.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

SPECIALIZED STATISTICS WORDS

This is perhaps the most straightforward of the categories to address, since both novices and experts are readily aware of when a problem exists. For the novice, the word is unfamiliar and immediately begs the question of clarification. For the expert, the word is associated with a specialized technique, distribution or attribute, for which there is a (perhaps distant) memory of having to learn about this topic.

Careful introduction of new terminology, organizing glossaries and references with the details of the concept can all prove useful as strategies for helping novices become familiar and comfortable with them. For students communicating, it is helpful to encourage them to discuss their understanding of concepts both using the statistics specific words as well as without using any technical jargon. Dictionaries (see Everitt, 2006; Upton & Cook, 2008) and encyclopedias (see Dodge, 2008) of statistics tend to focus on this category of terms.

DISTINCT DEFINITION WORDS

Perhaps my first awareness of this potential category of words occurred within weeks of my arrival at Los Alamos National Laboratory. In a meeting of LANL statisticians, a fast-paced conversation was taking place about experimentation within the lab. I heard "DOE in the DOE has really evolved". Needless to say, I was quite lost! Well, the usage of DOE (Design of Experiments) and DOE (Department of Energy) in different parts of the sentence made perfect sense, when fully up to speed. However, for the uninitiated, the abbreviations were identical and knowing only the first interpretation meant that the conversation made no sense to me at all.

Even though here we are not focused on acronyms, the above example illustrates the potential problem. Acronyms by definition have hidden information in them, namely what the letters stand for. Some potential for problems can be disguised by related but distinct meanings, which have both standard English and statistics-specific definitions.

One class of words has a standard English definition and a distinct statistical definition. Examples of this type of key word include "uniform" and "confounding".

- *Uniform* – I. Standard English: without variations in detail
  II. Statistics: evenly spread over a range (e.g., uniform distribution or design)
- *Confounding* – I. Standard English: confusing or baffling
  II. Statistics: associated with both the probable cause and the outcome (e.g., a confounding or lurking variable in regression)

The solution in this case can involve careful initial clarification of the statistics-specific definition with discussion of how the term is related to the standard English usage, but distinct with a more technical definition. In subsequent usage, developing conventions to separate the multiple definitions, such as replacing "uniform" with "statistically uniform" can help distinguish terms, and provide a natural trigger for the novice that the intended definition is the statistically-specific one.

A second class of words has both a separate English and multiple distinct statistical meanings. We consider several common words:

- *Control* – I. Standard English: to exercise restraint or direction over
  II. Statistics: 1. give no treatment (e.g., the control group received a placebo), 2. adjust for (e.g., to control for a variable to try to separate its effect from the treatment effect), 3. a setting where factor levels can be assigned to experimental units (e.g., a controlled experiment allowed for direct comparison of treatments).
- *Independent* – I. Standard English: not influenced by the thought or action of others
  II. Statistics: 1. the results of one event do not affect the results of a second (e.g., the probability of "getting an ace" and "getting a heart" are independent for a standard deck of cards, 2. having the characteristic of being useful to explain the response (e.g., an independent variable in regression).
- *Random* – I. Standard English: haphazard
  II. Statistics: 1. not predictable (e.g., random event), 2. associated with numeric values assigned for different outcomes (e.g., random variable), 3. an experiment where the assignment of treatments is deliberately determined by chance (e.g. a randomized controlled experiment), 4. non-systematic (e.g., a random error from a measurement system).
- *Normal* – I. Standard English: conforming to the general standard

II. Statistics: 1. related to the bell-shaped curve (e.g., the Normal distribution or random variable), 2. to standardize (e.g., the residuals were normalized to have constant variance).

The Glossary of Statistical Terms (Stark, 2009) provides a careful discussion of these and many other such multiple definition words. Because the words have standard English uses as well as several statistically-specific definitions, the previous method of predicating with "statistically" is only a partial solution. Upfront discussion of the different usages of the words through examples (such as those given above) encourages students to contemplate the differences directly. When words are initially introduced it can also be helpful to explain to students why statisticians use that term the way they do and how it is connected to the Standard English usage of the work. In an interdisciplinary collaborative environment, avoiding the use of these terms during explanations can remove unnecessary distractions and maintain focus on the true problem of interest.

CONNOTATION LADEN WORDS

This final category of words is similar to the distinct definition words in that they are words commonly used in standard English, but also have some statistical-specific definitions. The additional feature that these words have is that they have been adopted by other disciplines as well and have connotations and associations specific to that discipline as well. This can lead to even greater confusion with many alternate interpretations. Below are listed a number of words that we have encountered which have quite rich interpretations in multiple disciplines:

- *Design* – I. Standard English: 1. a plan or project, 2. the organization of formal elements.

  II. Statistics: a data collection strategy for obtaining observations where levels of the input factors can be controlled by the experimenter.

  III. Engineering: the process of creating a working version or prototype of the system with all of the key elements of the final version.

  IV. Business: 1. the realization of a concept into a plan, which helps to achieve the final objectives of the project, 2. the format of a questionnaire for obtaining sampling data.
- *Parameters* – I. Standard English: 1. limits or boundaries, 2. characteristics or aspects

  II. Statistics: unknown terms in a statistical model, which we wish to estimate using data (and possibly priors). Note that statisticians can view parameters either as fixed but unknown (frequentist) or variable (Bayesian).

  III. Mathematics: constants or variable terms in a function that when specified determine the specific characteristics.

  IV. Computer: variables that are given a specific value during execution of the program or code. The variables are simply placeholders until the desired values are identified.

  V. Engineering: limits or constraints on reasonable or planned system operation. These bounding conditions provide an inherent scope to where it is sensible to exercise the system.
- *Model* – I. Standard English: 1. a standard for imitation or comparison, 2. a miniature representation to show the general characteristics.

  II. Statistics: The mathematical form or expression for characterizing the general family of equations from which we wish to estimate a relationship. This form includes contributions from the data and from the parameters which we wish to estimate. There are typically assumptions that are associated with the model which relate the physical application to the assumed mathematical form.

  III. Engineering: a small-scale version of the system with sufficient details about the key characteristics to illustrate its intended use and functionality.

  IV. Business: a representation of a concept or system to facilitate understanding by eliminating superfluous details, or to aid in decision making by allowing exploration of hypothetical scenarios.
- *Analysis* – I. Standard English: 1. separating material into its constituent elements, 2. the process of studying the nature of essential features of an object.

  II. Statistics: the process of combining the observed data with the functional form of the model, to obtain estimates of the model parameters with their associated uncertainties. There are different approaches (maximization or integration) for obtaining these optimal parameter values that best match what has been observed in the data.

III. Engineering: 1. the process of decomposing compounds into their intrinsic elements, 2. computer codes to simulate parts of a system for improved understanding.

IV. Business: 1. the systematic examination of data by breaking it into its components to uncover interrelationships, 2. the examination of data to uncover and understand cause-and-effect relationships.

- *Factor* – I.Standard English: one of the elements contributing to a result or situation

II. Statistics: a controllable input to a designed experiment, which is manipulated in during the data collection process, and is thought to contribute to changes in the observed response.

III. Biochemistry: a necessary substance for a process to progress.

IV. Business: 1. a variable under examination in a study (input or response), 2. an element that causes a certain effect.

For this category of words, there are two potential problems: 1. the statistical connotations will be missed, and 2. the connotations from another discipline will be inappropriately applied. For example in a discussion about "analysis", a collaborator with a business background may miss the aspect of combining the data with the model form to obtain estimates, and insert the notion of having established a causal relationship between inputs and response. Similarly when discussing design, an engineer will be focusing on attributes of the system under study, while statisticians are focused on data collection strategies for understanding the system. It is also interesting to note how interconnected the above five terms are for statisticians–the statistical definitions of many of the terms naturally could include one or more of the other connotation laden words. This interrelationship between the concepts can lead to a compounding of the potential confusion. An effective strategy for minimizing confusion is to repeatedly discuss the richer associations for these key words in the context of specific examples to describe which connotations are appropriately included or excluded. Similar to the distinct definition words, it can be helpful to connect the disciplines specific definitions to the Standard English definition to show their association.

CONCLUSIONS

We have described several categories of words commonly presented in statistical applications and courses. As you will have no doubt noted, the boundaries between categories are somewhat blurred with standard English usage and discipline specific usage being intertwined. Our familiarity with these words as we become increasingly conversant in our discipline served to make the problem for the novice learning about statistics more subtle and us more likely to overlook the potential for misinterpretation. It is important for us to realize that many statistical novices will not be able to resolve these inconsistencies and ambiguities on their own. It may often be the case that the more expertise that a person has in another discipline the more they are aware that there can be differences in the usage of some terminology between Standard English and in the context of a technical field. Hence, particular care should be taken for undergraduate students who are in an initial statistics course and may not have an already established area of expertise of their own. With increased awareness of the possible problems and a plan to openly discuss areas of confusion, some of the misunderstandings can be avoided or minimized. Direct discussion about terminology and connotations, including a careful introduction to the terms which clarifies connections and differences in the differ contexts, can be made memorable for students and help them retain the distinctions between the various disciplines.

REFERENCES

Dodge, Y. (2008). *The Concise Encyclopedia of Statistics*. New York: Springer.

Everitt, B. S. (2006). *The Cambridge Dictionary of Statistics* (3rd edition). New York: Cambridge University Press.

Stark, P. B. (2009) *Glossary of Statistics Terms* (referenced 01/05/2010). http://www.stat.berkeley.edu/~stark/SticiGui/Text/gloss.htm.

Upton, G., & Cook, I. (2008). *A Dictionary of Statistics* New York: Oxford University Press.