

Plan of the Course

Rules of Evidence follows a straightforward outline, beginning, in Volume I, with the study of single variables and advancing, in Volume II, to two variables. With one variable it introduces the use of distributions of the data and the use of summary statistics for the center and the variation. With two variables it introduces the concept of correlation between variables and linear relations including regression.

Interpretation

Within the outline, one distinction of *Rules* is its emphasis on interpretation, in English, accompanied by visual and graphical displays. It is easy for a student to become absorbed in technical details — harder to connect the data to reality. But data analysis is without purpose if it does not connect to reality.

In data analysis technical errors — dropping a decimal place, or forgetting a square root — are not mere technical errors of no consequence. An error usually leads to an absurd interpretation — to estimated growth rates that are ridiculous, to predictions of wealth that are impossible, to drug treatments that kill — and if the analyst does not see the absurdity introduced by “technical error” it means that the connection between the numbers and the data has been lost.

For this reason no example presented to students, and no homework presented by students, should be considered complete without an intelligible write up, in English: A number, like “3”, is not the answer to any exercise in data analysis. No matter how long the file of computer output, no matter how pretty the graphs — the answer has to make sense. “3 dollars,” or “3 pounds of potatoes per pound of fertilizer,” or “3% increase in population per year,” or “a dose of 3 grams of antidote per kilogram of body weight” — may be an appropriate statement about data.

Homework

This has an important implication for homework: For my own classes I assign nightly homework but, many decades later, I have accepted the

prissy doctrine of grade-school English teachers to the effect that “If you can’t put it in writing, then you don’t understand it.” And to “put it in writing,” to produce an intelligent interpretation of data rendered in clear technical English takes time. It is not clear why good technical writing is difficult and time consuming. But, empirically, it is. This means that I assign very few problems — usually no more than one per night. This certainly makes the course easier to teach but, more important, it is the right approach: If I were to assign half a dozen interesting problems for one night’s homework, the magnitude of the assignment would, by itself, tell students the opposite of what I intend. The size of the load would tell students to get the numbers right and ignore meaning because it is basically impossible to work out five or six examples and write them up in intelligent English — all in one night.

Exploratory Analysis

Rules of Evidence is heavily influenced by John Tukey’s introductory text *Exploratory Data Analysis* in which there is no mention, nor any need to mention probability or Gaussian distributions or, for that matter, least squares or even means and standard deviations — an introduction that Tukey accomplishes without sacrifice of either rigor or precision.

Rules will not go that far, but one reward of reading Tukey’s text is to discover, or re-discover, how much can be accomplished well, how much can be accomplished with both rigor and precision — if the analyst has a firm and clear understanding of the basics. When a student can “eyeball” numerical estimates of the center and the deviation, when a student can “read” estimates of slope and goodness of fit right off of a well formed graph, when a student can interpret these things in terms of income, or education, or time, or temperature — as appropriate to the data — then the student is ready for the technology of “r-squared” and standard errors of estimate.

There is a line of thought that says you can not learn the meaning of a thing unless you’ve got two of them: It forces you to abstract from the examples to the principle. For this reason, I have tried to make it a rule that every time I demonstrate one technical solution to a problem in data analysis I also demonstrate a second: A mean is one realization of the center of a distribution, the median is another. For this reason too, I have tried to introduce data analysis with one broad interdisciplinary course — to be followed by courses that adopt the special practices of the separate disciplines: If the student sees only one solution to a

problem, as is usually the case within a mature scientific discipline, it is hard to see that there are others, and then to understand the choices that have been made and the trade-offs — choices that make one solution good or better than another.

Prerequisites

I assume a background in mathematics and, equally important, a willingness to use that background. I do not assume that the student is at ease with that background when it comes to using it, in the real world. That is the business of this course. For example, I use logarithms. I know full well that students are in varying degree “uncomfortable” with such things. Even those who are perfectly comfortable with the “math” need to learn to work with these logarithms when they apply to the size of real world populations, “log people”, or when they apply to real world wealth “log dollars”. I expect students to work with what they have already learned in secondary school mathematics, and to learn to use it with data. Similarly, I make some (optional) use of calculus, specifically the use of derivatives to find a minimum or a maximum. And, again, I know full well that there will be some discomfort. But the reward is to see how important problems are solved, using the derivative, and to see how

calculus invests coherence and strategy to what is otherwise no more than a collection of numerical techniques.

These are reasonable demands of the “average” student in an above average university. It is difficult for faculty to believe this because the “average” student of this generation has a stronger math background than the “average” student of our own (the faculty’s) generation. But it is true. If anything, our students may have too much *faith* in mathematics. They have acquired the lay person’s idea that mathematics *is* science and they have learned, somewhere, that the more mathematical something looks, the more scientific it is. For a lay person, knowing little of either science or mathematics, that may be a reasonable approximation. For a scientist it is a fatal error.

Technology in The Liberal Arts Curriculum

The reader will see clear traces of my own teaching environment embodied in this text: I teach at an institution that sees itself as a liberal arts college. I do not interpret liberal arts to mean “artsy”, and I certainly do not interpret liberal arts to mean “non technical”. I interpret it to mean that, at least initially, a liberal arts education attempts to do something broader than teach students a trade. I cannot, it would violate

the “rules” of a liberal arts education for me to say to my students, “Here are the symbols and rituals of your trade — learn them!” That’s not allowed. The use of the tools and the meaning of the tools go together.

Moreover, data analysis occupies an important niche in the culture. Data analysis is part of the scientific method: Humans have many ways, often strange, for establishing truth. We may rely on authority. We may believe what “everyone” knows to be true. We may find answers in culture and ideology. We may use pure logic. By contrast, science offers, first, skepticism, and then a way to put questions to reality. Data analysis is applied epistemology wherein the questions “What do we know?” and “How do we know it?” become inescapable and must be answered. In later life the skills of the scientist, the business analyst, the policy maker become specialized and differentiated. But in school, at the undergraduate level where these skills begin, they have a common root in the skepticism and the methods of inquiry summarized by the phrase “scientific method”.

People like ourselves created these methods to begin with. And people like ourselves have to understand what the methods are for, and how they were invented, and how they may be re-

invented or modified when that is what’s called for. That’s why technical work, including data analysis, belongs in the liberal arts curriculum.