INTEGRATIVE STATISTICS

Customized Statistical Analysis
and Survey Research

Long, *Daubert*, and Acceptance of Methods

Roland B. Stark, M.Ed.

Statistician and Research Consultant

September 13, 2023

Statistical evidence with respect to police impartiality or bias has become central to many criminal cases in the wake of *Commonwealth v. Lora*, 451 Mass. 425 (2008) and *Commonwealth v. Long*, 485 Mass. 711 (2020). In the tradition of *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993), defense attorneys, prosecutors, and judges increasingly seek to evaluate the soundness, validity, or credibility of the findings of statisticians serving as experts.

The questions posed as part of these evaluations often, in this statistician's experience, focus excessively and sometimes inappropriately on a single aspect among the five aspects of such evaluation presented by *Daubert.* These aspects are nicely summarized by Cornell Law School's Legal Information Institute[1] as involving testing of techniques and *falsifiability* of theories; record of publication and peer review; error rates; standards for use; and degree of acceptance within the relevant scientific community.

Rather than concentrating on the primary object discussed in *Daubert* – the reasoning and principles underlying a statistical analysis – the wording typically used in red-herring questions focuses inordinately on whether the methods used have been widely accepted. Such wording, and the manner of thinking that it may engender, tends to hinder rather than support adequate evaluation.

Examples of such problematic wording include the following:

"Are the methods used for this report accepted [widely or generally] [in the scientific community] [in your professional community] [in your scientific discipline] [among statisticians] [among analysts] [among researchers]?"

Such a global question too often results in oversimplified, imprecise, and/or naive thinking that can confuse or derail the evaluation and thus be counter-productive.

It is much more effective for eliciting helpful information to ask about past acceptance separately with regard to specific methods used.

[1] Cornell Law School (2023, August update). Daubert Standard. https://www.law.cornell.edu/wex/daubert_standard.

Just as important is the need to look beyond past acceptance of specific methods to properly evaluate what might be termed "packages of methods" customized to a unique situation.[2]

In the arena of statistical evidence, much better than asking global "wide acceptance" questions is to recognize a distinction between two categories of statistical methods.

I. "Category I" statistical methods are ways of performing distinct, discrete, set procedures or operations on specific types of data.

I.A. No doubt readers will be familiar with the following examples, which may be safely skimmed.

- Determining what fraction (or percentage) one set of events constitutes, out of a larger set. Computing that

  4 events out of 10 = 2/5, and

  2/5 = 40%,

  employs no method (in any meaningful sense) other than division.

- Calculating an average. What is most commonly termed the "average" (the *mean* or *arithmetic mean*) of 40 and 10 is calculated as the sum of the numbers divided by the sample size (i.e., how many items are being included.)

  The average of 40% and 10% =

  (40% + 10%) / 2 =

  50% / 2 =

  25%.

- Calculating a *weighted average*. If Location 1 is instead assigned twice as much weight (i.e., twice as large a role) in a certain analysis as is Location 2, and if the quantities being averaged are 40% for Location 1 and 10% for Location 2, then the combined, weighted average will of course be closer to Location 1's percentage than to Location 2's. This weighted average will equal

  [(2 x 40%) + (1 x 10%)] / [2 + 1] =

  90% / 3 =

  30%.

I.B. Other examples that hopefully are familiar:

- The *absolute* difference, sometimes called the difference in *percentage points*, between 40% and 10% is simply

  (40 – 10) =

  30 percentage points.

[2] This piece treats as virtually interchangeable the terms "method," "technique," and "approach." The term "methodology" is opaquer, best defined as and used in *Daubert* as a body, system, or package of methods.

- The *ratio* between the two numbers is equal to

  40% / 10% =

  4.0.

- This ratio can be restated such that

  40%, compared to 10%, is 4.0 times *as great* =

  400% *as great*.

- This ratio can also be converted into a *relative difference* of

  (40% - 10%) / 10% =

  3 / 1,

  which shows that 40% is, compared to 10%,

  3.0 times *greater* or 300% *greater*.

All the statements and calculations above follow from Category I methods and in fact make no substantial use of any mathematical operations beyond the four basic operations of addition, subtraction, multiplication, and division. All such statements and calculations are completely uncontroversial and should be universally accepted among anyone litigating or adjudging a legal matter. At this level, *Daubert* considerations are simply not applicable.

Also fitting into Category I are a group of more specialized data-analysis methods. These methods, developed by leading statisticians worldwide over the past 120 years, also have gained extremely wide acceptance as established, sound, valid ways of answering questions using quantitative data. This acceptance has been true across nations, types of institutions (e.g., commercial, governmental, academic, non-profit), and content-area disciplines (e.g., the physical and biological and social sciences, law, criminal justice, health care, and education).

These more intensively statistical Category I procedures may involve assessing group differences, as above; distinguishing patterns that occur over time; identifying the ways in which different phenomena might be correlated with each other; evaluating the ways in which outputs might vary depending on key inputs; and so on. These procedures include, among many others, Chi-Square Tests, Tests of Proportions, *T*-Tests, Analyses of Variance, Mann-Whitney *U* Tests, Pearson Correlations, Regressions, and Monte Carlo Simulations.

Such methods are ubiquitous in statistical textbooks, peer-reviewed articles, and publications of all sorts that discuss either a) effective ways of analyzing data in general and in theory, or b) examples in which these methods have been applied to real-world contexts. In short, a considerable number of techniques used by statisticians and other quantitative analysts and researchers are extremely widely accepted and would invariably pass *Daubert* tests as long as the methods in question are appropriately matched to the questions at hand and to the types of data available.

Techniques designed for *benchmarking analysis* deserve a special note. In benchmarking, we might seek to compare, e.g., the proportion of motorists belonging to a certain protected group --

- among all drivers who were stopped by a police officer

  and

- among all those who were driving on the roads that an officer patrolled.

An imbalance between these two proportions (or *fractions* or *percentages* or *rates*) would suggest discriminatory enforcement of the law. This is crystallized in the statement,

> "[A]bsent impermissible discrimination, the enforcement rates should reflect the demographic composition of all drivers."
> *Long*, 485 Mass. at 731.

The first of these two bullet-point percentages might be obtained in a straightforward manner from police records. The second, however, must be approximated via some type of Driving Population Estimate, or DPE. The Massachusetts Supreme Judicial Court, in *Lora* and *Long*, unquestionably found benchmarking and the associated estimation to be a worthwhile endeavor. The Court treated such estimation as the most prominent, probably the most important of all statistical paths to assessing an officer for inequitable treatment of protected groups. Without the Court's reliance on and lengthy discussion of such estimation, the *Lora* and *Long* opinions would be scarcely recognizable.

II. "Category II" statistical methods, on the other hand, are ways of resourcefully customizing or adapting Category I methods in order to meet the idiosyncratic needs of an analytic context.

II.A. Before discussing the more complex matter of demographic benchmarking, an initial illustrative example involves a simpler test of whether outcomes of police stops tend to break down differently for drivers of different ages. A hypothetical *sample* of stops, taken from some larger *population* of them, could produce the following:

Group of Drivers	Warning	Civil Citation	Criminal Citation	Arrest	Row Total
Teenage	63	22	11	4	100
Adult	68	26	4	2	100

In every shaded column (i.e., comparing vertically), the two age groups show different breakdowns. The Adult drivers received more of each of the first two, more-lenient outcomes; the Teenagers, more of the two harsher outcomes.

The default approach to formally addressing this question for statistical significance would use a Chi-Square Test. It would show how commonly such a set of differences would occur if nothing but mere chance were responsible, and by extension just how inconsistent or non-credible this set of numbers would be under that "mere chance" assumption. Such results would help us generalize from this limited *sample*, which is of lesser interest, to a wider or longer-term *population* of events, which is of greater interest.

However, these data pose a problem for such a test. Too few people of any age were arrested in order for this default, standard version of the Chi-Square Test to produce a strictly accurate and trustworthy result. The data in the "Arrest" column are too sparse, and perhaps in the Criminal Citation column as well.

Here, statistical analysts would implement a workaround. Most commonly, they would condense the table slightly, by combining numbers for Criminal Citation and Arrest. This would create sufficient sample sizes per individual cell there (15 and 6) to obtain an accurate result from *that* version of a Chi-Square Test – even though it would necessarily be disregarding a certain amount of the original information.

Some analysts would prefer to condense further, into just the two outcome categories of Non-Criminal and Criminal. This third version of a Chi-Square Test would need to sacrifice additional information but would potentially allow an even more definitive statement, buttressed by larger sample sizes within cells, as to whether there was *some* kind of age difference in the larger population of which this was a sample.

There are pros and cons to choices such as these. And of course, in good reporting one would include as complete and transparent a treatment of these decisions as the situation and audience seem to call for.

In this stop-outcomes-by-age example, either of the above adjustments would be widely considered by professional analysts to be legitimate ways of testing for statistical significance. The default method would not. So we see that it is not just the traditional, default, *named* method that is necessarily the most deserving of scrutiny as one evaluates the merits of a chosen approach; the way that approach may have been customized to fit a situation matters as well.

II.B. We can and should extend this principle – of the importance of considering customization – to more complex *Long*-type situations, in which multiple departures from standard data-analytic conditions may call for far more individualized and far more numerous adjustments to one or more accepted approaches.

Often a question being investigated concerns the expected range of results that might occur if a variety of inputs, i.e., factors affecting results, were difficult to pin down exactly. Each such uncertain (or variable or error-laden) input might need to be assigned its own range of plausible values that it might take on. For instance, unavoidable uncertainty might dictate that, rather than claiming to know that Worcester residents make up exactly Y% of the drivers found in Boston, we should assign a likely range for that percentage, from X% to Z%. Again, the basic structure of such an analysis could be modeled on a long-established, extremely well-accepted method: Monte Carlo Simulation[3], a versatile, *multivariate* technique first developed as part of the Manhattan Project in the 1940s.

This writer's benchmarking analyses often take advantage of simulation. The passage below is drawn from a working paper currently being prepared for peer review. The paper endorses two types of reasoning, or two types of statistical principles, linking back to *Daubert*:

> "Statistical analyses that carry with them great consequences for criminal defendants or police officers need to do two things. First, they need to acknowledge uncertainty. Those

[3] Often called simply "simulation."

> who estimate critical quantities involving racial justice or police accountability need to communicate skillfully and candidly about the plausibility of alternatives to their estimates. Second, these analyses should, to the extent possible, take into account multiple factors that may have affected an outcome. Both of these goals can often be met through Monte Carlo simulation when other methods fail.
>
> By customizing statistical models, simulation analysis allows for uncertainty or variation in the data in ways not accounted for by standard [i.e., Category I] statistical procedures. Most inferential statistical tests [again, in Category I] make allowance for a single type of error: sampling error. In contrast, demographic estimates, especially in benchmarking studies, often entail some degree of error not just via sampling error but in multiple relevant factors. This imprecision cannot be accurately captured formulaically. Simulation allows one to see the gamut of ways in which these multiple sources of uncertainty, perhaps acting in concert, may have affected the outcome."[4]

The referenced paper and reports this analyst has created for specific criminal cases cite several sources that further explain simulation and affirm its usefulness and wide acceptance among statisticians and others. One could cite thousands more. Simulation is used in practically every arena that involves quantitative estimation – in both theoretical and empirical contexts. In fact, for the more cutting-edge analyses prevalent in the top-tier statistical journals, simulations can be found in roughly a third of articles published in recent years. There is no question among statisticians that simulation is an essential tool for arriving at answers when other approaches leave too much room for doubt.

Even though the Monte Carlo simulation approach is ubiquitous and widely accepted, an analysis such as described below might require customization in so many respects that no previously applied simulation from another context could furnish an adequate template. No comparison between such methodologies would constitute an apples-to-apples comparison.

Nor would it be adequate to conduct an extensive series of comparisons of very specific decisions as made in one context vs. in another. To break down the object of evaluation into small fragments, assessing them piecemeal, is termed *analytic assessment*. Multifaceted, customized situations require, at least in part, *holistic assessment*.

As an illustrative example, let us suppose that in a benchmarking analysis we seek to compare

a) the proportion who were teenagers among all motorists stopped by two officers in their respective careers, to

b) the proportion who were teenagers among all motorists driving on the roads these officers patrolled in the same time periods.

Again, the narrow task of testing for statistically significant differences between two proportions is extremely well accepted. Comparison of an officer's % to a benchmark % is common and accepted as well, having been used in *Lora* and *Long* (although this is not so much a "professional statistical community" issue as a choice of what is worth comparing – an extra-statistical choice of interest specifically in criminal justice and law enforcement.)

[4] Stark, RB (manuscript in preparation). Monte Carlo Simulation for Driving Population Estimates.

To enable this comparison, however, we must first develop one or more Driving Population Estimates (or DPEs) to approximate item b). Ultimately we would consult the US Census or other reputable sources to find, and then to combine using some set of weights, statistics on the population size and age for the residents of each municipality that contributed substantial numbers of drivers to the roadways patrolled. Using systematically collected data to learn locations' demographics has a history of acceptance in the US that goes back to 1790.[5]

Which municipalities should be deemed important contributors, and in what proportions, may well be debatable. Now, the *Lora* question – of whether to consider only Auburn, MA residents in studying drivers who happened to be driving within Auburn's city limits on a major highway (Route 290) – is another extra-statistical choice. Who would believe that any major highway could limit a motorist to travelling only within their own town's borders? Or that every driver always stays within their own town?

But let us suppose that we properly recognized the need to identify *multiple* locations, both within the patrol area and outside of it ("feeders"), whose demographic statistics needed to be incorporated. Before we could identify the locations and conduct the associated calculations, for our analysis we might need to examine many factors in order to estimate just what portion of the drivers seen in the patrol area would have been residents of each of many potentially relevant cities and towns. The many factors needing attention in this hypothetical study might include the following:

- The length of each officer's career in traversing the roads; we might need to give more weight to one patrol history than to the other.
- The different but overlapping lists of cities and towns patrolled by each officer.
- The fraction of time each officer spent patrolling major highways, which draw drivers from innumerable feeder locations near and far, vs. patrolling neighborhood streets, which draw more locally.
- The distance from each patrol area itself to each location identified as a feeder (whether we choose to measure from, e.g., each one's center or its nearest border). All else being equal, closer locations tend to supply an area with more drivers.
- The patterns of automobile access to each of the two areas patrolled. Even if one feeder town is farther than another from a patrol area (preferably measuring roadway, not as-the-crow-flies, distance), the first town might have a more direct and convenient route in. This would affect each feeder town's role in calculations.
- The fraction of time each officer typically spent in each location. All else being equal, more time known to have been monitoring an area with a higher concentration of teenagers will of course increase our eventual estimates of the % in the driving population who were teenagers.
    - Different sources may indicate different fractions of the total time spent by the officer in each location. For example, frequently a police department's or unit's public-website characterization will differ from information obtained in discovery

[5] United States Census Bureau (2023). 1790 Overview. https://www.census.gov/history/www/through_the_decades/overview/1790.html.

or through a public-records request. Even among the latter two types, sources may not align, as when an official summary statement about the officer's focus by time and location differs noticeably from what is calculable from spreadsheet data.

The reader might suppose that, once an analyst establishes that the driving population consists largely of residents from a certain set of locations in a certain proportion, then, at least, characterizing each such contingent demographically will be straightforward. In actuality, the required choices would not end there. Instead, one may need to grapple with additional issues involving demographic sources:

- Demographic statistics might need to be obtained –
    - at different geographic levels (e.g., city or town; neighborhood; ZIP code; Census tract; or Census block), with potential incompatibilities arising;
    - from different sources –
        - whose level of credibility might differ;
        - whose numbers might disagree;
        - which include different sub-locations as making up a given location, or which use different labels for the same location, or vice-versa, or both;
        - which report statistics from different timeframes; and/or
        - which otherwise employ different demographic methods.

It would be fantasy to suppose that we could consult some almanac or other reference to obtain – qualified by all these conditions, taking into account all such variables that appear relevant, and subject to all of these myriad decisions and calculations – the true fraction of the drivers encountered who were teenagers.

Nor would any set of previous scientific studies have considered all the conditions, variables, and informed decisions involved in such a study to the point where scientific consensus had already affirmed the best way of customizing the analysis for this idiosyncratic situation. The *Daubert*-type question of "prior acceptance" would not apply. It would hardly make sense to ask whether a just-produced, highly customized "package" of analytic methods has *already* been met with wide acceptance. And so one seeking to assess the worth of such a study would need to assess its various aspects in context, paying special attention to the underlying reasoning and scientific principles followed.

II.C. Recognition of Category II approaches merits a statement about *false positives*. Needless to say, it is undesirable if, based on a complex body of statistical work, law-enforcement inequities are found that are spurious. Therefore, this statistician's comparisons document extra measures taken, tailored to the context, to be fair to police officers in particular. These measures make allowance for factors that, if considered, might show officers' patterns of decisions to be quite equitable with respect to race or ethnicity after all. Such efforts, too, from *Long*-type analyses seldom have direct analogs to what has been published in the scientific literature and so are seldom subject to the question of past acceptance there.

However, more generally, there is an extraordinarily rich history among the work of statisticians and lay people alike to try to account for multiple factors, perhaps some less than obvious, that might affect conclusions. To take into account, or *to control for*, possible *confounding* or *nuisance variables* is universally endorsed as being a hallmark of good research, good science, and good statistical analysis.