The Federal Rules of Evidence; Some Concluding Remarks

Psychology (Statistics) 484

Statistics, Ethics, and the Social and Behavioral Sciences

June 19, 2013

Beginning Quotations

Gentlemen of the jury, there are three kinds of liars: the common liar, the damned liar, and the scientific expert.
– W. L. Foster

If it doesn't fit, you must acquit.
If it doesn't make sense, you should find for the defense.
– Johnnie Cochran

Nothing is so unbelievable that oratory cannot make it acceptable.
– Cicero

The superior man understands what is right; the inferior man understands what will sell.
– Confucius

Some drink deeply from the river of knowledge. Others only gargle.
– Woody Allen

Week 15: The Federal Rules of Evidence; Some Concluding Remarks

— the Federal Rules of Evidence and the court admissibility of expert witnesses and scientific data; the Daubert trilogy of Supreme Court decisions

— the importance of context and framing in the presentation of data; the work of Tversky and Kahneman, and the more recent points made by Gigerenzer and his colleagues ("Helping Doctors and Patients Make Sense of Health Statistics," in the series sponsored by the Association for Psychological Science, Psychological Science in the Public Interest)

— Daubert v. Merrell Dow Pharmaceuticals (1993): The Federal Rules of Evidence govern the admission of scientific evidence in a trial held in federal court. They require the trial judge to act as a gatekeeper before admitting the evidence, determining that the evidence is scientifically valid and relevant to the case at hand.

Required Reading:
SGEP (449–492)–
Junk Science
The Consequences of Daubert and the Data Quality Act (of 2001)

Popular Articles–
Science and Society: The Interdependence of Science and Law, Stephen Breyer (*Science*, April 24, 1998)

Something Rotten At the Core of Science? David F. Horrobin (*Trends in Pharmacological Sciences*, February, 2001)

Is Science Different for Lawyers? David L. Faigman (*Science*, July 19, 2002)

Scientific Evidence and Public Policy, David Michaels (*American Journal of Public Health*, Supplement 1, 2005)

Doubt Is Their Product, David Michaels (*Scientific American*, June, 2005)

Suggested Reading:
Suggested Reading on the Admissibility of Expert Testimony, The Federal Rules of Evidence, and Related Topics

Film: *West of Memphis* (2 hours)

The Federal Rules of Evidence: Introduction

A common aspect of many modern court proceedings, particularly in litigation involving toxic torts, product liability, contaminating environmental agents, and the like, is the presence of expert witnesses who provide evidence relevant to the matter at hand.

Very often such evidence is given through statistical argumentations, possibly through estimated statistical models (that might, for example, assess a probability of causation), meta-analyses, or other methods of data presentation and interpretation developed through graphical or tabular means.

To be defensible ethically, such evidence when presented statistically must avoid the various pitfalls we have discussed throughout (for example, regression toward the mean, the ecological fallacy, confusing test "impact" with test "bias," and "lurking" third variable confoundings).

The issues involved in expert witness admissibility, however, are much broader than just in how data are presented.

The scientific reliability and validity of the available evidence are of major interest and are subject to the *Federal Rules of Evidence*.

Here, we discuss some of the issues involved in the admissibility of evidence and the proffering of expert witnesses, both as they are currently understood and practiced and how they have evolved historically over the years.

Trial courts have had to evaluate the admissibility of expert testimony ever since a judicial system was established in the United States. Courts in the nineteenth and early twentieth centuries generally asked only whether an expert was "qualified" before expert testimony was considered admissible.

Whenever a subject was beyond the ken of an average juror, a qualified expert's opinion was considered crucial to a jury's determination of the facts at issue.

Usually, experts were qualified by dint of success in a relevant profession or occupation.

Nothing more than this was required, assuming that what was proffered was relevant to the case at hand;

the expertise or body of evidence admitted in trial was generally viewed as inseparable from the expert.

In the early 1920s, a ruling was made, commonly referred to as the *Frye* standard or opinion, that would frame expert witness admissibility for most of the twentieth century.

Even now the *Frye* standard may still hold sway, particularly for those states where the current *Federal Rules of Evidence* (FRE) have not been adopted for use in state courts.

In *Frye v. United States* (1923), the defendant offered an early form of a polygraph lie detection test (based on systolic blood pressure) to support a plea of innocence to a charge of murder.

The relevant part of the ruling, commonly referred to as the "general acceptance" standard, follows:

Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs.

The *Frye* opinion did several major new things.

First, an explicit separation was made between expertise and the expert, creating the important precedent that a body of asserted knowledge could exist apart from the proferring expert.

Second, expert testimony must arise from a knowledge base that has "gained general acceptance in the particular field to which it belongs."

Thus, various forms of pseudoscience (or in the modern parlance, "junk science") were inadmissible, irrespective of the person providing such information.

A reputation alone as a "hired gun" was not enough; the weapons had to be real and accepted in the specific field of interest.

The FRE govern how facts are admitted and how parties in federal courts of the United States may prove their cases.

These rules were the outcome of a long academic, legislative, and judicial process. They became federal law on January 2, 1975, when President Ford signed the *Act to Establish Rules of Evidence for Certain Courts and Proceedings*.

Rule 702 governs testimony by experts, and by most accounts, supersedes the *Frye* standard.

We give Rule 702 below, with a part indicated in brackets that was added (in 2000) in response to a Supreme Court decision that we discuss shortly (*Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993)):

Rule 702. Testimony by Experts:

If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, [if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case].

In the Supreme Court's decision in the case mentioned above, *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, the Court considered the standard for evaluating the admissibility of scientific expert testimony, and held that under the FRE (superseding *Frye*), trial court judges were the responsible "gatekeepers."

These trial judges were to evaluate the validity of the basis for the scientific expertise before the expert would be allowed to testify.

These pretrial determinations of admissibility are usually referred to as *Daubert* hearings (or Rule 104(a) hearings), in reference to the Supreme Court opinion (or in reference to the FRE).

The Supreme Court opinion in *Daubert v. Merrell* defined "scientific methodology" as the process of formulating hypotheses and conducting experiments to prove or falsify these hypotheses.

In the process, a "flexible" test was provided for establishing "validity" based on four (*Daubert*) factors:

(1) Empirical testing: the theory or technique must be falsifiable, refutable, and testable;
(2) Subject to peer review and publication;
(3) Known or potential error rate;
(4) The degree to which the theory and technique is generally accepted by a relevant scientific community.

Two additional Supreme Court opinions have further articulated the *Daubert* ruling:

General Electric Co. v. Joiner (1997), and *Kumho Tire Co. v. Carmichael* (1999).

The opinion in *General Electric v. Joiner* held that an abuse of discretion standard of review was the correct one for appellate courts to use in the review of a trial court's decision either to admit or not expert testimony.

The phrase "abuse of discretion" refers to a trial judge making an error in judgment that is clearly against the evidence or established law.

The Supreme Court ruling in *General Electric v. Joiner* held that the trial judge's exclusion of expert witness testimony did not constitute an abuse of discretion.

The third opinion in *Kumho Tire v. Carmichael* completes what is commonly named the *Daubert* trilogy.

Here, a trial judge's gatekeeping function, identified in *Daubert*, is extended to all expert testimony, including that which is putatively nonscientific.

Junk Science

One of the supposed downsides of the *Daubert* standard is that lay judges, now being the "gatekeepers" of scientific evidence, may prevent respected scientists from offering testimony.

As a consequence, corporate defendants are increasingly emboldened to accuse their adversaries of merely offering "junk science."

The label of "junk science" is easy to apply when there is a need to discount scientific findings that might be an impediment to short-term corporate profit.

Thus, Big Tobacco has applied the term to research on the harmful effects of smoking and second-hand smoke;

or the Fox News columnist, Steven Milloy, who applies the "junk science" label to research on many topics, including global warming, ozone depletion, DDT, Alar, and mad cow disease.

The Consequences of Daubert and the Data Quality Act (of 2001)

A spending bill that passed Congress in 2001, named the *Consolidated Appropriations Act*, included a brief two-sentence rider that has since become known (pretentiously) as the *Data Quality Act*.

This act directed the Office of Management and Budget (OMB) to develop government-wide guidelines:

The Director of the Office of Management and Budget shall . . . with public and Federal agency involvement, issue guidelines . . . that provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by Federal agencies.

In practice, the Data Quality Act has been used more as a ploy by corporations and their affiliates to suppress and/or delay the release of government reports that would be contrary to their collective economic interests.

We quote Chris Mooney from his 2005 book, *The Republican War on Science*:

As subsequently interpreted by the Bush administration . . . the so-called Data Quality Act creates an unprecedented and cumbersome process by which government agencies must field complaints over the data, studies, and reports they release to the public. It is a science abuser's dream come true.

Some Concluding Remarks

As we have maintained from the outset, a graduate course in statistics should prepare students in a number of areas that have immediate implications for the practice of ethical reasoning.

In these concluding slides, we review six broad topics that should be part of any competently taught sequence in the behavioral sciences:

(1) formal tools to help think through ethical situations;

(2) a basic understanding of the psychology of reasoning and how it may differ from that based on a normative theory of probability;

(3) how to be (dis)honest in the presentation of information, and to avoid obfuscation;

(4) some ability to ferret out specious argumentation when it has a supposed statistical basis;

(5) the deleterious effects of culling in all its various forms (for example, the identification of "false positives"), and the subsequent failures either to replicate or cross-validate;

(6) identifying plausible but misguided reasoning from data, or from other information presented graphically.

Doing the Math

One of the trite quantitative sayings that may at times drive individuals "up a wall" is when someone says condescendingly, "just do the math."

This saying can become a little less obnoxious when reinterpreted to mean working through a situation formally rather than just giving a quick answer based on first impressions.

An example of this may help.

In 1990, Craig Whitaker wrote a letter to Marilyn vos Savant's column in *Parade* magazine stating what has been named the Monty Hall problem:

Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, 'Do you want to pick door No. 2?' Is it to your advantage to switch your choice?

The answer almost universally given to this problem is that switching does not matter, presumably with the reasoning that there is no way for the player to know which of the two unopened doors is the winner, and each of these must then have an equal probability of being the winner.

By writing down three doors hiding one car and two goats, and working through the options in a short simulation, it becomes clear quickly that the opening of a goat door changes the information one has about the original situation, and that always changing doors doubles the probability of winning from 1/3 to 2/3.

Any beginning statistics class should always include a number of formal tools to help work through puzzling situations.

Several of these have been mentioned earlier:

Bayes' theorem and implications for screening using sensitivities, specificities, and prior probabilities; conditional probabilities more generally and how probabilistic reasoning might work for facilitative and inhibitive events;

sample sizes and variability in, say, a sample mean, and how a confidence interval might be constructed that could be made as accurate as necessary by just increasing the sample size, and without any need to consider the size of the original population of interest;

how statistical independence operates or doesn't;

the pervasiveness of natural variability and the use of simple probability models (such as the binomial) to generate stochastic processes;

the computations involved in corrections for attenuation;

the use of Taylor–Russell charts;

the formal distinction between (test) bias and impact, where the latter is not evidence of test "unfairness" per se.

Reasoning Heuristics

A second area of interest in developing statistical literacy and learning to reason ethically is the large body of work produced by psychologists.

This work compares the normative theory of choice and decisions derivable from probability theory, and how this may not be the best guide to the actual reasoning processes individuals use.

The contributions of Tversky and Kahneman (for example, 1971, 1974, 1981) are particularly germane to our understanding of reasoning.

People rely on various simplifying heuristic principles to assess probabilities and engage in judgments under uncertainty.

The representativeness heuristic operates where probabilities are evaluated by the degree to which A is representative of B;

if highly representative, the probability that A originates from B is assessed to be higher.

When representativeness heuristics are in operation, a number of related characteristics of the attendant reasoning processes become apparent:
prior probabilities (baserates) are ignored;
insensitivity develops to the operation of sample size on variability;
an expectation that a sequence of events generated by some random process, even when the sequence is short, will still possess all the essential characteristics of the process itself.

This leads to the "gambler's fallacy" (or, "the doctrine of the maturity of chances"), where certain events must be "due" to bring the string more in line with representativeness; as one should know, corrections are not made in a chance process but only diluted as the process unfolds.

When a belief is present in the "law of small numbers," even small samples must be highly representative of the parent population; thus, researchers put too much faith in what is seen in small samples and overestimate replicability.

Also, people may fail to recognize regression toward the mean because predicted outcomes should be maximally representative of the input and therefore be exactly as extreme.

A second powerful reasoning heuristic is *availability*. We quote from Tversky and Kahneman (1974):

Lifelong experience has taught us that, in general, instances of large classes are recalled better and faster than instances of less frequent classes; that likely occurrences are easier to imagine than unlikely ones; and that the associative connections between events are strengthened when the events frequently co-occur. As a result, man has at his disposal a procedure (the availability heuristic) for estimating the numerosity of a class, the likelihood of an event, or the frequency of co-occurrences, by the ease with which the relevant mental operations of retrieval, construction, or association can be performed.

When required to reason about an individual's motives in some ethical context, it is prudent to remember the operation of the *fundamental attribution error*, where people presume that actions of others are indicative of the true ilk of a person, and not just that the situation compels the behavior.

As one example from the courts, even when confessions are extracted that can be demonstrably shown false, there is still a greater likelihood of inferring guilt compared to the situation where a false confession was not heard.

A particulary egregious example of making the fundamental attribution error (and moreover, for nefarious political purposes), is Liz Cheney and her ad on the website "Keep America Safe" regarding those lawyers currently at the Justice Department who worked as advocates for "enemy combatants" at Guantanamo Bay, Cuba.

The Presentation of Data

The presentation of data is an obvious area of concern when developing the basics of statistical literacy.

Some aspects may be obvious, such as not making up data or suppressing analyses or information that don't conform to prior expectations.

At times, however, it is possible to contextualize (or to "frame") the same information in different ways that might lead to differing interpretations.

When data are presented to make a health-related point, it is common practice to give the argument in terms of a "surrogate endpoint."

Instead of providing direct evidence based on a clinically desired outcome (for example, if you engage in this recommended behavior, the chance of dying from, say, a heart attack is reduced by such and such amount), the case is stated in terms of a proxy (for example, if you engage in this recommended behavior, your cholesterol levels will be reduced).

In general, a surrogate end point or biomarker is a measure of a certain treatment that may correlate with a real clinical endpoint, but the relationship is not guaranteed.

This caution can be rephrased as "a correlate does not a surrogate make."

It is a common misconception that something correlated with the true clinical outcome must automatically then be usable as a valid surrogate end point and can act as a proxy replacement for the clinical outcome of primary interest.

As is true for all correlational phenomena, causal extrapolation requires further argument.

In this case, it is that the effect of the intervention on the surrogate directly predicts the clinical outcome.

Obviously, this is a more demanding requirement.

Identifying Illegitimate Claims

A fourth statistical literacy concern is to have enough of the formal skills and context to separate legitimate claims from those that might represent more specious arguments.

As examples, one should recognize when a case for cause is made in a situation for which regression toward the mean is as likely an explanation, or when test unfairness is argued for based on differential performance (that is, impact) and not on actual test bias (that is, same ability levels performing differently).

A more recent illustration of the questionable promotion of a methodological approach, called optimal data analysis (ODA), is given by Yarnold and Soltysik (2004). We quote from their preface:

[T]o determine whether ODA is the appropriate method of analysis for any particular dataset, it is sufficient to consider the following question: When you make a prediction, would you rather be correct or incorrect? If your answer is "correct," then ODA is the appropriate analytic methodology—by definition. That is because, for any given dataset, ODA explicitly obtains a statistical model that yields the theoretical maximum possible level of predictive accuracy (for example, number of correct predictions) when it is applied to those data. That is the motivation for ODA; that is its purpose. Of course, it is a matter of personal preference whether one desires to make accurate predictions. In contrast, alternative non-ODA statistical models do not explicitly yield theoretical maximum predictive accuracy. Although they sometimes may, it is not guaranteed as it is for ODA models. It is for this reason that we refer to non-ODA models as being *suboptimal*.

Sophistic arguments such as these have no place in the methodological literature (even when a text, such as this one, has been reviewed and published by the American Psychological Association).

It is inappropriate to call one's method "optimal" and refer pejoratively to others as therefore "suboptimal."

The simplistic approach to classification underlying "optimal data analysis" is known not to cross-validate well (see, for example, Stam, 1997);

it is a large area of operations research where the engineering effort is always to squeeze a little more out of an observed sample.

What is most relevant in the behavioral sciences is stability and cross-validation (of the type reviewed in Dawes [1979] on proper and improper linear models);

and to know which variables discriminate and how, and to thereby "tell the story" more convincingly and honestly.

False Positives

The penultimate area of review is a reminder of the ubiquitous effects of searching/selecting/optimization, and the identification of "false positives."

We have mentioned some blatant examples earlier—the weird neuroscience correlations; the small probabilities (mis)reported in various legal cases (such as the Dreyfus small probability for the forgery coincidences, or that for the de Berk hospital fatalities pattern); repeated clinical experimentation until positive results are reached in a drug trial—but there are many more situations that would fail to replicate.

We need to be ever-vigilant of results obtained by "culling" and then presented as evidence.

A general version of the difficulties encountered when results are culled is labeled the *file-drawer problem*.

This refers to the practice of researchers putting away studies with negative outcomes (that is, studies not reaching reasonable statistical significance or when something is found contrary to what the researchers want or expect, or those rejected by journals that will consider publishing only articles demonstrating significant positive effects).

The file-drawer problem can seriously bias the results of a meta-analysis, particularly if only published sources are used (and not, for example, unpublished dissertations or all the rejected manuscripts lying on a pile in someone's office).

The subtle effects of culling with subsequent failures to replicate can have serious consequences for the advancement of our understanding of human behavior.

A recent important case in point involves a gene–environment interaction studied by a team led by Avshalom Caspi (Caspi et al., 2003).

A polymorphism related to the neurotransmitter serotonin was identified that apparently could be triggered to confer susceptibility to life stresses and resulting depression.

Needless to say, this behavioral genetic link caused quite a stir in the community devoted to mental health research.

Unfortunately, the result could not be replicated in a subsequent meta-analysis (could this possibly be due to the implicit culling over the numerous genes affecting the amount of serotonin in the brain?).

Argument From Ignorance

Our final statistical literacy issue is the importance of developing abilities to spot and avoid falling prey to the trap of specious reasoning known as an "argument from ignorance," or *argumentum ad ignorantiam*, where a premise is claimed to be true only because it has not been proven false, or that it is false because it has not been proven true.

Sometimes this is also referred to as "arguing from a vacuum", where what is purported to be true is supported not by direct evidence but by attacking an alternative possibility.

Thus, a clinician might say: "because the research results indicate a great deal of uncertainty about what to do, my expert judgment can do better in prescribing treatment than these results."

Or to argue that people "need" drugs just because they haven't solved their problems before taking them.

In making policy decisions based on arguments of causality in areas such as medicine and the environment, it is best to remember the *precautionary principle*:

whenever a policy or action has the potential of causing harm, say, to the environment or people, and a scientific consensus does not exist that the policy or action is harmful, the burden of proof that it is *not* harmful rests with those wishing to take the action.

A related fallacy is *argument from personal incredulity*, where because one personally finds a premise unlikely or unbelievable, the premise can be assumed false, or that another preferred but unproven premise is true instead.

In both of these instances, a person regards the lack of evidence for one view as constituting proof that another is true.

Related fallacies are (a) the *false dilemma* where only two alternatives are considered when there are, in fact, other options.

The famous Eldridge Cleaver quotation from his 1968 presidential campaign is a case in point: "You're either part of the solution or part of the problem."

Or, (b) the Latin phrase *falsum in uno, falsum in omnibus* (false in one thing, false in everything) implying that someone found to be wrong on one issue, must be wrong on all others as well.

In a more homey form, "when a clock strikes thirteen, it raises doubt not only to that chime, but to the twelve that came before."

Unfortunately, we may have a current example of this in the ongoing climate-change debate that we have discussed earlier—

the one false statistic proffered by a 2007 report from the Intergovernmental Panel on Climate Change (IPCC) on Himalayan glacier melt may serve to derail the whole science-based argument that climate change is real.

There are several fallacies with a strong statistical tinge related to *argumentum ad ignorantiam*.

One is the "margin of error folly,": if it could be, it is.

Or, in a hypothesis-testing context, if a difference isn't significant, it is zero.

It is important not to confuse a statement of "no evidence of an effect" with one of "evidence of no effect."

We now can refer to all of these reasoning anomalies under the umbrella term "truthiness," coined by Stephen Colbert from Comedy Central's, *The Colbert Report*.

Here, truth comes from the gut, not books, and refers to the preferring of concepts or facts one wishes to be true, rather than known to be true.

For example, in 2009 we had the "birthers," who claimed that Barack Obama was not born in the United States, so constitutionally he cannot be President; or that the Health Care Bill included "death squads" ready to "pull the plug on granny," or earlier in the 2000s, there were weapons of mass destruction that justified the Iraq war; and on and on.

In developing skills to avoid specious reasoning, or to be able to label such reasoning as fallacious, it is helpful to have a number of useful words and phrases at one's disposal. We give a representative few here, with others used in context throughout the book:

ex cathedra: spoken with authority; with the authority of the office. From Latin *ex cathedra* (from the chair), from *cathedra* (chair). In the Roman Catholic Church, when the Pope speaks *ex cathedra* he is considered infallible. The word cathedral is short for the full-term cathedral church, meaning the principal church of a diocese, one containing a bishop's throne. The term is often used ironically or sarcastically to describe self-certain statements, alluding to the Pope's supposed infallibility, as if an office or position conferred immunity from error.

ipse dixit: from the Latin, "he himself said it"; a statement asserted but not proved; to be accepted on faith; a *dictum*.

obiter dictum: a comment or remark made in passing, particularly by a judge on an issue not directly relevant to the case at hand; from the Latin, "something said in passing".

via regia: a reference to an imperial and ancient road; from the Latin, "King's Road".

chimerical: created by fanciful imagination; highly improbable.

equivocal: capable of differing interpretations; ambiguous.

equivocate: to use equivocal language intentionally, and avoid making an explicit statement.

ad hominem: appealing to one's prejudices, emotions, or other personal considerations rather than to intellect or reason. Attacking an opponent personally instead of countering the argument. From Latin, literally "to the person".