Data Presentation and Interpretation

Psychology (Statistics) 484

Statistics, Ethics, and the Social and Behavioral Sciences

June 17, 2013

Beginning Quotations

What we got here is . . . failure to communicate.
– The Captain in *Cool Hand Luke*

Ranges are for cattle! Give me a number.
– Lyndon Baines Johnson

For ye shall know the truth, and the truth shall set you free.
– Motto of the CIA (from John 8:31–32)

Data! Data! Data! I can't make bricks without clay.
– Sir Arthur Conan Doyle (*The Adventures of the Copper Beeches*)

Week 8: Data Presentation and Interpretation

— weight-of-the-evidence argumentation in presenting and interpreting data, particularly for medical and regulatory issues

— Brown v. Board of Education (1954): Segregation of students in public schools violates the Equal Protection Clause of the Fourteenth Amendment, because separate facilities are inherently unequal.

— Matrixx Initiatives, Inc. v. Siracusano (2011): A plaintiff may state a claim against a pharmaceutical company for securities fraud under the Securities Exchange Act of 1934 based on the company's failure to disclose reports of adverse events even when the reports do not disclose a "statistically significant" number of such adverse events.

Required Reading:
SGEP (259–282) —
Weight-of-the-Evidence Arguments in the Presentation and Interpretation of Data
A Case Study in Data Interpretation: Brown v. Board of Education (1954)
A Case Study in Data Interpretation: Matrixx Initiatives, Inc. v. Siracusano (2011)

Popular Articles —
Head Case: Can Psychiatry Be a Science? Louis Menand (*New Yorker*, March 1, 2010)

Talking Back to Prozac, Frederick C. Crews (*New Yorker*, December 6, 2007)
Do We Really Know What Makes Us Healthy? Gary Taubes (*New York Times*, September 16, 2007)
The Plastic Panic, Jerome Groopman (*New Yorker*, May 31, 2010)
John Rock's Error, Malcolm Gladwell (*New Yorker*, March 10, 2000)

Suggested Reading:
Suggested Reading on Data Presentation and Interpretation
Appendix: Brown v. Board of Education (Decided: May 17, 1954)
Appendix: Matrixx Initiatives, Inc. v. Siracusano (Decided: March 22, 2011)

Film:

The Manhattan Project (42 minutes)

The Town That Never Was (16 minutes)

Introduction

The goal of statistics is to gain understanding from data.

The methods of presentation and analyses used should not only allow us to "tell the story" in the clearest and fairest way possible, but more primarily, to help uncover what the story is in the first place.

When results are presented, there is a need to be sensitive to the common and perhaps not-so-common missteps that result from a superficial understanding and application of the methods in statistics.

It is insufficient just to "copy and paste" without providing context for how good or bad the methods are that are being used, and understanding what is behind the procedures producing the numbers.

Smaller Statistical Missteps to Avoid

(1) Even trivial differences between groups will be statistically significant when sample sizes are large. Significance should never be confused with importance;

the current emphasis on the use of confidence intervals and the reporting of effect sizes reflects this point.

Conversely, lack of statistical significance does not mean that the effect is therefore zero.

Such a confusion was behind the ongoing debacle for Vioxx, the now withdrawn pain killer believed responsible for thousands of deaths by lethal heart attack.

(2) As some current textbooks still report inappropriately, a significance test does not evaluate whether a null hypothesis is true. A *p*-value measures the "surprise value" of a particular observed result conditional on the null hypothesis being true.

(3) Degrees of freedom do not refer to the number of independent observations within a dataset. The term indicates how restricted the quantities are that are being averaged in computing various statistics, such as sums of squares between or within groups.

(4) Although the central limit theorem justifies assertions of robustness when dealing with means, the same is not true for variances. The common tests on variances are notoriously nonrobust and should not be used; robust alternatives are available in the form of sample reuse methods such as the jackknife and the bootstrap.

(5) Do not carry out a test for equality of variances before performing a two-independent samples t-test. A quotation from George Box contrast the good robustness properties of the t-test with the nonrobustness of the usual tests for variances:

"To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port.

(6) Measures of central tendency and dispersion, such as the mean and variance, are not *resistant* in that they are influenced greatly by extreme observations. The median and interquartile range, on the other hand, are resistant, and each observation counts the same in the calculation of the measure.

In reporting an "average," it is important to distinguish between the types that could be given. For example, an arithmetic mean may provide one form for what is considered "average," but it may not be what is also "typical." For that, the mode or median could be a better means of communication. This can be especially true when dealing with money having underlying skewed distributions, such as for average tax cuts, tax increases, salaries, and home sale prices.

(7) Do not ignore the repeated-measures nature of your data, and use methods appropriate for independent samples. For example, don't perform an independent samples t-test on "before" and "after" data in a time-series intervention study. Generally, the standard error of a mean difference must include a correction for correlated observations, as is routinely done in a paired (matched samples) t-test.

(8) The level of measurement used for your observations limits the inferences that are meaningful. For example, interpreting the relative sizes of differences makes little sense on nominal or ordinal data. Also, performing arithmetic operations on data that are nominal, or at best ordinal, may produce inappropriate descriptive interpretations as well.

(9) Do not issue blanket statements as to the impossibility of carrying out reasonable testing, confidence interval construction, or cross-validation. It is almost always now possible to use resampling methods that do not rely on parametric models or restrictive assumptions, and which are computer-implemented for immediate application.

(10) Keep in mind the distinctions between fixed and random effects models and the differing test statistics they may necessitate. The output from some statistical package may use a default understanding of how the factors are to be interpreted. If your context is different, then appropriate calculations must be made, sometimes "by hand." To parody the Capital One Credit Card commercial: "What's in your denominator?"

(11) Do not report all of the eight or so decimal places given in typical computer output. Two decimal places are needed at most, and often, one is all that is really justified.

As an example, consider how large a sample is required to support the reporting of a correlation to more than one decimal place (answer: given the approximate standard error of $\frac{1}{\sqrt{n}}$, a sample size greater than 400 would be needed to give a 95% confidence interval of $\pm.1$).

Several quotations given below all pertain to the issue of approximation and accuracy:

It is the mark of an educated man to look for precision in each class of things just so far as the nature of the subject admits; it is evidently equally foolish to accept probable reasoning from a mathematician and to demand from a rhetorician scientific proofs.
– Aristotle, *Nicomachean Ethics*

A little inaccuracy sometimes saves tons of explanation.
– H. H. Munro (as Saki)

Truth is much too complicated to allow anything but approximations.
– John von Neumann

Although this may seem a paradox, all exact science is dominated by the idea of approximation.
– Bertrand Russell

Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise.
– John W. Tukey

(12) It is generally wise to avoid issuing statements that might appear to be right but with some deeper understanding are just misguided:

(a) "Given the size of a population, it is impossible to achieve accuracy with a sample"; this reappears regularly with the discussion of undercount and the census.

(b) "Always divide by $n-1$ when calculating a variance to give the 'best' estimator"; if you divide by n or $n+1$, the estimator has a smaller expected error of estimation, which to many is more important than being "unbiased."

Also, we note that no one ever really worries that the usual correlation coefficient is a "biased" estimate of its population counterpart.

c) "ANOVA is so robust that all of its assumptions can be violated at will"; although it is true that normality is not crucial if sample sizes are reasonable in size (and the central limit theorem is of assistance), and homogeneity of variances doesn't really matter as long as cell sizes are close, the independence of errors assumption is critical, and one can be led far astray when it doesn't hold; for example, in intact groups, spatial contexts, and repeated measures.

(d) Don't lament the dearth of one type of individual from the very upper scores on some test without first noting possible differences in variability. Even though mean scores may be the same for groups, those with even slightly larger variances will tend to have more representatives in both the upper and lower echelons.

(13) Unless a compelling reason exists, avoid using one-tailed tests.

Even the mechanisms for carrying out traditional one-tailed hypothesis tests, the chi-square and F distributions, have two tails, and both ought to be considered.

The logic of hypothesis testing is that if an event is sufficiently unlikely, we must reconsider the truth of the null hypothesis.

Thus, for example, if an event falls in the lower tail of the chi-square distribution, it implies that the model fits too well.

If investigators had used two-tailed tests, the data fabrications of Cyril Burt might have been uncovered much earlier (see Dorfman, 1978).

(14) A confidence level does not give the probability that repeated estimates would fall into that particular confidence interval.

Instead, a confidence level is an indication of the proportion of time the intervals cover the true value of the parameter under consideration if we repeat the complete confidence interval construction process.

Some Comments on PowerPoint

In concluding these introductory comments about the smaller missteps to be avoided, we note the observations of Edward Tufte on the ubiquity of PowerPoint (PP) for presenting quantitative data, and the degradation it produces in our ability to communicate (italics in the original):

The PP slide format has the worst signal/noise ratio of any known method of communication on paper or computer screen. Extending PowerPoint to embrace paper and internet screens pollutes those display methods.

Generally, PowerPoint is poor at presenting statistical evidence, and is not a good replacement for technical reports, numerical data presented in detailed handouts, and the like.

It is now part of our "pitch culture," where, for example, we are sold on what drugs to take, but are not provided with the type of detailed numerical evidence we should have for an informed decision about benefits and risks.

In commenting on the incredible obscuration of important (numerical) data that surrounded the use of PowerPoint-type presentations in the crucial briefings on the first Shuttle accident of Challenger in 1986, Richard Feynman noted:

Then we learned about 'bullets'—little black circles in front of phrases that were supposed to summarize things. There was one after another of these little goddamn bullets in our briefing books and on slides.

Weight-of-the-Evidence Arguments

Before discussing how weight-of-the-evidence (WOE) arguments might be framed, we start with several admonitions regarding what data should be presented in the first place.

To begin, it is questionable professionally to engage in "salami science," where a single body of work is finely subdivided into "least publishable units," known by its acronym of LPUs.

To avoid such salami science, an emphasis exists in the better publication outlets for the behavioral sciences on the combined reporting of multiple studies delineating a common area so that a compelling WOE argument might be constructed.

Second, in medically related studies, the underreporting of research can be seen as scientific misconduct. As noted in the article by Iain Chalmers (1990, "Underreporting Research is Scientific Misconduct," *Journal of the American Medical Association*, *263*, 1405–1408), the results of many clinical trials never appear in print, and among many of those that do, there is insufficient detail to assess the validity of the study.

The consequences of underreporting can be serious. It could compromise treatment decisions for patients; do injustice to those patients who participated in the trials; and waste scarce resources and funds available for medically relevant trials.

As Chalmers remarks: "Studies should be accepted or rejected on the basis of whether they have been well conceptualized and competently executed, not on the basis of the direction or magnitude of any differences observed between comparison groups" .

A final admonition is noted in a 2011 article by Simmons, Nelson, and Simonsohn, "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant" (*Psychological Science*, *22*, 1359–1366).

Simmons et al. (2011) provide several suggestions for authors and reviewers to mitigate the problem of false-positive publication;

we paraphrase these recommendations in the required reading.

For those disciplines of primary relevance to us—the social and behavioral sciences, and the health and medical fields—it is rare, if ever, to have a crucial experiment that would decisively resolve an issue at hand.

Instead, we hear more about the "weight of the evidence" (WOE) being the eventual decider.

In law, the WOE may merely refer to preponderance; that is, something that is more likely than not.

In other areas, however, a WOE argument is usually advanced to make a (causal) claim, or to indicate that the risks far outweigh the benefits of a medical procedure, drug, supplement, pesticide, and so on.

In making decisions it is important to consider how big the potential benefits are, what the risks and costs might be, and what the weight of the evidence really shows.

In issues of toxicology, for example, we have the famous quotation from Paracelsus (1493–1541):

German: Alle Ding' sind Gift, und nichts ohn' Gift; allein die Dosis macht, dass ein Ding kein Gift ist.

(All things are poison and nothing is without poison, only the dose permits something not to be poisonous.)

Historic WOE Arguments

Historically, there have been many changed practices due to an eventual WOE argument, even though possibly no identifiable group of individuals ever made it explicit. We give several examples of this below:

(1) The Halsted radical mastectomy surgery is a procedure for breast cancer where the breast, underlying chest muscle, and lymph nodes of the axilla are removed.

It was developed by William Halsted in 1882. About 90% of women treated for breast cancer in the United States from 1895 to the mid-1970s had radical mastectomies.

This morbid surgery is now rarely performed except in extreme cases.

(2) Phrenology is a system where the personality traits of a person can be inferred from the shape of the skull.

Until the depression era of the 1930s, phrenology was a widely held "theory".

The phrenological argument was straightforward:

people have diverse mental capacities or faculties, localized in the brain. The strength of faculty expression was related to the size of that part of the brain, in turn affecting the contours on the surface of the skull.

(3) In the 1940s, William Herbert Sheldon pioneered the use of anthropometry to categorize people into somatotypes: endomorphic (soft and round), mesomorphic (stocky and muscular), and ectomorphic (thin and fragile).

People could be graded on one-to-seven point scales as to the degree they exhibited each of the three somatotypes.

For example, a "pure" mesomorph would be a 1-7-1, a "pure" ectomorph, a 1-1-7, and so on.

Sheldon divided personality characteristics into three categories, where each body type had a corresponding personality profile: endotonia (physical comfort, food, and socializing); mesotonia (physical action and ambition); ectotonia (privacy and restraint).

Sheldon saw a strong correlation between mesotonia and mesomorphs, and concluded that these mesotonic individuals would descend into criminality.

Related to Sheldon's connection between somatotypes and personality, there is the much earlier (late 1800s) Cesare Lombrosco theory of anthropological criminality.

This idea held that body characteristics and criminality are connected, and "born criminals" could be identified by physical defects (or stigmata), thus confirming a criminal as savage, or atavistic (that is, a throwback).

(4) Graphology is the study and analysis of handwriting in relation to human psychology and personality assessment. Given the paucity of empirical studies that show any validity, it is generally now considered a pseudoscience.

(5) Polygraph lie detection has much the same status as graphology.

A perusal of the National Research Council report, *The Polygraph and Lie Detection* (2003), should be enough to place polygraph examination into the same category as phrenology.

6) The practice of bloodletting was the most common medical procedure performed by doctors from antiquity until the early 20th century.

Barber poles, with the intermixture of white and red, signify where bloodletting could be performed.

Leeches also became popular in the early 19th century, with hundreds of millions being used by physicians throughout the century.

(7) Many health-related and other beliefs and practices have been debunked by WOE arguments.

One practice no longer engaged in because of a WOE understanding of radiation risk was in buying new shoes for school during the 1950s.

Typically, a shoe-fitting fluoroscope was available, where you literally stood on top of an x-ray tube (with all of your organs exposed), to see how your foot bones wiggled nicely in your new shoes.

And possibly, once your shoes were bought, you could then play on the machine somewhat longer while your mother tried on shoes for herself.

As another (funny) example, was once believed that listening to classical music could make you smarter – the "Mozart Effect"

A short section from the Wikipedia article on the political impact of the Mozart effect, follows:

The popular impact of the theory was demonstrated on January 13, 1998, when Zell Miller, governor of Georgia, announced that his proposed state budget would include $105,000 a year to provide every child born in Georgia with a tape or CD of classical music. Miller stated "No one questions that listening to music at a very early age affects the spatial-temporal reasoning that underlies math and engineering and even chess."

Miller played legislators some of Beethoven's "Ode to Joy" on a tape recorder and asked "Now, don't you feel smarter already?" Miller asked Yoel Levi, music director of the Atlanta Symphony, to compile a collection of classical pieces that should be included. State representative Homer M. DeLoach said "I asked about the possibility of including some Charlie Daniels or something like that, but they said they thought the classical music has a greater positive impact. Having never studied those impacts too much, I guess I'll just have to take their word for that."

A Case Study in Data Interpretation: Brown v. Board of Education (1954)

Brown v. Board of Education (1954) was a landmark decision of the United States Supreme Court, declaring that state laws establishing separate public schools for black and white students were unconstitutional and violated the Equal Protection Clause of the Fourteenth Amendment.

The Warren Court's unanimous (9–0) decision stated categorically that "separate educational facilities are inherently unequal," thus overturning the 1896 (7–1) decision in *Plessy v. Ferguson* (1896).

This turn-of-the-century ruling held that the "separate but equal" provision of private services mandated by state government is constitutional under the Equal Protection Clause.

The crucial difference in deciding *Brown v. Board of Education* in 1954 versus *Plessy v. Ferguson* in 1896 was the wealth of dispositive behavioral science research and data then available on the deleterious effects on children of segregation and the philosophy of "separate but equal."

As stated by Chief Justice Earl Warren in the unanimous opinion: "To separate them from others of similar age and qualifications solely because of their race generates a feeling of inferiority as to their status in the community that may affect their hearts and minds in a way unlikely to ever be undone."

The experimental data collected by Kenneth and Mamie Clark, in particular, were central to the Court's reasoning and decision; the Clark "doll test" studies were especially convincing as to the effects that segregation had on black school children's mental status.

A Case Study in Data Interpretation: Matrixx Initiatives, Inc. v. Siracusano (2011)

We conclude with a lengthy redaction from the recent Supreme Court case of *Matrixx Initiatives, Inc. v. Siracusano* (2011).

It is a remarkable example of cogent causal and statistical reasoning.

Although assisted by many "friend of the court" briefs, and probably more than a few bright law clerks, this opinion delivered by the "wise Latina woman," Sonia Sotomayor, is an exemplary presentation of a convincing WOE argument.

It serves as a model for causal reasoning in the presence of a "total mix" of evidence.

The Matrixx case involved the nasal spray and gel Zicam manufactured by Matrixx Initiatives, Inc.

From 1999 to 2004, the company received reports that Zicam might have led users to suffer the "adverse event" of losing a sense of smell—anosmia.

The company did not disclose these adverse event reports to potential investors under the excuse they were not "statistically significant."

There was no attempt to explain how "statistically significance" could even be determined given that no "control group" was available for comparison.

The suit against Matrixx Initiatives was for securities fraud and for making statements that omitted material information, defined as information that reasonable investors would consider as substantially altering the "total mix" of available information.

Here, "total mix" refers to all the information typically relied on in a WOE argument.

In arguing their position, Matrixx commented that in the several randomized clinical trials carried out for Zicam, adverse effects of anosmia did not appear (or, at least, the number of adverse events were not statistically different from those identified in the control groups).

What Matrixx did not acknowledge is that the small sample sizes of the clinical trials may well have failed to identify any rare events.

But once on the market and used by more individuals, rare events might well happen, and in relatively substantial numbers given the size of the treatment base.

These post-marketing or phase IV trials are supposed to monitor the continued safety of over-the-counter and prescription products available to the public.

The FDA has an automated reporting system in place (the Adverse Events Reporting System (AERS)), that tracks adverse events.

There is no need to invoke the idea of "statistical significance."

Indeed, it is simply not applicable or even computable in this case.

Once enough events get reported and a credible causal argument made, the product might be recalled or, at least, additional warning labels included—

think of Avendia, Vioxx, or any of a number of prescription drugs now absent from the market.