Statistical Sleuthing and Explanation

Psychology (Statistics) 484

Statistics, Ethics, and the Social and Behavioral Sciences

June 19, 2013

Beginning Quotations

The human understanding when it has once adopted an opinion . . . draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside or rejects . . . in order that by this great and pernicious predeterminations, the authority of its former conclusions may remain inviolate.
– Francis Bacon, *Novum Organum* (1620)

Faced with the choice between changing one's mind and proving that there is no need to do so, almost everyone gets busy on the proof.
– John Kenneth Galbraith

Week 12: Statistical Sleuthing and Explanation

— statistical sleuthing with formal models: Poisson clumping, Benford's law, survival analysis, Kaplan–Meier curves

— McCleskey v. Kemp (1987): Despite statistical evidence of a profound racial disparity in application of the death penalty, such evidence is insufficient to invalidate defendant's death sentence

Required Reading:
SGEP (359–384) —
Sleuthing Interests and Basic Tools
Survival Analysis
Statistical Sleuthing and the Imposition of the Death Penalty: McCleskey v. Kemp (1987)

Popular Articles —
The Treatment, Malcolm Gladwell (*New Yorker*, May 17, 2010)
The Ghost's Vocabulary: How the Computer Listens for Shakespeare's "Voiceprint," Edward Dolnick (*The Atlantic*, October, 1991)

Suggested Reading:
Suggested Reading on Statistical Sleuthing
Appendix: U.S. Supreme Court, McCleskey v. Kemp (Decided: April 22, 1987): Majority Opinion and Dissent

Film: *A Cry in the Dark* (121 minutes)

Introduction

Some of the more enjoyable intellectual activities statisticians engage in might be called *statistical sleuthing*—the use of various statistical techniques and methods to help explain or "tell the story" about some given situation.

We first give a flavor of several areas where such sleuthing has been of explanatory assistance:

(a) The irregularities encountered in Florida during the 2000 Presidential election and why; see, for example, Alan Agresti and Brett Presnell, "Misvotes, Undervotes, and Overvotes: The 2000 Presidential Election in Florida" (*Statistical Science*, *17*, 2002, 436–440).

(b) The attribution of authorship for various primary sources; for example, we have the seminal work by Mosteller and Wallace (1964) on the disputed authorship of some of the Federalist Papers.

(c) Searching for causal factors and situations that might influence disease onset; for example, "Statistical Sleuthing During Epidemics: Maternal Influenza and Schizophrenia" (Nicholas J. Horton & Emily C. Shapiro, *Chance*, *18*, 2005, 11–18);

(d) Evidence of cheating and corruption, such as the Justin Wolfers (2006) article on point shaving in NCAA basketball as it pertains to the use of Las Vegas point spreads in betting (but, also see the more recent article by Bernhardt and Heston [2010] disputing Wolfers' conclusions);

(e) The observations of Quetelet's from the middle 1800s that based on the very close normal distribution approximations for human characteristics, there were systematic understatements of height (to below 5 feet, 2 inches) for French conscripts wishing to avoid the minimum height requirement needed to be drafted;

(f) Defending someone against an accusation of cheating on a high-stakes exam when the "cheating" was identified by a "cold-hit" process of culling for coincidences, and with subsequent evidence provided by a selective search (that is, a confirmation bias). A defense that a false positive has probably occurred requires a little knowledge of Bayes' theorem and the positive predictive value.

(g) Demonstrating the reasonableness of results that seem "too good to be true" without needing an explanation of fraud or misconduct. An exemplar of this kind of argumentation is in the article, "A Little Ignorance: How Statistics Rescued a Damsel in Distress" (Peter Baldwin and Howard Wainer, *Chance*, 2009, *22*, 51–55).

Some Tools

A variety of sleuthing approaches are available to help explain what might be occurring over a variety of different contexts.

Some of these have been introduced already:

Simpson's Paradox,
Bayes' rule and baserates,
bounds provided by corrections for attenuation,
regression toward the mean,
the effects of culling on the identification of false positives and the subsequent inability to cross-validate,
the ecological fallacy,

the operation of randomness and the difficulty in "faking" such a process,
confusions caused by misinterpreting conditional probabilities,
illusory correlations,
restrictions of range for correlations,
and so on.

We mention a few other tools below that may provide some additional assistance:

the use of various discrete probability distributions, such as the binomial, Poisson, or those for runs, in constructing convincing explanations for some phenomena;
the digit regularities suggested by what is named Benford's law;
a reconception of some odd probability problems by considering pairs (what might be labeled as the "the birthday probability model";
and the use of the statistical techniques in survival analysis to model time-to-event processes.

Inspection Paradox

There are several quantitative phenomena useful in sleuthing but which are less than transparent to understand.

One particularly bedeviling result is called the Inspection Paradox.

Suppose a light bulb now burning above your desk (with an average rated life of, say, 2000 hours), has been in operation for a year.

It now has an expected life longer than 2000 hours because it has already been on for a while, and therefore cannot burn out at any earlier time than right now.

The same is true for life spans in general. Because we have not, as they say, "crapped out" as yet, and we cannot die at any earlier time than right now, our lifespans have an expectancy longer than what they were when we were born.

The Binomial Distribution

The simplest probability distribution has only two event classes (for example, success/fail, live/die, head/tail, 1/0).

A process that follows such a distribution is called Bernoulli; typically, our concern is with repeated and independent Bernoulli trials.

Using an interpretation of the two event classes of heads (H) and tails (T), assume $P(H) = p$ and $P(T) = 1 - p$, with p being invariant over repeated trials (that is, the process is stationary).

The probability of any sequence of size n that contains k heads and $n - k$ tails is $p^k(1 - p)^{n-k}$.

Commonly, our interest is in the distribution of the number of heads (say, X) seen in the n independent trials.

This random variable follows the binomial distribution:

$$P(X = r) = \binom{n}{r} p^r (1-p)^{n-r} \ ,$$

where $0 \leq r \leq n$, and $\binom{n}{r}$ is the binomial coefficient:

$$\binom{n}{r} = \frac{n!}{(n-r)!r!} \ ,$$

using the standard factorial notation.

Both the binomial distribution and the underlying repeated Bernoulli process offer useful background models against which to compare observed data, and to evaluate whether a stationary Bernoulli process could have been responsible for its generation.

The Poisson Distribution

A number of different discrete distributions prove useful in statistical sleuthing.

We mention two others here, the Poisson and a distribution for the number of runs in a sequence. A discrete random variable, X, that can take on values 0, 1, 2, 3, . . . , follows a Poisson distribution if

$$P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!} ,$$

where λ is an intensity parameter, and r can take on any integer value from 0 onward.

Although a Poisson distribution is usually considered a good way to model the number of occurrences for rare events, it also provides a model for spatial randomness as the example adapted from Feller (1968, Vol. 1, pp. 160–161) illustrates:

Flying-bomb hits on London. As an example of a spatial distribution of random points, consider the statistics of flying-bomb hits in the south of London during World War II.

The entire area is divided into 576 small areas of 1/4 square kilometers each.

Table 14.1 (in your required reading) records the number of areas with exactly k hits.

The total number of hits is 537, so the average is .93 (giving an estimate for the intensity parameter, λ).

The fit of the Poisson distribution is surprisingly good. As judged by the χ^2-criterion, under ideal conditions, some 88 per cent of comparable observations should show a worse agreement.

It is interesting to note that most people believed in a tendency of the points of impact to cluster.

If this were true, there would be a higher frequency of areas with either many hits or no hits and a deficiency in the intermediate classes.

Table 14.1 indicates a randomness and homogeneity of the area, and therefore, we have an instructive illustration of the established fact that to the untrained eye, randomness appears as regularity or tendency to cluster (the appearance of this regularity in such a random process is sometimes referred to as "Poisson clumping")

Run Distribution

To develop a distribution for the number of runs in a sequence, suppose we begin with two different kinds of objects (say, white (W) and black (B) balls) arranged randomly in a line.

We count the number of runs, R, defined by consecutive sequences of all Ws or all Bs (including sequences of size 1).

If there are n_1 W balls and n_2 B balls, the distribution for R under randomness can be constructed.

We note the expectation and variance of R, and the normal approximation:

$$E(R) = \frac{2n_1 n_2}{n_1 + n_2} + 1 \; ;$$

$$V(R) = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} \; ;$$

and

$$\frac{R - E(R)}{\sqrt{V(R)}}$$

is approximately (standard) normal with mean zero and variance one.

Based on this latter distributional approximation, an assessment can be made as to the randomness of the process that produced the sequence, and whether there are too many or too few runs for the continued credibility that the process is random.

Run statistics have proved especially important in monitoring quality control in manufacturing, but these same ideas could be useful in a variety of statistical sleuthing tasks.

Benford's Law

Besides the use of formal probability distributions, there are other related ideas that might be of value in the detection of fraud or other anomalies.

One such notion, called Benford's law, has captured some popular attention; for example, see the article by Malcolm W. Browne, "Following Benford's Law, or Looking Out for No. 1" (*New York Times*, August 4, 1998).

Benford's law gives a "probability distribution" for the first digits (1 to 9) found for many (naturally) occurring sets of numbers.

If the digits in some collection (such as tax returns, campaign finances, (Iranian) election results, or company audits) do not follow this distribution, there is a *prima facie* indication of fraud.

Benford's law gives a discrete probability distribution over the digits 1 to 9 according to:

$$P(X = r) = \log_{10}(1 + \frac{1}{r}) \ ,$$

for $1 \leq r \leq 9$. Numerically, we have the following:

r	Probability	r	Probability
1	.301	6	.067
2	.176	7	.058
3	.125	8	.051
4	.097	9	.046
5	.079		

Although there may be many examples of using Benford's law for detecting various monetary irregularities,

one of the most recent applications is to election fraud, such as in the 2009 Iranian Presidential decision.

A recent popular account of this type of sleuthing is Carl Bialik's article, "Rise and Flaw of Internet Election-Fraud Hunters" (*Wall Street Journal*, July 1, 2009).

It is always prudent to remember, however, that heuristics, such as Benford's law and other digit regularities, might point to a potentially anomalous situation that should be studied further, but violations of these presumed regularities should never be considered definitive "proof."

The Birthday Problem

Another helpful explanatory probability result is commonly referred to as the "birthday problem":

what is the probability that in a room of n people, at least one pair of individuals will have the same birthday.

As an approximation, we have $1 - e^{-n^2/(2\times 365)}$;

for example, when $k = 23$, the probability is .507; when $k = 30$, it is .706.

These surprisingly large probability values result from the need to consider matchings over all pairs of individuals in the room; that is, there are $\binom{n}{2}$ chances to consider for a matching, and these inflate the probability beyond what we might intuitively expect.

We give an example from Leonard Mlodinow's book, *The Drunkard's Walk* (2009):

Another lottery mystery that raised many eyebrows occurred in Germany on June 21, 1995. The freak event happened in a lottery named Lotto 6/49, which means that the winning six numbers are drawn from the numbers 1 to 49. On the day in question the winning numbers were 15-25-27-30-42-48. The very same sequence had been drawn previously, on December 20, 1986. It was the first time in 3,016 drawings that a winning sequence had been repeated. What were the chances of that? Not as bad as you'd think. When you do the math, the chance of a repeat at some point over the years comes out to around 28 per cent.

Survival Analysis

We begin with the epigram at the start of this section in your required reading:

The reason that the term "censored" is used is that in the pessimistic vocabulary of survival-analysis, life is a temporary phenomenon and someone who is alive is simply not dead yet. What the statistician would like to know is how long he or she lived but this information is not (yet) available and so is censored.
— Stephen Senn (*Dicing with Death*, 2003)

The area of statistics that models the time to the occurrence of an event, such as death or failure, is called *survival analysis*.

Some of the questions survival analysis is concerned with include:

what is the proportion of a population that will survive beyond a particular time;

among the survivors, at what (hazard) rate will they die (or fail);

how do the circumstances and characteristics of the population change the odds of survival;

can multiple causes of death (or failure) be taken into account.

The primary object of interest is the survival function, specifying the probability that time of death (the term to be used generically from now on), is later than some specified time.

Formally, we define the survival function as: $S(t) = P(T > t)$, where t is some time, and T is a random variable denoting the time of death.

The function must be nonincreasing, so: $S(u) \leq S(v)$, when $v \leq u$.

This reflects the idea that survival to some later time requires survival at all earlier times as well.

The most common way to estimate $S(t)$ is through the now ubiquitous Kaplan–Meier (1958) estimator, which allows a certain (important) type of right-censoring of the data.

This censoring is where the corresponding objects have either been lost to observation or their lifetimes are still ongoing when the data were analyzed.

Statistical Sleuthing and the Imposition of the Death Penalty: McCleskey v. Kemp (1987)

Those whom we would banish from society or from the human community itself often speak in too faint a voice to be heard above society's demand for punishment. It is the particular role of courts to hear these voices, for the Constitution declares that the majoritarian chorus may not alone dictate the conditions of social life.
— Supreme Court Justice Brennan (dissenting in *McCleskey v. Kemp*)

The United States has had a troubled history with the imposition of the death penalty.

Two amendments to the Constitution, the Eighth and the Fourteenth, operate as controlling guidelines for how death penalties are to be decided on and administered (if at all).

The Eighth Amendment prevents "cruel and unusual punishment"; the Fourteenth Amendment contains the famous "equal protection" clause:

No State shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any State deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws.

Various Supreme Court rulings over the years have relied on the Eighth Amendment to forbid some punishments entirely and to exclude others that are excessive in relation to the crime or the competence of the defendant.

One of the more famous such rulings was in *Furman v. Georgia* (1972), which held that an arbitrary and inconsistent imposition of the death penalty violates both the Eighth and Fourteenth Amendments, and constitutes cruel and unusual punishment.

This ruling lead to a moratorium on capital punishment throughout the United States that extended to 1976 when another Georgia case was decided in *Gregg v. Georgia* (1976).

The Supreme Court case of *Gregg v. Georgia* reaffirmed the use of the death penalty in the United States.

It held that the imposition of the death penalty does not automatically violate the Eighth and Fourteenth Amendments.

If the jury is furnished with standards to direct and limit the sentencing discretion, and the jury's decision is subjected to meaningful appellate review, the death sentence may be constitutional.

If, however, the death penalty is mandatory, so there is no provision for mercy based on the characteristics of the offender, then it is unconstitutional.

This short background on *Furman v. Georgia* and *Gregg v. Georgia* brings us to the case of *McCleskey v. Kemp* (1987), of primary interest in this section.

For us, the main importance of *McCleskey v. Kemp* is the use and subsequent complete disregard of a monumental statistical study by David C. Baldus, Charles Pulaski, and George G. Woodworth, "Comparative Review of Death Sentences: An Empirical Study of the Georgia Experience" (*Journal of Criminal Law and Criminology*, 1983, *74*, 661–753).

For a book length and extended version of this article, and an explicit discussion of *McCleskey v. Kemp*, see *Equal Justice and the Death Penalty: A Legal and Empirical Analysis*. David C. Baldus, George Woodworth, and Charles A. Pulaski, Jr., Boston: Northeastern University Press, 1990.

In *McCleskey v. Kemp*, the Court held that despite statistical evidence of a profound racial disparity in application of the death penalty, such evidence is insufficient to invalidate a defendant's death sentence.

The syllabus of this ruling is given in your required reading (and the actual opinion and dissent in the Supplementary Readings).

To see additional contemporary commentary, an article by Anthony Lewis lamenting this ruling appeared in the *New York Times* (April 28, 1987), entitled "Bowing To Racism."

We make a number of comments about the majority opinion in *McCleskey v. Kemp* summarized in the syllabus and noted in the article by Anthony Lewis.

First, it is rarely the case that a policy could be identified as the cause for an occurrence in one specific individual.

The legal system in its dealings with epidemiology and toxicology has generally recognized that an agent can never be said to have been the specific cause of, say, a disease in a particular individual.

This is the notion of specific causation, which is typically unprovable.

As an alternative approach to causation, courts have commonly adopted a criterion of general causation defined by relative risk being greater than 2.0 (as discussed earlier) to infer that a toxic agent was more likely than not the cause of a specific person's disease (and thus open to compensation).

In his dissent, Justice Brennan makes this exact point when he states:
"For this reason, we have demanded a uniquely high degree of rationality in imposing the death penalty. A capital sentencing system in which race more likely than not plays a role does not meet this standard."

To require that a defendant prove that the decision makers in his particular case acted with discriminatory malice is to set an unreachable standard.

So is an expectation that statistics could ever absolutely prove "that race enters into any capital sentencing decisions or that race was a factor in petitioner's case."

Statistical sleuthing can at best identify anomalies that need further study, for example, when Benford's law is used to identify possible fraud, or the Poisson model is used to suggest a lack of spatial clustering.

But, irrespective, the anomalies cannot be just willed away as if they never existed.

The *New York Review of Books* in its December 23, 2010 issue scored a coup by having a lead article entitled "On the Death Sentence," by retired Supreme Court Justice John Paul Stevens.

Stevens was reviewing the book, *Peculiar Institution: America's Death Penalty in an Age of Abolition* (by David Garland).

In the course of his essay, Stevens comments on *McCleskey v. Kemp* and notes that Justice Powell (who wrote the majority opinion) in remarks he made to his biographer, said that he should have voted the other way in the *McCleskey* 5 to 4 decision.

It's too bad we cannot retroactively reverse Supreme Court rulings, particularly given the doctrine of *stare decisis*, according to which judges are obliged to respect the precedents set by prior decisions.

The doctrine of *stare decisis* suggests that no amount of statistical evidence will ever be sufficient to declare the death penalty in violation of the "equal protection" clause of the Fourteenth Amendment.

The relevant quotation from the Stevens review follows:

In 1987, the Court held in McCleskey v. Kemp that it did not violate the Constitution for a state to administer a criminal justice system under which murderers of victims of one race received death sentences much more frequently than murderers of victims of another race. The case involved a study by Iowa law professor David Baldus and his colleagues demonstrating that in Georgia murderers of white victims were eleven times more likely to be sentenced to death than were murderers of black victims. Controlling for race-neutral factors and focusing solely on decisions by prosecutors about whether to seek the death penalty, Justice Blackmun observed in dissent, the effect of race remained "readily identifiable" and "statistically significant" across a sample of 2,484 cases.

That the murder of black victims is treated as less culpable than the murder of white victims provides a haunting reminder of once-prevalent Southern lynchings. Justice Stewart, had he remained on the Court, surely would have voted with the four dissenters. That conclusion is reinforced by Justice Powell's second thoughts; he later told his biographer that he regretted his vote in McCleskey.