Module 2: The (Un)reliability of Clinical and Actuarial Predictions of (Dangerous) Behavior

I would not say that the future is necessarily less predictable than the past. I think the past was not predictable when it started.
– Donald Rumsfeld

Abstract: The prediction of dangerous and/or violent behavior is important to the conduct of the United States justice system in making decisions about restrictions of personal freedom such as preventive detention, forensic commitment, or parole. This module discusses behavioral prediction both through clinical judgement as well as actuarial assessment. The general conclusion drawn is that for both clinical and actuarial prediction of dangerous behavior, we are far from a level of accuracy that could justify routine use. To support this later negative assessment, two topic areas are discussed at some length: 1) the MacArthur Study of Mental Disorder and Violence, including the actuarial instrument developed as part of this project (the Classification of Violence Risk (COVR)), along with all the data collected that helped develop the instrument; 2) the Supreme Court case of *Barefoot v. Estelle* (1983) and the American Psychiatric Association "friend of the court" brief on the (in)accuracy of clinical prediction for the commission of future violence. An elegant Justice Blackmun dissent is given in its entirety that contradicts the majority decision that held: There is no merit to petitioner's argument that psychiatrists, individually and as a group, are incompetent to predict with an acceptable degree of reliability that a particular criminal will

commit other crimes in the future, and so represent a danger to the community.

Contents

1 Introduction

An ability to predict and treat dangerous or violent behavior in criminal offenders is important to the administration of the criminal justice system in the United States. This prediction might be in the context of preventive detentions, parole decisions, forensic commitments, or other legal forms of restriction on personal liberty. Behavioral prediction might rely on clinical judgement (usually through trained psychologists or other medically versed individuals) or by actuarial (statistical) assessments. In any case, concern should be on the reliability of such predictions, and more pointedly, on the state of clinical and actuarial prediction. So, the question: are we at such a level of predictive accuracy that as a society we can justify the necessary false positives that would inappropriately restrict the personal liberty of those who would prove to be neither dangerous or violent. Unfortunately, the conclusion reached in this module is that for both clinical or actuarial prediction of dangerous behavior, we are quite far from a level that could sanction routine use.

Evidence on prediction accuracy can typically be presented in the form of a 2×2 contingency table defined by a cross-classification of individuals according to the events A and $\bar{A}$ (whether the person proved dangerous (A) or not ($\bar{A}$)); and B and $\bar{B}$ (whether the person was predicted to be dangerous (B) or not ($\bar{B}$)):

Prediction:

B (dangerous)
$\bar{B}$ (not dangerous)

Outcome (Post-Prediction):

A (dangerous)
$\bar{A}$ (not dangerous)

A generic 2×2 table presenting the available evidence on prediction accuracy might then be given in the following form (arbitrary cell frequencies are indicated using the appropriate subscript combinations of A and $\bar{A}$ and B and $\bar{B}$):

			Outcome	
		A (dangerous)	$\bar{A}$ (not dangerous)	row sums
Prediction	B (dangerous)	n_{BA}	$n_{B\bar{A}}$	n_B
	$\bar{B}$ (not dangerous)	$n_{\bar{B}A}$	$n_{\bar{B}\bar{A}}$	$n_{\bar{B}}$
	column sums	n_A	$n_{\bar{A}}$	n

2 Clinical Prediction

The 2×2 contingency table given immediately below illustrates the poor prediction of dangerous behavior when based on clinical assessment. These data are from Kozol, Boucher, and Garofalo (1972), "The Diagnosis and Treatment of Dangerousness":

			Outcome	
		A (dangerous)	$\bar{A}$ (not dangerous)	row sums
Prediction	B (dangerous)	17	32	49
	$\bar{B}$ (not dangerous)	31	355	386
	column sums	48	387	435

For these data, 2 out of 3 predictions of "dangerous" are wrong ($.65 = 32/49$ to be precise). This is the source of the "error rate" reported in the Supreme Court opinion in Barefoot v. Estelle (1983), discussed at great length later. Also, 1 out of 12 predictions of "not dangerous" are wrong ($.08 = 31/386$).

In his dissent opinion in the Barefoot v. Estelle case, Justice Blackmun quotes the American Psychiatric Association *amicus curiae* brief as follows: " [the] most that can be said about any individual is that a history of past violence increases the probability that future violence will occur." In other words, the best we can say is that "past violence" (B) is facilitative of "future violence" (A) but the error in that prediction can be very large as it is here for the Kozol et al. data: $P(A|B) = \frac{17}{49} = .35$ is greater than $P(A) = \frac{48}{435} = .11$. But this implies that 2 out of 3 such predictions of "dangerous" are wrong (or, 1 out of 3 are correct). To us, the accuracy of these behavioral predictions is insufficient to justify any incarceration policy based on them; the same conclusion will hold for the type of actuarial prediction of "dangerous" discussed in the section to follow.

In Module 4 on diagnostic testing, the Meehl and Rosen (1955) notion of "clinical efficiency" is formally discussed, or when a diagnostic test is more accurate than just predicting using base rates. For these data, prediction by base rates would be to say everyone will be "not dangerous" because the number of people who are "not dangerous" (387) is larger than the number of people who are "dangerous" (48). Here, we would be correct in our predictions 89% of the time (.89 = 387/435). Based on clinical prediction, we would be correct a *smaller* 86% percentage of the time (.86 = (17 + 355)/435). So, according to the Meehl and Rosen characterization, clinical prediction is *not* "clinically efficient" because one can do better by just predicting according to base rates.

In commenting on the Kozol, et al. study, Monahan (1973) takes issue with the article's principal conclusion that "dangerousness can be reliably diagnosed and effectively treated" and notes that it "is, at best, misleading and is largely refuted by their own data." Mohahan concludes his critique with the following quotation from Wenk, Robison, and Smith (1972):

> Confidence in the ability to predict violence serves to legitimate intrusive types of social control. Our demonstration of the *futility* of such prediction should have consequences as great for the protection of individual liberty as a demonstration of the utility of violence prediction would have for the

protection of society. (p. 402)

3 Actuarial Prediction

Paul Meehl in his iconic 1954 monograph, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, created quite a stir with his convincing demonstration that mechanical methods of data combination, such as multiple regression, outperform (expert) clinical prediction. The enormous amount of literature produced since the appearance of this seminal contribution has uniformly supported this general observation; similarly, so have the extensions suggested for combining data in ways other than by multiple regression, for example, by much simpler unit weighting schemes, or those using other prior weights. It appears that individuals who are conversant in a field are better at selecting and coding information than they are at actually integrating it. Combining such selected information in a more mechanical manner will generally do better than the person choosing such information in the first place.[1]

[1] A 2005 article by Robyn Dawes in the *Journal of Clinical Psychology* (*61*, 1245–1255) has the intriguing title "The Ethical Implications of Paul Meehl's Work on Comparing Clinical Versus Actuarial Prediction Methods." Dawes' main point is that given the overwhelming evidence we now have, it is unethical to use clinical judgment in preference to the use of statistical prediction rules. We quote from the abstract:

> Whenever statistical prediction rules . . . are available for making a relevant prediction, they should be used in preference to intuition. . . . Providing service that assumes that clinicians "can do better" simply based on self-confidence or plausibility in the absence of evidence that they can actually do so is simply unethical. (p. 1245)

The MacArthur Study of Mental Disorder and Violence

The MacArthur Research Network on Mental Health and the Law was created in 1988 by a major grant to the University of Virginia from the John D. and Catherine T. MacArthur Foundation. The avowed aim of the Network was to construct an empirical foundation for the next generation of mental health laws, assuring the rights and safety of individuals and society. New knowledge was to be developed about the relation between the law and mental health; new assessment tools were to be developed along with criteria for evaluating individuals and making decisions affecting their lives. The major product of the Network was the MacArthur Violence Risk Assessment Study; its principal findings were published in the very well-received 2001 book, *Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence* (John Monahan, et al., Oxford University Press). More importantly for us (and as a source of illustrations used throughout), the complete data set (on 939 individuals over 134 risk factors) is available on the web.

The major analyses reported in *Rethinking Risk Assessment* are based on constructed classification trees; in effect, these are branching decision maps for using risk factors to assess the likelihood that a particular person will commit violence in the future. All analyses were carried out with an SPSS classification-tree program, called CHAID, now a rather antiquated algorithm (the use of this method without a modern means of cross-validation most likely led to the overfitting difficulties to be discussed shortly). Moreover, these same classification tree analyses have been incorporated into a proprietary software product called the Classification of Violence Risk (COVR);

it is available from the Florida-based company PAR (Psychological Assessment Resources). The program is to be used in law enforcement/mental health contexts to assess “dangerousness to others,” a principal standard for inpatient or outpatient commitment or commitment to a forensic hospital.

One of the authors of the current module taught a class entitled Advanced Multivariate Methods for the first time in the Fall of 2011, with a focus on recent classification and regression tree methods developed over the last several decades and implemented in the newer environments of Matlab and R (but not in SPSS). These advances involve “random forests,” “bootstrap aggregation (bagging),” “boosting algorithms,” “ensemble methods,” and a number of techniques that avoid the dangers of overfitting and allow several different strategies of internal cross-validation. To provide interesting projects for the class to present, a documented data set was obtained from the statistician on the original MacArthur study; this was a more transparent packaging of the data already available on the web. This ”cleaned-up” data set could provide a direct replication of the earlier SPSS analyses (with CHAID); but in addition and more importantly, all the “cutting-edge” methods could now be applied that were unavailable when the original MacArthur study was completed in the late 1990s. At the end of the semester, five subgroups of the graduate students in the class reported on analyses they did on the MacArthur data set (each also had a different psychological test battery to focus on, for example, Brief Psychiatric Rating Scale, Novaco Anger Scale, Novaco Provocation Inventory, Big Five Personality Inventory, Psychopathy Checklist (Screening Version)). Every one of the talks essentially reported a “wash-out” when cross-validation and prediction was the emphasis as opposed to just fitting the classification structures. We could not do better than just predicting with base rates. This was a first indication that the prediction of "dangerousness" was possibly not as advanced as the MacArthur Network might have us believe.

The second major indication of a difficulty with prediction even with the newer MacArthur assessment tools was given by a close read of the first small cross-validation study done to justify this actuarial software COVR (mentioned earlier): "An Actuarial Model of Violence Risk Assessment for Persons with Mental Disorders" (John Monahan, et al., *Psychiatric Services*, 2005, 56, 810–815). The abstract of this article is given below:

> Objectives: An actuarial model was developed in the MacArthur Violence Risk Assessment Study to predict violence in the community among patients who have recently been discharged from psychiatric facilities. This model, called the multiple iterative classification tree (ICT) model, showed considerable accuracy in predicting violence in the construction sample. The purpose of the study reported here was to determine the validity of the multiple ICT model in distinguishing between patients with high and low risk of violence in the community when applied to a new sample of individuals.
>
> Methods: Software incorporating the multiple ICT model was administered with independent samples of acutely hospitalized civil patients. Patients who were classified as having a high or a low risk of violence were followed in the community for 20 weeks after discharge. Violence included any battery with physical injury, use of a weapon, threats made with a weapon in hand, and sexual assault.
>
> Results: Expected rates of violence in the low- and high-risk groups were 1 percent and 64 percent, respectively. Observed rates of violence in the low- and high-risk groups were 9 percent and 35 percent, respectively, ... These findings may reflect the "shrinkage" expected in moving from construction to validation samples.

> Conclusions: The multiple ICT model may be helpful to clinicians who are faced with making decisions about discharge planning for acutely hospitalized civil patients.

John Monahan in his influential NIMH monograph, *The Clinical Prediction of Violent Behavior* (1977), observed that, even allowing for possible distortions in the research data, "it would be fair to conclude that the best clinical research currently in existence indicates that psychiatrists and psychologists are accurate in no more than one out of three predictions of violent behavior over a several year period among institutionalized populations that had both committed violence in the past (and thus had high base rates for it) and who were diagnosed as mentally ill." In other words, predictions that someone will be violent (and therefore subject to detention) will be wrong two out of three times. With such a dismal record of clinical prediction, there were high expectations that the MacArthur Network could produce a much better (actuarial) instrument in COVR. Unfortunately, that does not appear to be the case. The figures of 64% and 35% given in the abstract suggest two conclusions: in the training sample, the error in predicting dangerousness is 1 out of 3; whether this shows "considerable accuracy in predicting violence in the construction sample" is highly questionable, even assuming the inflated value is correct. The cross-classified proportion of 35% gives the error of being wrong in prediction of dangerousness as 2 out of 3. It is quite an understatement to then say: "These findings may reflect the "shrinkage" expected in moving from construction to validation samples." What it reflects is that actuarial prediction of violence is exactly as bad as clinical prediction. This may be one of the only (if not the only) examples from the behavioral science literature in which actuarial prediction doesn't do better than clinical prediction.

The complete 2×2 table from the COVR validation study follows:

		Outcome		
		A (dangerous)	$\bar{A}$ (not dangerous)	row sums
Prediction	B (dangerous)	19	36	55
	$\bar{B}$ (not dangerous)	9	93	102
	column sums	28	129	157

As noted above, a high prediction of "dangerous" is wrong 65% (= 36/55) of the time. A prediction of "not dangerous" is incorrect 9% (= 9/102) of the time (again, this is close to the 1 out of 12 incorrect predictions of "not dangerous" typically seen for purely clinical predictions). The accuracy or "hit-rate" is (10 + 93)/157 = .71. If everyone were predicted to be nondangerous, we would be correct 129 out of 157 times, the base rate for $\bar{A}$: $P(\bar{A}) = 129/157 =$ .82. Obviously, the accuracy of prediction using base rates (82%) is better than for the COVR (71%), making the COVR not "clinically efficient" according to the Meehl and Rosen terminology.

We give two more examples from the MacArthur data set mentioned earlier that involve the variables of "prior arrest" or "prior violence" as diagnostic "tests" in their own right. The first adopts prior arrest as a diagnostic "test": dangerous—one or more prior arrests (B); not dangerous—no prior arrests ($\bar{B}$).

		Outcome		
		A (dangerous)	$\bar{A}$ (not dangerous)	row sums
Prediction	B (dangerous)	103	294	397
	$\bar{B}$ (not dangerous)	39	354	393
	column sums	142	648	790

Here, 3 out of 4 predictions of "dangerous" are wrong (.74 = 294/397); 1 out of 10 predictions of "not dangerous" are wrong (.10 = 39/393).

The accuracy of the test is (103 + 354)/790 = .50, and the correctness of prediction by base rates is 648/790 = .82; thus, “prior arrest” is not a clinically efficient “test.”

The second example uses prior violence as a diagnostic “test”: dangerous—prior violence (B); not dangerous—no prior violence ($\bar{B}$).

		Outcome		
		A (dangerous)	$\bar{A}$ (not dangerous)	row sums
Prediction	B (dangerous)	48	106	154
	$\bar{B}$ (not dangerous)	128	657	785
	column sums	176	763	939

In this case, 7 out of 10 predictions of “dangerous” are wrong (.69 = 106/154); 1 out of 6 predictions of “not dangerous” are wrong (.16 = 128/785). The accuracy of the test, (48 + 657)/939 = .75, is less than the the correctness of prediction by base rates: 763/939 = .81; thus, “prior violence” is not a clinically efficient “test.”

4 Barefoot v. Estelle (1983)

The present discussion on probabilistic reasoning concerns the unreliability of clinical and actuarial behavioral prediction, particularly for violence; the particular section to follow includes two extensive redactions in appendices: one is the majority opinion in the Supreme Court case of *Barefoot v. Estelle* (1983) and an eloquent Justice Blackmun dissent; the second is an *amicus curiae* brief in this same case from the American Psychiatric Association on the accuracy of clinical prediction of future violence. Both of these documents are detailed, self-explanatory, and highly informative about our current lack of ability to make clinical assessments that lead to accurate and reliable predictions of future behavior. To set the background for the *Barefoot v. Estelle* case, the beginning part of the *amicus curiae* brief follows; a redaction of the remainder of the brief, as already noted, is given in an appendix at the end of the chapter.

Brief for American Psychiatric Association as *Amicus Curiae* Supporting Petitioner, Barefoot v. Estelle

Petitioner Thomas A. Barefoot stands convicted by a Texas state court of the August 7, 1978 murder of a police officer—one of five categories of homicides for which Texas law authorizes the imposition of the death penalty. Under capital sentencing procedures established after this Court's decision in Furman v. Georgia, the "guilt" phase of petitioner's trial was followed by a separate sentencing proceeding in which the jury was directed to answer three statutorily prescribed questions. One of these questions—and the only question of relevance here—directed the jury to determine: whether there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society. The jury's affirmative response to this question resulted in petitioner being sentenced to death.

The principle evidence presented to the jury on the question of petitioner's "future dangerousness" was the expert testimony of two psychiatrists, Dr. John T. Holbrook and Dr. James Grigson, both of whom testified for the prosecution. Petitioner elected not to testify in his own defense. Nor did he present any evidence or testimony, psychiatric or otherwise, in an attempt to rebut the state's claim that he would commit future criminal acts of violence.

Over defense counsel's objection, the prosecution psychiatrists were permitted to offer clinical opinions regarding petitioner, including their opinions on the ultimate issue of future dangerousness, even though they had not performed a psychiatric examination or evaluation of him. Instead, the critical psychiatric testimony was elicited through an extended hypothetical question propounded by the prosecutor. On the basis of the assumed facts stated in the hypothetical, both Dr. Holbrook and Dr. Grigson gave essentially the same testimony.

First, petitioner was diagnosed as a severe criminal sociopath, a label variously defined as describing persons who “lack a conscience,” and who “do things which serve their own purposes without regard for any consequences or outcomes to other people.” Second, both psychiatrists testified that petitioner would commit criminal acts of violence in the future. Dr. Holbrook stated that he could predict petitioner’s future behavior in this regard “within reasonable psychiatric certainty.” Dr. Grigson was more confident, claiming predictive accuracy of “one hundred percent and absolute.”

The prosecutor’s hypothetical question consisted mainly of a cataloguing of petitioner’s past antisocial behavior, including a description of his criminal record. In addition, the hypothetical question contained a highly detailed summary of the prosecution’s evidence introduced during the guilt phase of the trial, as well as a brief statement concerning petitioner’s behavior and demeanor during the period from his commission of the murder to his later apprehension by police.

In relevant part, the prosecutor’s hypothetical asked the psychiatrists to assume as true the following facts: First, that petitioner had been convicted of five criminal offenses—all of them nonviolent, as far as the record reveals—and that he had also been arrested and charged on several counts of sexual offenses involving children. Second, that petitioner had led a peripatetic existence and “had a bad reputation for peaceful and law abiding citizenship” in each of eight communities that he had resided in during the previous ten years. Third, that in the two-month period preceding the murder, petitioner was unemployed, spending much of his time using drugs, boasting of his plans to commit numerous crimes, and in various ways deceiving certain acquaintances with whom he was living temporarily. Fourth, that petitioner had murdered the police officer as charged, and that he had done so with “no provocation whatsoever” by shooting the officer in the head “from a distance of no more than six inches.” And fifth, that subsequent to the murder, petitioner was observed by one witness, “a homosexual,” who stated that petitioner “was not in any way acting unusual or that anything was bothering him or upsetting him . . . ”

Testimony of Dr. Holbrook:

Dr. Holbrook was the first to testify on the basis of the hypothetical question. He stated that the person described in the question exhibited "probably six or seven major criterias (sic) for the sociopath in the criminal area within reasonable medical certainty." Symptomatic of petitioner's sociopathic personality, according to Dr. Holbrook, was his consistent "antisocial behavior" from "early life into adulthood," his willingness to take any action which "serves [his] own purposes" without any regard for the "consequences to other people," and his demonstrated failure to establish any "loyalties to the normal institutions such as family, friends, politics, law or religion."

Dr. Holbrook explained that his diagnosis of sociopathy was also supported by petitioner's past clinical violence and "serious threats of violence," as well as an apparent history of "escaping or running away from authority" rather than "accepting a confrontation in the legal way in a court of law." And finally, Dr. Holbrook testified that petitioner had shown a propensity to "use other people through lying and manipulation ... " According to Dr. Holbrook, by use of such manipulation the sociopath succeeds in "enhancing [his] own ego image ... It makes [him] feel good."

After stating his diagnosis of sociopathy, Dr. Holbrook was asked whether he had an "opinion within reasonable psychiatric certainty as to whether or not there is a probability that the Thomas A. Barefoot in that hypothetical will commit criminal acts of violence in the future that would constitute a continuing threat to society?" Without attempting to explain the implied clinical link between his diagnosis of petitioner and his prediction of future dangerousness, Dr. Holbrook answered simply: "In my opinion he will."

Testimony of Dr. Grigson:

On the basis of the prosecutor's hypothetical question, Dr. Grigson diagnosed petitioner as "a fairly classical, typical, sociopathic personality disorder" of the "most severe category." The most "outstanding characteristic" of persons fitting this diagnosis, according to Dr. Grigson, is the complete "lack of a conscience." Dr. Grigson stated that such persons "repeatedly break the rules, they con, manipulate and use people, [and] are only interested in their own self pleasure [and] gratification."

Although Dr. Grigson testified that some sociopathic individuals do not pose a continuing threat to society, he characterized petitioner as "your most severe sociopath." Dr. Grigson stated that persons falling into this special category are "the ones that ... have complete disregard for another human being's life." Dr. Grigson further testified that "there is not anything in medicine or psychiatry or any other field that will in any way at all modify or change the severe sociopath."

The prosecutor then asked Dr. Grigson to state his opinion on the ultimate issue—"whether or not there is a probability that the defendant ... will commit criminal acts of violence that would constitute a continuing threat to society?" Again, without explaining the basis for his prediction or its relationship to the diagnosis of sociopathy, Dr. Grigson testified that he was "one hundred percent" sure that petitioner "most certainly would" commit future criminal acts of violence. Dr. Grigson also stated that his diagnosis and prediction would be the same whether petitioner "was in the penitentiary or whether he was free."

The psychiatrist featured so prominently in the opinions for *Barefoot v. Estelle* and the corresponding American Psychiatric Association *amicus* brief, James Grigson, played the same role repeatedly in the Texas legal system. For over three decades before his retirement in 2003, he testified when requested at death sentence hearings to a high certainty as to "whether there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society." An affirmative answer by the sentencing jury imposed the death penalty automatically, as it was on Thomas Barefoot; he was executed on October 30, 1984. When asked if he had a last statement to make, he replied:

> Yes, I do. I hope that one day we can look back on the evil that we're doing right now like the witches we burned at the stake. I want everybody to know that I hold nothing against them. I forgive them all. I hope everybody I've done anything to will forgive me. I've been praying all day for Carl Levin's wife to drive the bitterness from her heart because that bitterness that's in her heart will send her to Hell just as surely as any other sin. I'm sorry for everything I've ever done to anybody. I hope they'll forgive me.

James Grigson was expelled in 1995 from the American Psychiatric Association and the Texas Association of Psychiatric Physicians for two chronic ethics violations: making statements in testimony on defendants he had not actually examined, and for predicting violence with 100% certainty. The press gave him the nickname of "Dr. Death."

Barefoot v. Estelle has another connection to the distinction between actuarial and clinical prediction, and where the former is commonly better than the latter. There is evidence mentioned in the APA brief that actuarial predictions of violence carried out by statistically informed laymen might be better than those of a clinician. This may be due to a bias that psychiatrists might (unsuspectingly) have in overpredicting violence because of the clients they see or for other reasons related to their practice. There is a pertinent passage from the court opinion (not given in our redactions):

> That psychiatrists actually may be less accurate predictors of future violence than laymen, may be due to personal biases in favor of predicting violence arising from the fear of being responsible for the erroneous release of a violent individual. ... It also may be due to a tendency to generalize from experiences with past offenders on bases that have no empirical relationship to future violence, a tendency that may be present in Grigson's and Holbrook's testimony. Statistical prediction is clearly more reliable than clinical prediction ... and prediction based on statistics alone may be done by anyone.

The two psychiatrists mentioned in *Barefoot v. Estelle*, James Grigson and John Holbrook, appeared together repeatedly in various capital sentencing hearings in Texas during the later part of the 20th century. Although Grigson was generally the more outrageous of the two with predictions of absolute certitude based on a sociopath diagnosis, Holbrook was similarly at fault ethically. This pair of psychiatrists of Texas death penalty fame might well be nicknamed "Dr. Death" and "Dr. Doom." They were both culpable in the famous exoneration documented in the award winning film by Errol Morris, *The Thin Blue Line*. To tell this story, we give the summary of the Randall Dale Adams exoneration from the Northwestern Law School's Center on Wrongful Convictions (written by Robert Warden with Michael L. Radelet):

> Sentenced to death in 1977 for the murder of a police officer in Dallas, Texas, Randall Dale Adams was exonerated as a result of information uncovered by film-maker Errol Morris and presented in an acclaimed 1988 documentary, *The Thin Blue Line*.
>
> Patrolman Robert Wood was shot to death during a traffic stop on November 28, 1976, by sixteen-year-old David Ray Harris, who framed Adams to avoid prosecution himself. Another factor in the wrongful conviction was the surprise—and partly perjured—testimony of three eyewitnesses whose existence had been concealed from the defense until the witnesses appeared in the courtroom. A third factor was a statement Adams signed during interrogation that the prosecution construed as an admission that he had been at the scene of the crime.
>
> The day before the murder, Adams was walking along a Dallas street after his car had run out of gasoline. Harris happened by, driving a stolen car. He offered Adams a ride and the two wound up spending the afternoon and evening together, drinking beer, smoking marijuana, pawning various items Harris had stolen, and going to a drive-in movie theater to watch porn movies. Adams then returned to a motel where he was staying.
>
> Shortly after midnight, Wood and his partner, Teresa Turko, spotted Harris driving a blue car with no headlights. The officers stopped the car and, as Wood approached the driver's side, Harris shot him five times. Wood died on the spot. As the car sped off, Turko fired several shots, but missed. She did not get a license number. She seemed certain that there was only one person in the car—the driver.
>
> Harris drove directly to his home in Vidor, 300 miles southeast of Dallas. Over the next several days, he bragged to friends that he had "offed a pig"

in Dallas. When police in Vidor learned of the statements, they took Harris in for questioning. He denied having had anything to do with the murder, claiming he had said otherwise only to impress his friends. But when police told him that a ballistics test established that a pistol he had stolen from his father was the murder weapon, Harris changed his story. He now claimed that he had been present at the shooting, but that it had been committed by a hitchhiker he had picked up—Adams.

Adams, an Ohio native working in Dallas, was taken in for questioning. He denied any knowledge of the crime, but he did give a detailed statement describing his activities the day before the murder. Police told him he had failed a polygraph test and that Harris had passed one, but Adams remained resolute in asserting his innocence.

Although polygraph results are not admissible in Texas courts, the results provided some rationale for questioning Harris's story. However, when a police officer is murdered, authorities usually demand the most severe possible punishment, which in Texas, and most other United States jurisdictions, is death. Harris was only sixteen and ineligible for the death penalty; Adams was twenty-seven and thus could be executed.

At trial before Dallas County District Court Judge Don Metcalfe and a jury, Turko testified that she had not seen the killer clearly, but that his hair was the color of Adams's. She also said that the killer wore a coat with a fur collar. Harris had such a coat, but Adams did not.

Adams took the stand and emphatically denied having any knowledge of the crime. But then the prosecution sprang two surprises. The first was the introduction of Adams's purported signed statement, which police and prosecutors claimed was a confession, although it said only—falsely, according to Adams—that when he was in the car with Harris, they had at one point been near the crime scene. The second was the testimony of three purported eyewitnesses whose existence had until then been unknown to the defense. One of these witnesses, Michael Randell, testified that he had driven by the scene shortly before the murder and, in the car that had been stopped by the officers, had seen two persons, one of whom he claimed was Adams. The other two witnesses, Robert and Emily Miller, had happened by at about the same time, but claimed to have seen only one person in the car—Adams.

Because the eyewitnesses were called only to rebut Adams's testimony, prosecutors claimed that Texas law did not require them to inform the defense of their existence before they testified. The weekend after their surprise testimony, however, the defense learned that Emily Miller had initially told police that the man she had seen appeared to be Mexican or a light-skinned African American. When the defense asked to recall the Millers to testify, the prosecution claimed that the couple had left town. In fact, the Millers had only moved from one part of Dallas to another. When the defense asked to introduce Emily Miller's statement, Judge Metcalfe would not allow it. He said it would be unfair to impeach her credibility when she was not available for further examination.

The jury quickly returned a verdict of guilty and turned to sentencing. Under Texas law, in order for Adams to be sentenced to death, the jury was required to determine, among other things, whether there was "beyond a reasonable doubt [a] probability" that he or she would commit future acts of violence. To establish that Adams met that oxymoronic criterion, the prosecution called Dr. James Grigson, a Dallas psychiatrist known as "Dr. Death," and Dr. John Holbrook, former chief of psychiatry for the Texas Department of Corrections.

Although the American Psychiatric Association has said on several occasions that future dangerousness was impossible to predict, Grigson and Holbrook testified that Adams would be dangerous unless executed. Grigson testified similarly in more than 100 other Texas cases that ended in death sentences. After hearing the psychiatrists, Adams's jury voted to sentence him to death. Twenty one months later, at the end of January 1979, the Texas Court of Criminal Appeals affirmed the conviction and death sentence. Judge Metcalfe scheduled the execution for May 8, 1979.

Adams was three days away from execution when United States Supreme Court Justice Lewis F. Powell Jr. ordered a stay. Powell was troubled that prospective jurors with moral qualms about the death penalty had been excluded from service, even though they had clearly stated that they would follow the Texas law.

To most observers—including, initially, Dallas District Attorney Henry Wade (of Roe v. Wade fame) the Supreme Court's language meant that

Adams was entitled to a new trial. But a few days later Wade announced that a new trial would be a waste of money. Thus, he said, he was asking Governor Bill Clements to commute Adams's sentence to life in prison. When the governor promptly complied, Wade proclaimed that there now would be no need for a new trial. Adams, of course, thought otherwise, but the Texas Court of Criminal Appeals agreed with Wade. As a result of the governor's action, said the court, "There is now no error in the case."

In March 1985, Errol Morris arrived in Dallas to work on a documentary about Grigson—"Dr. Death." Morris's intent had not been to question the guilt of defendants in whose cases Grigson had testified but only to question his psychiatric conclusions. When Morris met Adams, the focus of the project changed.

Morris learned from Randy Schaffer, a volunteer Houston lawyer who had been working on the case since 1982, that Harris had not led an exemplary life after helping convict Adams. Harris had joined the Army and been stationed in Germany, where he had been convicted in a military court of a series [of] burglaries and sent to prison in Leavenworth, Kansas. A few months after his release, Harris had been convicted in California of kidnapping, armed robbery, and related crimes.

After his release from prison in California, and five months after Morris arrived in Dallas, Harris tried to kidnap a young woman named Roxanne Lockard in Beaumont, Texas. In an effort to prevent the abduction, Lockard's boyfriend, Mark Mays, exchanged gunfire with Harris. Mays was shot to death and Harris was wounded. For the Mays murder—a crime that would not have occurred if Dallas authorities convicted the actual killer of Officer Wood eight years earlier—Harris was sentenced to death.

Meanwhile, Morris and Schaffer discovered that Officer Turko had been hypnotized during the investigation and initially had acknowledged that she had not seen the killer—facts that the prosecution had illegally withheld from the defense. Morris and Schaffer also found that robbery charges against the daughter of eyewitness Emily Miller had been dropped after Miller agreed to identify Adams as Wood's killer. The new information, coupled with the fact that Miller initially had described the killer as Mexican or African American, became the basis for a new trial motion.

In 1988, during a three-day hearing on the motion before Dallas District Court Judge Larry Baraka, Harris recanted. “Twelve years ago, I was a kid, you know, and I’m not a kid anymore, and I realize I’ve been responsible for a great injustice,” Harris told Baraka. “And I felt like it’s my responsibility to step forward, to be a man, to admit my part in it. And that’s why I’m trying to correct an injustice.”

On December 2, 1988, Judge Baraka recommended to the Texas Court of Criminal Appeals that Adams be granted a new trial, and two months later he wrote a letter to the Texas Board of Pardons and Paroles recommending that Adams be paroled immediately. The board refused, but on March 1 the Texas Court of Criminal Appeals unanimously concurred with Baraka that Adams was entitled to a new trial. Three weeks later, Adams was released on his own recognizance, and two days after that, Dallas District Attorney John Vance, who had succeeded Wade, dropped all charges.

Harris was never tried for the murder of Officer Woods. On June 30, 2004, he was executed for the Mays murder.

The *Federal Rules of Evidence* and the admissibility of expert witnesses and scientific data was influenced heavily by the case of *Daubert v. Merrell Dow Pharmaceuticals* (1993) that promulgates the *Daubert* standard for admitting expert testimony in federal courts. The majority opinion in *Daubert* was written by Justice Blackman, the same justice who wrote the dissent in *Barefoot v. Estelle.* The court stated that Rule 702 of the *Federal Rules of Evidence* was the governing standard for admitting scientific evidence in trials held in federal court (and now in most state courts as well). Rule 702, Testimony by Experts, states:

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.

We give a redaction of that part of the Wikipedia article on *Daubert v. Merrell Dow Pharmaceuticals* devoted to the discussion of the *Daubert* standard governing expert testimony. We doubt that clinical predictions of violence based on a sociopath diagnosis would be admissible under the *Daubert* standard.

The Standard Governing Expert Testimony: Three key provisions of the Rules governed admission of expert testimony in court. The first was *scientific knowledge.* This means that the testimony must be scientific in nature, and that the testimony must be grounded in "knowledge." Of course, science does not claim to know anything with absolute certainty; science "represents a *process* for proposing and refining theoretical explanations about the world that are subject to further testing and refinement." The "scientific knowledge" contemplated by Rule 702 had to be arrived at by the scientific method.

Second, the scientific knowledge must *assist the trier of fact* in understanding the evidence or determining a fact in issue in the case. The trier of fact is often either a jury or a judge; but other fact-finders may exist within the contemplation of the federal rules of evidence. To be helpful to the trier of fact, there must be a "valid scientific connection to the pertinent inquiry as a prerequisite to admissibility." Although it is within the purview of scientific knowledge, knowing whether the moon was full on a given night does not typically assist the trier of fact in knowing whether a person was sane when he or she committed a given act.

Third, the Rules expressly provided that the judge would make the threshold determination regarding whether certain scientific knowledge would indeed assist the trier of fact in the manner contemplated by Rule 702. "This entails a preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue." This preliminary assessment can turn on whether something has been tested, whether an idea has been subjected to scientific peer review or published in scientific journals, the rate of error involved in the technique, and even general acceptance, among other things. It focuses on methodology and principles, not the ultimate conclusions generated.

The Court stressed that the new standard under Rule 702 was rooted in the judicial process and intended to be distinct and separate from the search for scientific truth. "Scientific conclusions are subject to perpetual revision. Law, on the other hand, must resolve disputes finally and quickly. The scientific project is advanced by broad and wide-ranging consideration of a multitude of hypotheses, for those that are incorrect will eventually be shown to be so, and that in itself is an advance." Rule 702 was intended to resolve legal disputes, and thus had to be interpreted in conjunction with other rules of evidence and with other legal means of ending those disputes. Cross examination within the adversary process is adequate to help legal decision makers arrive at efficient ends to disputes. "We recognize that, in practice, a gate-keeping role for the judge, no matter how flexible, inevitably on occasion will prevent the jury from learning of authentic insights and innovations. That, nevertheless, is the balance that is struck by Rules of Evidence designed not for the exhaustive search for cosmic understanding but for the particularized resolution of legal disputes."

As noted in the various opinions and *amicus* brief given in *Barefoot v. Estelle*, the jury in considering whether the death penalty should be imposed, has to answer affirmatively one question: whether there was a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society. The use of the word "probability" without specifying any further size seems odd to say the least, but Texas courts have steadfastly refused to delimit it any further. So, presumably a very small probability of future violence would be sufficient for execution if this small probability could be proved "beyond a reasonable doubt."

The point of much of the current discussion has been to emphasize that actuarial evidence about future violence involving variables such as age, race, or sex, is all there really is in making such predictions. More pointedly, the assignment of a clinical label, such as "sociopath," adds nothing to an ability to predict, and to suggest that it does is to use the worst "junk science," even though it may be routinely assumed true in the larger society. All we have to rely on is the usual psychological adage that the best predictor of future behavior is past behavior. Thus, the best predictor of criminal recidivism is a history of such behavior, and past violence suggests future violence. The greater the amount of past criminal behavior or violence, the more likely that such future behavior or violence will occur (a behavioral form of a "dose-response" relationship). At its basis, this is statistical evidence of such a likely occurrence and no medical or psychological diagnosis is needed or useful.

Besides the specious application of a sociopath diagnosis to predict future violence, after the Supreme Court decision in *Estelle v. Smith* (1981) such a diagnosis had to be made on the basis of a hypothetical question and not on an actual psychological examination of the defendant. In addition to a 100% incontrovertible assurance of future violence, offering testimony without actually examining a defendant proved to be Grigson's eventual downfall and one reason for the expulsion from his professional psychiatric societies. This prevention of an actual examination of a defendant by the Supreme Court case, *Estelle v. Smith* (1981), also involved James Grigson. Ernest Smith, indicted for murder, had been examined by Grigson in jail and who determined he was competent to stand trial. In the psychiatric report on Smith, Grigson termed him "a severe sociopath" but gave no other statements as to future dangerousness. Smith was sentenced to death based on the sociopath label given by Grigson. In *Estelle v. Smith* the Supreme Court held that because of the well-known case of *Miranda v. Arizona* (1966), the state could not force a defendant to submit to a psychiatric examination for the purposes of sentencing because it violated a defendant's Fifth Amendment rights against self-incrimination and the Sixth Amendment right to counsel. Thus, the examination of Ernest Smith was inadmissible at sentencing. From that point on, predictions of violence were made solely on hypothetical questions and Grigson's belief that a labeling as a sociopath was sufficient to guarantee future violence on the part of a defendant, and therefore, the defendant should be put to death.

The offering of a professional psychiatric opinion about an individual without direct examination is an ethical violation of the Goldwater Rule, named for the Arizona Senator who ran for President in 1964 as a Republican. Promulgated by the American Psychiatric Association in 1971, it delineated a set of requirements for communication with the media about the state of mind of individuals. The Goldwater Rule was the result of a special September/October 1964 issue of *Fact:* magazine, published by the highly provocative Ralph Ginzburg. The issue title was "The Unconscious of a Conservative: Special Issue on the Mind of Barry Goldwater," and reported on a mail survey of 12,356 psychiatrists, of whom 2,417 responded: 24% said they did not know enough about Goldwater to answer the question; 27% said he was mentally fit; 49% said he was not. Much was made of Goldwater's "two nervous breakdowns," because such a person should obviously never be President because of a risk of recurrence under stress that might then lead to pressing the nuclear button.

Goldwater brought a $2 million libel suit against *Fact:* and its publisher, Ginzburg. In 1970 the United States Supreme Court decided in Goldwater's favor giving him $1 in compensatory damages and $75,000 in punitive damages. More importantly, it set a legal precedent that changed medical ethics forever. For an updated discussion of the Goldwater Rule, this time because of the many psychiatrists commenting on the psychological makeup of the former chief of the International Monetary Fund, Dominique Strauss-Kahn, after his arrest on sexual assault charges in New York, see Richard A. Friedman's article, "How a Telescopic Lens Muddles Psychiatric Insights" (*New York Times*, May 23, 2011).[2]

References

[1] Kozol, H. L., Boucher, R. J., & Garofalo, R. (1972). The diagnosis and treatment of dangerousness. *Crime and Delinquency*, *18*, 371–392.

[2] Meehl, P., & Rosen, A. (1955). Antecendent probabilitiy and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194–215.

[2] For a recent and thorough review of the literature on the prediction of dangerous or violent behavior as it relates to the death penalty, see Michael L. Perlin, *Mental Disability and the Death Penalty: The Shame of the States* (2013; Rowman & Littlefield); Chapter 3 is particularly relevant: "Future Dangerousness and the Death Penalty." A good resource generally for material on the prediction of dangerous behavior and related forensic matters is the Texas Defender Service (`www.texasdefender.org`), and the publications it has freely available at its web site:

A State of Denial: Texas Justice and the Death Penalty (2000)

Deadly Speculation: Misleading Texas Capital Juries with False Predictions of Future Dangerousness (2004)

Minimizing Risk: A Blueprint for Death Penalty Reform in Texas (2005)

[3] Monahan, J. (1973). Dangerous offenders: A critique of Kozol et al. *Crime and Delinquency*, *19*, 418–420.

[4] Monahan, J. (1977). *The Clinical Prediction of Violent Behavior*. Lanham, Maryland: Jason Aronson, Inc.

[5] Monahan, J., Steadman, H., et al. (2001). *Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence*. New York: Oxford University Press.

[6] Monahan, J., et al. (2005). An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services*, *56*, 810–815.

[7] Wenk, E. A., Robison, J. O., & Smith, G. W. (1972). Can violence be predicted? *Crime and Delinquency*, *18*, 393–402.

5 Appendix: Continuation of the American Psychiatric Association, *Amicus Curiae* Brief: Barefoot v. Estelle

INTRODUCTION AND SUMMARY OF ARGUMENT

The questions presented in this case are the logical outgrowth of two prior decisions by this Court. In the first, Jurek v. Texas, the Court dealt with the same Texas capital sentencing procedure involved here. The Court there rejected a constitutional challenge to the "future dangerousness" question, ruling that the statutory standard was not impermissibly vague. Although recognizing the difficulty inherent in predicting future behavior, the Court held that "[t]he task that [the] jury must perform ... is basically no different from the task performed countless times each day throughout the American system of criminal justice." The Jurek Court thus upheld the use of the Texas statutory question, but did not consider the types of evidence that could be presented to the jury for purposes of this determination.

Subsequently in Estelle v. Smith, the Court again dealt with the Texas sentencing scheme—this time in the context of a psychiatric examination to determine the defendant's competency to stand trial. The Court held that the Fifth Amendment's privilege against self-incrimination applied to such psychiatric examinations, at least to the extent that a prosecution psychiatrist later testifies concerning the defendant's future dangerousness. The Court reasoned that although a defendant has no generalized constitutional right to remain silent at a psychiatric examination properly limited to the issues of sanity or competency, full Miranda warnings must be given with respect to testimony concerning future dangerousness because of "the gravity of the decision to be made at the penalty phase ... " The Smith decision thus enables a capital defendant to bar a government psychiatric examination on the issue of future dangerousness.

The [present] case raises the two issues left unresolved in Jurek and Smith. These are, first, whether a psychiatrist, testifying as an expert medical witness, may ever be permitted to render a prediction as to a capital defendant's long-term future dangerousness. The second issue is whether such testimony may be elicited on the basis of hypothetical questions, even if there exists no general prohibition against the use of expert psychiatric testimony on the issue of long-term future dangerousness. *Amicus* believes that both of these questions should be answered in the negative.

I. Psychiatrists should not be permitted to offer a prediction concerning the long-term future dangerousness of a defendant in a capital case, at least in those circumstances where the psychiatrist purports to be testifying as a medical expert possessing predictive expertise in this area. Although psychiatric assessments may permit short-term predictions of violent or assaultive behavior, medical knowledge has simply not advanced to the point where long-term predictions—the type of testimony at issue in this case—may be made with even reasonable accuracy. The large body of research in this area indicates that, even under the best of conditions, psychiatric predictions of long-term future dangerousness are wrong in at least two out of every three cases.

The forecast of future violent conduct on the part of a defendant in a capital case is, at bottom, a lay determination, not an expert psychiatric determination. To the extent such predictions have any validity, they can only be made on the basis of essentially actuarial data to which psychiatrists, qua psychiatrists, can bring no special interpretative skills. On the other hand, the use of psychiatric testimony on this issue causes serious prejudice to the defendant. By dressing up the actuarial data with an "expert" opinion, the psychiatrist's testimony is likely to receive undue weight. In addition, it permits the jury to avoid the difficult actuarial questions by seeking refuge in a medical diagnosis that provides a false aura of certainty. For these reasons, psychiatric testimony on future dangerousness impermissibly distorts the fact-finding process in capital cases.

II. Even if psychiatrists under some circumstances are allowed to render an expert medical opinion on the question of future dangerousness, *amicus* submits that they should never be permitted to do so unless they have conducted a psychiatric examination of the defendant. It is evident from the testimony in this case that the key clinical determination relied upon by both psychiatrists was their diagnosis of "sociopathy" or "antisocial personality disorder." However, such a diagnosis simply cannot be made on the basis of a hypothetical question. Absent an in-depth psychiatric examination and evaluation, the psychiatrist cannot exclude alternative diagnoses; nor can he assure that the necessary criteria for making the diagnosis in question are met. As a result, he is unable to render a medical opinion with a reasonable degree of certainty.

These deficiencies strip the psychiatric testimony of all value in the present context. Even assuming that the diagnosis of antisocial personality disorder is probative of future dangerousness—an assumption which we do not accept—it is nonetheless clear that the limited facts given in the hypothetical fail to disprove other illnesses that plainly do not indicate a general propensity to commit criminal acts. Moreover, these other illnesses may be more amenable to treatment—a factor that may further reduce the likelihood of future aggressive behavior by the defendant.

...

6 Appendix: Opinion and Dissent in the U.S. Supreme Court, Barefoot v. Estelle (Decided, July 6, 1983)

Summary of the majority opinion:

(a) There is no merit to petitioner's argument that psychiatrists, individually and as a group, are incompetent to predict with an acceptable degree of reliability that a particular criminal will commit other crimes in the future, and so represent a danger to the community. To accept such an argument would call into question predictions of future behavior that are constantly made in other contexts. Moreover, under the generally applicable rules of evidence covering the admission and weight of unprivileged evidence, psychiatric testimony predicting dangerousness may be countered not only as erroneous in a particular case but also as generally so unreliable that it should be ignored. Nor, despite the view of the American Psychiatric Association supporting petitioner's view, is there any convincing evidence that such testimony is almost entirely unreliable, and that the factfinder and the adversary system will not be competent to uncover, recognize, and take due account of its shortcomings.

(b) Psychiatric testimony need not be based on personal examination of the defendant, but may properly be given in response to hypothetical questions. Expert testimony, whether in the form of an opinion based on hypothetical questions or otherwise, is commonly admitted as evidence where it might help the factfinder do its job. Although this case involves the death penalty, there is no constitutional barrier to applying the ordinary rules of evidence governing the use of expert testimony.

. . .

Justice Blackmun dissenting:

I agree with most of what Justice Marshall has said in his dissenting opinion. I, too, dissent, but I base my conclusion also on evidentiary factors that the Court rejects with some emphasis. The Court holds that psychiatric testimony about a defendant's future dangerousness is admissible, despite the fact that such testimony is wrong two times out of three. The Court reaches this result—even in a capital case—because, it is said, the testimony is subject to cross-examination and impeachment. In the present state of psychiatric knowledge, this is too much for me. One may accept this in a routine lawsuit for money damages, but when a person's life is at stake—no matter how heinous his offense—a requirement of greater reliability should prevail. In a capital case, the specious testimony of a psychiatrist, colored in the eyes of an impressionable jury by the inevitable untouchability of a medical specialist's words, equates with death itself.

To obtain a death sentence in Texas, the State is required to prove beyond a reasonable doubt that "there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society." As a practical matter, this prediction of future dangerousness was the only issue to be decided by Barefoot's sentencing jury.

At the sentencing hearing, the State established that Barefoot had two prior convictions for drug offenses and two prior convictions for unlawful possession of firearms. None of these convictions involved acts of violence. At the guilt stage of the trial, for the limited purpose of establishing that the crime was committed in order to evade police custody, the State had presented evidence that Barefoot had escaped from jail in New Mexico where he was being held on charges of statutory rape and unlawful restraint of a minor child with intent to commit sexual penetration against the child's will. The prosecution also called several character witnesses at the sentencing hearing, from towns in five States. Without mentioning particular examples of Barefoot's conduct, these witnesses testified that Barefoot's reputation for being a peaceable and law-abiding citizen was bad in their respective communities.

Last, the prosecution called Doctors Holbrook and Grigson, whose testimony extended over more than half the hearing. Neither had examined Barefoot or requested the opportunity to examine him. In the presence of the jury, and over defense counsel's objection, each was qualified as an expert psychiatrist witness. Doctor Holbrook detailed at length his training and experience as a psychiatrist, which included a position as chief of psychiatric services at the Department of Corrections. He explained that he had previously performed many "criminal evaluations," and that he subsequently took the post at the Department of Corrections to observe the subjects of these evaluations so that he could "be certain those opinions that [he] had were accurate at the time of trial and pretrial." He then informed the jury that it was "within [his] capacity as a doctor of psychiatry to predict the future dangerousness of an individual within a reasonable medical certainty," and that he could give

> "an expert medical opinion that would be within reasonable psychiatric certainty as to whether or not that individual would be dangerous to the degree that there would be a probability that that person would commit criminal acts of violence in the future that would constitute a continuing threat to society."

Doctor Grigson also detailed his training and medical experience, which, he said, included examination of "between thirty and forty thousand individuals," including 8,000 charged with felonies, and at least 300 charged with murder. He testified that, with enough information, he would be able to "give a medical opinion within reasonable psychiatric certainty as to the psychological or psychiatric makeup of an individual," and that this skill was "particular to the field of psychiatry, and not to the average layman."

Each psychiatrist then was given an extended hypothetical question asking him to assume as true about Barefoot the four prior convictions for nonviolent offenses, the bad reputation for being law-abiding in various communities, the

New Mexico escape, the events surrounding the murder for which he was on trial and, in Doctor Grigson's case, the New Mexico arrest. On the basis of the hypothetical question, Doctor Holbrook diagnosed Barefoot "within a reasonable psychiatr[ic] certainty," as a "criminal sociopath." He testified that he knew of no treatment that could change this condition, and that the condition would not change for the better but "may become accelerated" in the next few years. Finally, Doctor Holbrook testified that, "within reasonable psychiatric certainty," there was "a probability that the Thomas A. Barefoot in that hypothetical will commit criminal acts of violence in the future that would constitute a continuing threat to society," and that his opinion would not change if the "society" at issue was that within Texas prisons, rather than society outside prison.

Doctor Grigson then testified that, on the basis of the hypothetical question, he could diagnose Barefoot "within reasonable psychiatric certainty" as an individual with "a fairly classical, typical, sociopathic personality disorder." He placed Barefoot in the "most severe category of sociopaths (on a scale of one to ten, Barefoot was "above ten"), and stated that there was no known cure for the condition. Finally, Doctor Grigson testified that whether Barefoot was in society at large or in a prison society there was a "one hundred percent and absolute" chance that Barefoot would commit future acts of criminal violence that would constitute a continuing threat to society.

On cross-examination, defense counsel questioned the psychiatrists about studies demonstrating that psychiatrists' predictions of future dangerousness are inherently unreliable. Doctor Holbrook indicated his familiarity with many of these studies, but stated that he disagreed with their conclusions. Doctor Grigson stated that he was not familiar with most of these studies, and that their conclusions were accepted by only a "small minority group" of psychiatrists—"[i]t's not the American Psychiatric Association that believes that.

After an hour of deliberation, the jury answered "yes" to the two statutory questions, and Thomas Barefoot was sentenced to death.

The American Psychiatric Association (APA), participating in this case as *amicus curiae*, informs us that "[t]he unreliability of psychiatric predictions of long-term future dangerousness is by now an established fact within the profession." The APA's best estimate is that two out of three predictions of long-term future violence made by psychiatrists are wrong. The Court does not dispute this proposition, and indeed it could not do so; the evidence is overwhelming. For example, the APA's Draft Report of the Task Force on the Role of Psychiatry in the Sentencing Process (1983) states that

> "[c]onsiderable evidence has been accumulated by now to demonstrate that long-term prediction by psychiatrists of future violence is an extremely inaccurate process."

John Monahan, recognized as "the leading thinker on this issue" even by the State's expert witness at Barefoot's federal habeas corpus hearing, concludes that

> "the 'best' clinical research currently in existence indicates that psychiatrists and psychologists are accurate in no more than one out of three predictions of violent behavior,"

even among populations of individuals who are mentally ill and have committed violence in the past. Another study has found it impossible to identify any subclass of offenders "whose members have a greater-than-even chance of engaging again in an assaultive act." Yet another commentator observes:

> "In general, mental health professionals ... are more likely to be wrong than right when they predict legally relevant behavior. When predicting violence, dangerousness, and suicide, they are far more likely to be wrong than right."

Neither the Court nor the State of Texas has cited a single reputable scientific source contradicting the unanimous conclusion of professionals in this field that psychiatric predictions of long-term future violence are wrong more often than they are right.

The APA also concludes, as do researchers that have studied the issue, that psychiatrists simply have no expertise in predicting long-term future dangerousness. A layman with access to relevant statistics can do at least as well, and possibly better; psychiatric training is not relevant to the factors that validly can be employed to make such predictions, and psychiatrists consistently err on the side of overpredicting violence. Thus, while Doctors Grigson and Holbrook were presented by the State and by self-proclamation as experts at predicting future dangerousness, the scientific literature makes crystal clear that they had no expertise whatever. Despite their claims that they were able to predict Barefoot's future behavior "within reasonable psychiatric certainty," or to a "one hundred percent and absolute" certainty, there was, in fact, no more than a one in three chance that they were correct.[3]

It is impossible to square admission of this purportedly scientific but actually baseless testimony with the Constitution's paramount concern for reliability in capital sentencing.[4] Death is a permissible punishment in Texas

[3] Like the District Court ... and the Court of Appeals, ... the Court seeks to justify the admission of psychiatric testimony on the ground that

> "[t]he majority of psychiatric experts agree that where there is a pattern of repetitive assaultive and violent conduct, the accuracy of psychiatric predictions of future dangerousness dramatically rises."

... The District Court correctly found that there is empirical evidence supporting the common sense correlation between repetitive past violence and future violence; the APA states that

> "[t]he most that can be said about any individual is that a history of past violence increases the probability that future violence will occur."

But psychiatrists have no special insights to add to this actuarial fact, and a single violent crime cannot provide a basis for a reliable prediction of future violence. ...

The lower courts and this Court have sought solace in this statistical correlation without acknowledging its obvious irrelevance to the facts of this case. The District Court did not find that the State demonstrated any pattern of repetitive assault and violent conduct by Barefoot. Recognizing the importance of giving some credibility to its experts' specious prognostications, the State now claims that the "reputation" testimony adduced at the sentencing hearing "can only evince repeated, widespread acts of criminal violence." ... This is simply absurd. There was no testimony worthy of credence that Barefoot had committed acts of violence apart from the crime for which he was being tried; there was testimony only of a bad reputation for peaceable and law-abiding conduct. In light of the fact that each of Barefoot's prior convictions was for a nonviolent offense, such testimony obviously could have been based on antisocial but nonviolent behavior. Neither psychiatrist informed the jury that he considered this reputation testimony to show a history of repeated acts of violence. Moreover, if the psychiatrists or the jury were to rely on such vague hearsay testimony in order to show a "pattern of repetitive assault and violent conduct," Barefoot's death sentence would rest on information that might "bear no closer relation to fact than the average rumor or item of gossip," ... and should be invalid for that reason alone. A death sentence cannot rest on highly dubious predictions secretly based on a factual foundation of hearsay and pure conjecture. ...

[4] Although I believe that the misleading nature of any psychiatric prediction of future violence violates due process when introduced in a capital sentencing hearing, admitting

only if the jury finds beyond a reasonable doubt that there is a probability the defendant will commit future acts of criminal violence. The admission of unreliable psychiatric predictions of future violence, offered with unabashed claims of "reasonable medical certainty" or "absolute" professional reliability, creates an intolerable danger that death sentences will be imposed erroneously.

The plurality in Woodson v. North Carolina, stated:

"Death, in its finality, differs more from life imprisonment than a 100-year prison term differs from one of only a year or two. Because of that qualitative difference, there is a corresponding difference in the need for reliability in the determination that death is the appropriate punishment in a specific case." The Court does not see fit to mention this principle today, yet it is as firmly established as any in our Eighth Amendment jurisprudence. Only two weeks ago, in Zant v. Stephens, the Court described the need for reliability in the application of the death penalty as one of the basic "themes ... reiterated in our opinions discussing the procedures required by the Constitution in capital sentencing determinations." (capital punishment must be "imposed fairly, and with reasonable consistency, or not at all"). State evidence rules notwithstanding, it is well established that, because the truth-seeking process may be unfairly skewed, due process may be violated even in a noncapital criminal case by the exclusion of evidence probative of innocence, or by the admission of certain categories of unreliable and prejudicial evidence ("[i]t is the reliability of identification evidence that primarily determines its admissi-

the predictions in this case—which were made without even examining the defendant—was particularly indefensible. In the APA's words, if prediction following even an in-depth examination is inherently unreliable,

"there is all the more reason to shun the practice of testifying without having examined the defendant at all. ... Needless to say, responding to hypotheticals is just as fraught with the possibility of error as testifying in any other way about an individual whom one has not personally examined. Although the courts have not yet rejected the practice, psychiatrists should."

... Such testimony is offensive not only to legal standards; the APA has declared that "[i]t is unethical for a psychiatrist to offer a professional opinion unless he/she has conducted an examination." ... The Court today sanctions admission in a capital sentencing hearing of "expert" medical testimony so unreliable and unprofessional that it violates the canons of medical ethics.

bility"). The reliability and admissibility of evidence considered by a capital sentencing factfinder is obviously of still greater constitutional concern.

The danger of an unreliable death sentence created by this testimony cannot be brushed aside on the ground that the "jury [must] have before it all possible relevant information about the individual defendant whose fate it must determine." Although committed to allowing a "wide scope of evidence" at presentence hearings, the Court has recognized that "consideration must be given to the quality, as well as the quantity, of the information on which the sentencing [authority] may rely." Thus, very recently, this Court reaffirmed a crucial limitation on the permissible scope of evidence: "[s]o long as the evidence introduced ... do[es] not prejudice a defendant, it is preferable not to impose restrictions." The Court all but admits the obviously prejudicial impact of the testimony of Doctors Grigson and Holbrook; granting that their absolute claims were more likely to be wrong than right, the Court states that "[t]here is no doubt that the psychiatric testimony increased the likelihood that petitioner would be sentenced to death." Indeed, unreliable scientific evidence is widely acknowledged to be prejudicial. The reasons for this are manifest. "The major danger of scientific evidence is its potential to mislead the jury; an aura of scientific infallibility may shroud the evidence, and thus lead the jury to accept it without critical scrutiny."[5]

[5] There can be no dispute about this obvious proposition:

"Scientific evidence impresses lay jurors. They tend to assume it is more accurate and objective than lay testimony. A juror who thinks of scientific evidence visualizes instruments capable of amazingly precise measurement, of findings arrived at by dispassionate scientific tests. In short, in the mind of the typical lay juror, a scientific witness has a special aura of credibility."

... "Scientific ... evidence has great potential for misleading the jury. The low probative worth can often be concealed in the jargon of some expert ... " This danger created by use of scientific evidence frequently has been recognized by the courts. Speaking specifically of psychiatric predictions of future dangerousness similar to those at issue, one District Court has observed that, when such a prediction

"is proffered by a witness bearing the title of 'Doctor,' its impact on the jury is much greater than if it were not masquerading as something it is not."

... In United States v. Addison, the court observed that scientific evidence may "assume a posture of mystic infallibility in the eyes of a jury of laymen." Another court has noted that scientific evidence "is likely to be shrouded with an aura of near infallibility, akin to the ancient oracle of Delphi." ...

Where the public holds an exaggerated opinion of the accuracy of scientific testimony, the prejudice is likely to be indelible. There is little question that psychiatrists are perceived by the public as having a special expertise to predict dangerousness, a perception based on psychiatrists' study of mental disease. It is this perception that the State in Barefoot's case sought to exploit. Yet mental disease is not correlated with violence, and the stark fact is that no such expertise exists. Moreover, psychiatrists, it is said, sometimes attempt to perpetuate this illusion of expertise, and Doctors Grigson and Holbrook—who purported to be able to predict future dangerousness "within reasonable psychiatric certainty," or absolutely—present extremely disturbing examples of this tendency. The problem is not uncommon.

Furthermore, as is only reasonable, the Court's concern in encouraging the introduction of a wide scope of evidence has been to ensure that accurate information is provided to the sentencing authority without restriction. The joint opinion announcing the judgment in Gregg explained the jury's need for relevant evidence in these terms:

"If an experienced trial judge, who daily faces the difficult task of imposing sentences, has a vital need for accurate information . . . to be able to impose a rational sentence in the typical criminal case, then accurate sentencing information is an indispensable prerequisite to a reasoned determination of whether a defendant shall live or die by a jury of people who may never before have made a sentencing decision."

So far as I am aware, the Court never has suggested that there is any interest in providing deceptive and inaccurate testimony to the jury. Psychiatric predictions of future dangerousness are not accurate; wrong two times out of three, their probative value, and therefore any possible contribution they might make to the ascertainment of truth, is virtually nonexistent (psychiatric testimony not sufficiently reliable to support finding that individual will be dangerous under any standard of proof). Indeed, given a psychiatrist's prediction that an individual will be dangerous, it is more likely than not that the defendant will not commit further violence. It is difficult to understand how the admission of such predictions can be justified as advancing the search for truth, particularly in light of their clearly prejudicial effect. Thus, the Court's remarkable observation that "[n]either petitioner nor the [APA] suggests that psychiatrists are always wrong with respect to future dangerousness, only most of the time," misses the point completely, and its claim that this testimony was no more problematic than "other relevant evidence against any defendant in a criminal case," is simply incredible. Surely, this Court's commitment to ensuring that death sentences are imposed reliably and reasonably requires that nonprobative and highly prejudicial testimony on the ultimate question of life or death be excluded from a capital sentencing hearing.

Despite its recognition that the testimony at issue was probably wrong and certainly prejudicial, the Court holds this testimony admissible because the Court is

> "unconvinced . . . that the adversary process cannot be trusted to sort out the reliable from the unreliable evidence and opinion about future dangerousness."

One can only wonder how juries are to separate valid from invalid expert opinions when the "experts" themselves are so obviously unable to do so. Indeed, the evidence suggests that juries are not effective at assessing the validity of scientific evidence.

There can be no question that psychiatric predictions of future violence will have an undue effect on the ultimate verdict. Even judges tend to accept psychiatrists' recommendations about a defendant's dangerousness with little regard for cross-examination or other testimony. The American Bar Association has warned repeatedly that sentencing juries are particularly incapable of dealing with information relating to "the likelihood that the defendant will commit other crimes," and similar predictive judgments. Relying on the ABA's conclusion, the joint opinion announcing the judgment in Gregg v. Georgia, recognized that,

> "[s]ince the members of a jury will have had little, if any, previous experience in sentencing, they are unlikely to be skilled in dealing with the information they are given."

But the Court in this case, in its haste to praise the jury's ability to find the truth, apparently forgets this well-known and worrisome shortcoming.

As if to suggest that petitioner's position that unreliable expert testimony should be excluded is unheard of in the law, the Court relies on the proposition that the rules of evidence generally

> "anticipate that relevant, unprivileged evidence should be admitted and its weight left to the factfinder, who would have the benefit of cross-examination and contrary evidence by the opposing party."

But the Court simply ignores hornbook law that, despite the availability of cross-examination and rebuttal witnesses,

> "opinion evidence is not admissible if the court believes that the state of the pertinent art or scientific knowledge does not permit a reasonable opinion to be asserted."

Because it is feared that the jury will overestimate its probative value, polygraph evidence, for example, almost invariably is excluded from trials despite the fact that, at a conservative estimate, an experienced polygraph examiner can detect truth or deception correctly about 80 to 90 percent of the time. In no area is purportedly "expert" testimony admitted for the jury's consideration where it cannot be demonstrated that it is correct more often than not. "It is inconceivable that a judgment could be considered an expert' judgment when it is less accurate than the flip of a coin." The risk that a jury will be incapable of separating "scientific" myth from reality is deemed unacceptably high.[6]

The Constitution's mandate of reliability, with the stakes at life or death, precludes reliance on cross-examination and the opportunity to present rebuttal witnesses as an antidote for this distortion of the truthfinding process. Cross-examination is unlikely to reveal the fatuousness of psychiatric predictions because such predictions often rest, as was the case here, on psychiatric categories and intuitive clinical judgments not susceptible to cross-examination and rebuttal. Psychiatric categories have little or no demon-

[6] The Court observes that this well-established rule is a matter of evidence law, not constitutional law. ... But the principle requiring that capital sentencing procedures ensure reliable verdicts, which the Court ignores, and the principle that due process is violated by the introduction of certain types of seemingly conclusive, but actually unreliable, evidence, ... which the Court also ignores, are constitutional doctrines of long standing. The teaching of the evidence doctrine is that unreliable scientific testimony creates a serious and unjustifiable risk of an erroneous verdict, and that the adversary process, at its best, does not remove this risk. We should not dismiss this lesson merely by labeling the doctrine nonconstitutional; its relevance to the constitutional question before the Court could not be more certain.

strated relationship to violence, and their use often obscures the unimpressive statistical or intuitive bases for prediction.[7] The APA particularly condemns the use of the diagnosis employed by Doctors Grigson and Holbrook in this case, that of sociopathy:

"In this area confusion reigns. The psychiatrist who is not careful can mislead the judge or jury into believing that a person has a major mental disease simply on the basis of a description of prior criminal behavior. Or a psychiatrist can mislead the court into believing that an individual is devoid of conscience on the basis of a description of criminal acts alone. ... The profession of psychiatry has a responsibility to avoid inflicting this confusion upon the courts, and to spare the defendant the harm that may result. ... Given our uncertainty about the implications of the finding, the diagnosis of sociopathy ... should not be used to justify or to support predictions of future conduct. There is no certainty in this area."

It is extremely unlikely that the adversary process will cut through the facade of superior knowledge. The Chief Justice [Burger] long ago observed:

"The very nature of the adversary system ... complicates the use of scientific opinion evidence, particularly in the field of psychiatry. This system of partisan contention, of attack and counterattack, at its best is not ideally suited to developing an accurate portrait or profile of the human personality, especially in the area of abnormal behavior. Although under ideal conditions the adversary system can develop for a jury most of the necessary fact material for an adequate decision, such conditions are rarely achieved in the courtrooms in this country. These ideal conditions would include a highly skilled and experienced trial judge and highly skilled lawyers on both sides of the case, all of whom, in addition to being well-trained in the law and in the techniques of advocacy, would be sophisticated in matters of medicine, psychiatry, and psychology. It is far too rare that all three of the legal actors in the cast meet these standards."

[7] In one study, for example, the only factor statistically related to whether psychiatrists predicted that a subject would be violent in the future was the type of crime with which the subject was charged. Yet the defendant's charge was mentioned by the psychiatrists to justify their predictions in only one-third of the cases. The criterion most frequently cited was "delusional or impaired thinking." ...

Another commentator has noted:

"Competent cross-examination and jury instructions may be partial antidotes ... but they cannot be complete. Many of the cases are not truly adversarial; too few attorneys are skilled at cross-examining psychiatrists, laypersons overweigh the testimony of experts, and, in any case, unrestricted use of experts promotes the incorrect view that the questions are primarily scientific. There is, however, no antidote for the major difficulty with mental health 'experts'—that they simply are not experts. ... In realms beyond their true expertise, the law has little special to learn from them; too often, their testimony is ... prejudicial."

Nor is the presentation of psychiatric witnesses on behalf of the defense likely to remove the prejudicial taint of misleading testimony by prosecution psychiatrists. No reputable expert would be able to predict with confidence that the defendant will not be violent; at best, the witness will be able to give his opinion that all predictions of dangerousness are unreliable. Consequently, the jury will not be presented with the traditional battle of experts with opposing views on the ultimate question. Given a choice between an expert who says that he can predict with certainty that the defendant, whether confined in prison or free in society, will kill again, and an expert who says merely that no such prediction can be made, members of the jury, charged by law with making the prediction, surely will be tempted to opt for the expert who claims he can help them in performing their duty, and who predicts dire consequences if the defendant is not put to death.[8]

Moreover, even at best, the presentation of defense psychiatrists will convert the death sentence hearing into a battle of experts, with the Eighth

[8] "Although jurors may treat mitigating psychiatric evidence with skepticism, they may credit psychiatric evidence demonstrating aggravation. Especially when jurors' sensibilities are offended by a crime, they may seize upon evidence of dangerousness to justify an enhanced sentence." ... Thus, the danger of jury deference to expert opinions is particularly acute in death penalty cases. Expert testimony of this sort may permit juries to avoid the difficult and emotionally draining personal decisions concerning rational and just punishment. ... Doctor Grigson himself has noted both the superfluousness and the misleading effect of his testimony: "I think you could do away with the psychiatrist in these cases. Just take any man off the street, show him what the guy's done, and most of these things are so clear-cut he would say the same things I do. But I think the jurors feel a little better when a psychiatrist says it—somebody that's supposed to know more than they know." ...

Amendment's well-established requirement of individually focused sentencing a certain loser. The jury's attention inevitably will turn from an assessment of the propriety of sentencing to death the defendant before it to resolving a scientific dispute about the capabilities of psychiatrists to predict future violence. In such an atmosphere, there is every reason to believe that the jury may be distracted from its constitutional responsibility to consider "particularized mitigating factors," in passing on the defendant's future dangerousness.

One searches the Court's opinion in vain for a plausible justification for tolerating the State's creation of this risk of an erroneous death verdict. As one Court of Appeals has observed:

"A courtroom is not a research laboratory. The fate of a defendant ... should not hang on his ability to successfully rebut scientific evidence which bears an 'aura of special reliability and trustworthiness,' although, in reality, the witness is testifying on the basis of an unproved hypothesis ... which has yet to gain general acceptance in its field."

Ultimately, when the Court knows full well that psychiatrists' predictions of dangerousness are specious, there can be no excuse for imposing on the defendant, on pain of his life, the heavy burden of convincing a jury of laymen of the fraud.[9]

The Court is simply wrong in claiming that psychiatric testimony respect-

[9] The Court is far wide of the mark in asserting that excluding psychiatric predictions of future dangerousness from capital sentencing proceedings "would immediately call into question those other contexts in which predictions of future behavior are constantly made." ... Short-term predictions of future violence, for the purpose of emergency commitment or treatment, are considerably more accurate than long-term predictions. In other contexts where psychiatric predictions of future dangerousness are made, moreover, the subject will not be criminally convicted, much less put to death, as a result of predictive error. The risk of error therefore may be shifted to the defendant to some extent. ... The APA, discussing civil commitment proceedings based on determinations of dangerousness, states that, in light of the unreliability of psychiatric predictions, "[c]lose monitoring, frequent follow-up, and a willingness to change one's mind about treatment recommendations and dispositions for violent persons, whether within the legal system or without, is the only acceptable practice if the psychiatrist is to play a helpful role in these assessments of dangerousness." ... In a capital case, there will be no chance for "follow-up" or "monitoring." A subsequent change of mind brings not justice delayed, but the despair of irreversible error. ...

ing future dangerousness is necessarily admissible in light of Jurek v. Texas, or Estelle v. Smith. As the Court recognizes, Jurek involved "only lay testimony." Thus, it is not surprising that "there was no suggestion by the Court that the testimony of doctors would be inadmissible," and it is simply irrelevant that the Jurek Court did not "disapprov[e]" the use of such testimony. In Smith, the psychiatric testimony at issue was given by the same Doctor Grigson who confronts us in this case, and his conclusions were disturbingly similar to those he rendered here. The APA, appearing as *amicus curiae*, argued that all psychiatric predictions of future dangerousness should be excluded from capital sentencing proceedings. The Court did not reach this issue, because it found Smith's death sentence invalid on narrower grounds: Doctor Grigson's testimony had violated Smith's Fifth and Sixth Amendment right. Contrary to the Court's inexplicable assertion in this case, Smith certainly did not reject the APA's position. Rather, the Court made clear that "the holding in Jurek was guided by recognition that the inquiry [into dangerousness] mandated by Texas law does not require resort to medical experts." If Jurek and Smith held that psychiatric predictions of future dangerousness are admissible in a capital sentencing proceeding as the Court claims, this guiding recognition would have been irrelevant.

The Court also errs in suggesting that the exclusion of psychiatrists' predictions of future dangerousness would be contrary to the logic of Jurek. Jurek merely upheld Texas' substantive decision to condition the death sentence upon proof of a probability that the defendant will commit criminal acts of violence in the future. Whether the evidence offered by the prosecution to prove that probability is so unreliable as to violate a capital defendant's rights to due process is an entirely different matter, one raising only questions of fair procedure.[10] Jurek's conclusion that Texas may impose the death penalty on capital defendants who probably will commit criminal acts of violence in no way establishes that the prosecution may convince a jury that this is so by

[10] The Court's focus in the death penalty cases has been primarily on ensuring a fair procedure: "In ensuring that the death penalty is not meted out arbitrarily or capriciously, the Court's principal concern has been more with the procedure by which the State imposes the death sentence than with the substantive factors the State lays before the jury as a basis for imposing death, once it has been determined that the defendant falls within the category of persons eligible for the death penalty."

misleading or patently unreliable evidence.

Moreover, Jurek's holding that the Texas death statute is not impermissibly vague does not lead ineluctably to the conclusion that psychiatric testimony is admissible. It makes sense to exclude psychiatric predictions of future violence while admitting lay testimony, because psychiatric predictions appear to come from trained mental health professionals, who purport to have special expertise. In view of the total scientific groundlessness of these predictions, psychiatric testimony is fatally misleading. Lay testimony, frankly based on statistical factors with demonstrated correlations to violent behavior, would not raise this substantial threat of unreliable and capricious sentencing decisions, inimical to the constitutional standards established in our cases; and such predictions are as accurate as any a psychiatrist could make. Indeed, the very basis of Jurek, as I understood it, was that such judgments can be made by laymen on the basis of lay testimony.

Our constitutional duty is to ensure that the State proves future dangerousness, if at all, in a reliable manner, one that ensures that "any decision to impose the death sentence be, and appear to be, based on reason rather than caprice or emotion." Texas' choice of substantive factors does not justify loading the factfinding process against the defendant through the presentation of what is, at bottom, false testimony.

Module 3: The Analysis of $2 \times 2 \times 2$ (Multiway) Contingency Tables: Explaining Simpson's Paradox and Demonstrating Racial Bias in the Imposition of the Death Penalty

It is the mark of a truly intelligent person to be moved by statistics.
– George Bernard Shaw

Abstract: This module discusses the two major topics of Simpson's paradox and the Supreme Court decision in *McCleskey v. Kemp* (1987). Simpson's paradox is ubiquitous in the misinterpretation of data; it is said to be present whenever a relationship that appears to exist at an aggregated level disappears or reverses when disaggregated and viewed within levels. A common mechanism for displaying data that manifests such a reversal phenomenon is through a multiway contingency table, often of the $2 \times 2 \times 2$ variety. For example, much of the evidence discussed in *McCleskey v. Kemp* was cross-categorized by three dichotomous variables: race of the victim (black or white), race of the defendant (black or white), and whether the death penalty was imposed (yes or no). Despite incontrovertible evidence that the race of the victim plays a significant role in whether the death penalty is imposed, the holding in *McClesky v. Kemp* was as follows: Despite statistical evidence of a profound racial disparity in application of the death penalty, such evidence is insufficient to invalidate defendant's death sentence.

Contents

1 A Few Introductory Examples of Simpson's Paradox

An enjoyable diversion on Saturday mornings is the NPR radio show, *Car Talk*, with Click and Clack, The Tappet Brothers (aka Ray and Tom Magliozzi). A regular feature of the show, besides giving advice on cars, is The Puzzler; a recent example on September 22, 2012 gives a nice introductory example of one main topic of this chapter, Simpson's paradox. It is called, Take Ray Out to the Ball Game, and is stated as follows on the Car Talk website:

Take Ray Out to the Ball Game:

RAY: As you might guess, I'm a baseball fan. And now that the season is in its waning days, I thought I'd use this baseball Puzzler I've been saving.

There are two rookie players, Bluto and Popeye, who started the season on opening day and made a wager as to which one would have the best batting average at the end of the season.

Well, the last day of the season arrives, and not much is going to change–especially considering that neither one of them is in the starting lineup.

Bluto says, “Hey, Popeye, what did you bat for the first half of the year?”

Popeye answers, “I batted .250.”

And Bluto responds, “Well, I got you there. I batted .300. How about after the All-Star break?”

Proudly, Popeye pipes up, “I batted .375.”

Bluto says, “Pretty good, but I batted .400. Fork over the 20 bucks that we bet.”

The bat boy, Dougie, saunters over and says, “Don’t pay the 20 bucks, Popeye. I think you won.”

TOM: Why is someone who batted .375 not playing in the last game of the season? That’s what I want to know!

RAY: Good point. But the question is this: How could Popeye have won?

RAY: Here’s the answer. Let’s assume that they both had 600 at-bats.

TOM: Yeah.

RAY: If Bluto batted .300 for the first half of the season and he had 500 at-bats during that first half of the season.

TOM: Oooh. Yeah.

RAY: He got 150 hits. One hundred fifty over 500 is a .300 average, right?

TOM: Mmm-hmm. So he would have gotten 150.

RAY: Yeah. OK? If Popeye batted .250 and had 100 at-bats, he would have had 25 for 100. The second half of the season, Bluto bats .400. How does he do that? Well, we know he had 500 at-bats in the first half.

TOM: So he's only been up 100 times in the second half of the season.

RAY: And he got 40 hits.

Popeye bats .375.

TOM: But he's up 500 times.

RAY: And he gets 187 and a half hits. One of them was a check-swing single over the infield. They only count that as half a hit. So now, let's ... let's figure it all out.

Bluto batted 600 times. How many total hits did he get?

TOM: 190.

RAY: Right. How about Popeye? How many hits did he get?

TOM: 212 and a half.

RAY: And when you figure that out, Bluto batted .316 for the season. Even though he batted .300 and .400 in each half.

TOM: Yeah.

RAY: And Popeye bats .353 and wins the batting title.

TOM: No kidding!

RAY: Pretty good, huh?

Putting the data about Bluto and Popeye in the form of a 2 × 2 table that gives batting averages both before and after the All-Star break as well as for the full year should help see what is happening:

	Before Break	After Break	Full Year
Bluto	$\frac{150}{500} = .300$	$\frac{40}{100} = .400$	$\frac{190}{600} = .317$
Popeye	$\frac{25}{100} = .250$	$\frac{187.5}{500} = .375$	$\frac{212.5}{600} = .354$

Thus, the batting averages of Popeye before and after the break (.250 and .375) can be less than for Bluto (.300 and .400), even though for the full year, Popeye’s average of .354 is better than Bluto’s .317. This type of counterintuitive situation is referred to as a “reversal paradox” or more usually by the term, “ Simpson’s paradox.”

The unusual phenomenon presented by the example above occurs frequently in the analysis of multiway contingency tables. Basically, various relations that appear to be present when data are conditioned on the levels of one variable, either disappear or change “direction” when aggregation occurs over the levels of the conditioning variable. A well-known real-life example is the Berkeley sex bias case applicable to graduate school (Bickel, Hammel, & O’Connell, 1975). The table below shows the aggregate admission figures for the fall of 1973:

	Number of applicants	Percent admitted
Men	8442	44
Women	4321	35

Given these data, there appears to be a *primae facie* case for bias because a lower percentage of women than men is admitted.

Although a bias seems to be present against women at the aggregate level, the situation becomes less clear when the data are broken down by major. Because no department is significantly biased against women, and in fact, most have a small bias against men, we have another instance of Simpson's paradox. Apparently, women tend to apply to competitive departments with lower rates of admission among qualified applicants (for example, English); men tend to apply to departments with generally higher rates of admission (for example, Engineering).[1]

A different example showing a similar point can be given using data on the differential imposition of a death sentence depending on the race of the defendant and the victim. These data are from twenty Florida counties during 1976-1977 (Radelet, 1981):

	Death Penalty	
Defendant	Yes	No
White	19 (12%)	141
Black	17 (10%)	149

Because 12% of white defendants receive the Death penalty and only 10% of blacks, at this aggregate level there appears to be no bias against blacks. But when the data are disaggregated, the situation appears to change:

[1] A question arises as to whether an argument for bias "falls apart" because of Simpson's paradox. Interesting, in many cases the authors have seen like this, there is a variable that if interpreted in a slightly different way would make a case for bias even at the disaggregated level. Here, why do the differential admission quotas interact with sex? In other words, is it inherently discriminatory to women if the majors to which they apply most heavily are also those with the most limiting admission quotas?

		Death Penalty	
Victim	Defendant	Yes	No
White	White	19 (13%)	132
White	Black	11 (17%)	52
Black	White	0 (0%)	9
Black	Black	6 (6%)	97

When aggregated over victim race, there is a higher percentage of white defendants (12%) receiving the death penalty than black defendants (10%), so apparently, there is a slight race bias against whites. But when looking within the race of the victim, black defendants have the higher percentages of receiving the death sentence compared to white defendants (17% to 13% for white victims; 6% to 0% for black victims). The conclusion is disconcerting: the value of a victim is worth more if white than if black, and because more whites kill whites, there appears to be a slight bias against whites at the aggregate level. But for both types of victims, blacks are more likely to receive the death penalty.[2]

A common way to explain what occurs in Simpson's paradox is to use contingency tables. For convenience, we restrict discussion

[2] Simpson's paradox is a very common occurrence, and even through it can be "explained away" by the influence of differential marginal frequencies, the question remains as to why the differential marginal frequencies are present in the first place. Generally, a case can be made that gives an argument for bias or discrimination in an alternative framework, for example, differential admission quotas or differing values on a life. A more recent study similar to Radelet (1981) is from the *New York Times*, April 20, 2001, reported in a short article by Fox Butterfield, "Victims' Race Affects Decisions on Killers' Sentence, Study Finds."

to the simple 2 × 2 × 2 case, and use the "death penalty" data as an illustration. There are two general approaches based on conditional probabilities. One that is presented below involves weighted averages; the second that we do not discuss relies on the language of events being conditionally positively correlated, but unconditionally negatively correlated (or the reverse).

To set up the numerical example, define three events: A, B, and C:

A: the death penalty is imposed;
B: the defendant is black;
C: the victim is white.

For reference later, we give a collection of conditional probabilities based on frequencies in the 2 × 2 × 2 contingency table:

$P(A|B) = .10$; $P(A|\bar{B}) = .12$; $P(A|B \cap C) = .17$;
$P(A|\bar{B} \cap C) = .13$; $P(A|\bar{B} \cap \bar{C}) = .00$;
$P(C|B) = .38$; $P(\bar{C}|B) = .62$; $P(C|\bar{B}) = .94$;
$P(\bar{C}|\bar{B}) = .38$; $P(C) = .66$; $P(\bar{C}) = .34$.

The explanation for Simpson's paradox based on a weighted average begins by formally stating the paradox through conditional probabilities: It is possible to have

$$P(A|B) < P(A|\bar{B}) \ ,$$

but

$$P(A|B \cap C) \geq P(A|\bar{B} \cap C) \ ;$$

$$P(A|B \cap \bar{C}) \geq P(A|\bar{B} \cap \bar{C}) \ .$$

So, conditioning on the C and $\bar{C}$ events, the relation reverses.

In labeling this reversal as anomalous, people reason that the conditional probability, $P(A|B)$, should be an average of

$$P(A|B \cap C) \text{ and } P(A|B \cap \bar{C}) \ ,$$

and similarly, that $P(A|\bar{B})$ should be an average of

$$P(A|\bar{B} \cap C) \text{ and } P(A|\bar{B} \cap \bar{C}) \ .$$

Although this is true, it is not a simple average but one that is weighted:

$$P(A|B) = P(C|B)P(A|B \cap C) + P(\bar{C}|B)P(A|B \cap \bar{C}) \ ;$$

$$P(A|\bar{B}) = P(C|\bar{B})P(A|\bar{B} \cap C) + P(\bar{C}|\bar{B})P(A|\bar{B} \cap \bar{C}) \ .$$

If B and C are independent events, $P(C|B) = P(C|\bar{B}) = P(C)$ and $P(\bar{C}|B) = P(\bar{C}|\bar{B}) = P(\bar{C})$. Also, under such independence, $P(C)$ and $P(\bar{C})$ $(= 1 - P(C))$ would be the weights for constructing the average, and no reversal would occur. If B and C are not independent, however, a reversal can happen, as it does for our "death penalty" example:

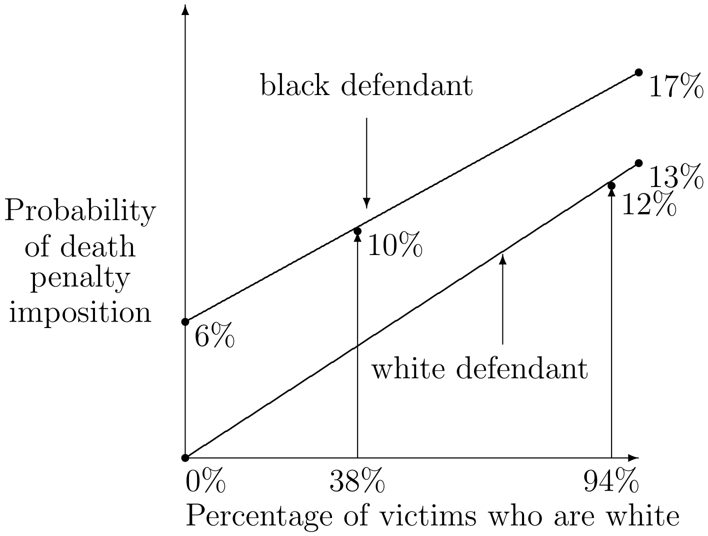
$.10 = P(A|B) = (.38)(.17) + (.62)(.06)$;
$.12 = P(A|\bar{B}) = (.94)(.13) + (.06)(.00)$.

So, instead of the weights of .66 $(= P(C))$ and .34 $(= P(\bar{C}))$, we use .38 $(= P(C|B))$ and .62 $(= P(\bar{C}|B))$; and .94 $(= P(C|\bar{B}))$ and .06 $(= P(\bar{C}|\bar{B}))$.

Figure 1 provides a convenient graphical representation for the reversal paradox in our "death penalty" illustration. This representation generalizes to any $2 \times 2 \times 2$ contingency table. The x-axis is labeled as percentage of victims who are white; the y-axis has a label indicating the probability of death penalty imposition. This probability generally increases along with the percentage of victims that are white. Two separate lines are given in the graph reflecting this increase, one for black defendants and one for white defendants. Note that the line for the black defendant lies wholly above that for the white defendant, implying that irrespective of the percentage of victims that may be white, the imposition of the death penalty has a greater probability for a black defendant compared to a white defendant.

The reversal paradox of having a higher death penalty imposition for whites (of 12%) compared to blacks (of 10%) in the 2×2 contingency table aggregated over the race of the victim, is represented by two vertical lines in the graphs. Because black defendants have 38% of their victims being white, the vertical line from the x-axis value of 38% intersects the black defendant line at 10%; similarly, because white defendants have 94% of their victims being white, the vertical line from the x-axis value of 94% intersects the white defendant line at (a higher value of) 12%. The reversal occurs because there is a much greater percentage of white victims for white defendants than for black defendants. (The two lines in the graph can be constructed readily by noting how the endpoints were obtained of 0% and 6%, and of 13% and 17%. When the percentage of white victims along the x-axis is 0%, that is the same as having a black victim [which immediately generates the graph values of 0% and 6%]; if the percent-

Figure 1: Graphical representation for the Florida death penalty data.



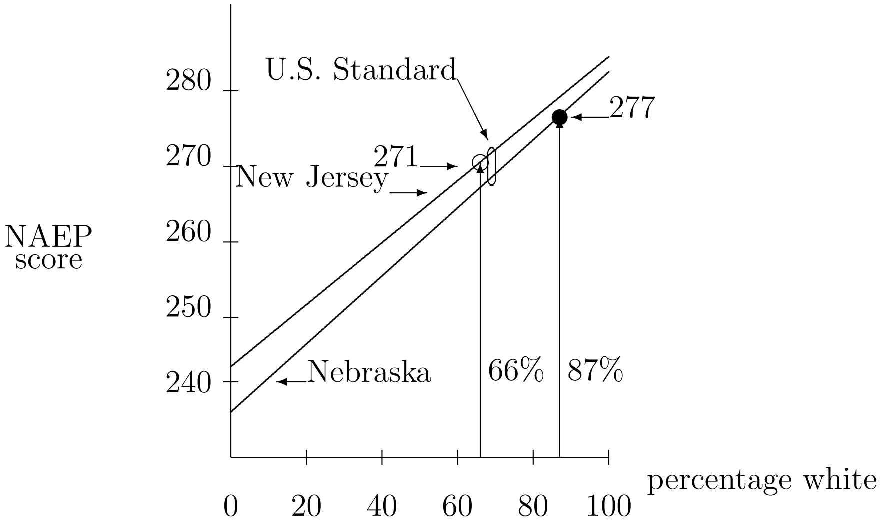
age of white victims is 100%, this is equivalent to the victim being white [and again, immediately provides the other two endpoints of 13% and 17%]).

We conclude with yet another example of Simpson's paradox (taken from Wainer, 2005, pp. 63–67) and a solution called standardization that makes the paradox disappear. Consider the results from the National Assessment of Educational Progress (NAEP) shown in Table 1. The 8th grade students in Nebraska scored 6 points higher in mathematics than their counterparts in New Jersey. White students do better in New Jersey, and so do black students; in fact, all students do better in New Jersey. How is this possible? Again, this is an example of Simpson's paradox. Because a much greater proportion of Nebraska's 8th grade students (87%) are from the higher scoring white population than in New Jersey (66%), their scores contribute more to the total.

Is ranking states on such an overall score sensible? It depends on the question that these scores are being used to answer. If the question is “I want to open a business. In which state will I find a higher proportion of high-scoring math students to hire?”, the unadjusted score is sensible. If, however, the question of interest is: “I want to enroll my children in school. In which state are they likely to do better in math?”, a different answer is required. Irrespective of race, children are more likely to do better in New Jersey. When questions of this latter type are asked more frequently, it makes sense to adjust the total to reflect the correct answer. One way to do this is through the method of standardization, where each state’s score is based upon a common demographic mixture. In this instance, a sensible mixture to use is that of the nation overall. After standardization, the result obtained is the score we would expect each state to have if it had the same demographic mix as the nation. When this is done, New Jersey’s score is not affected much (273 instead of 271), but Nebraska’s score shrinks substantially (271 instead of 277).

Although Simpson’s paradox is subtle, experience has taught us that a graphic depiction often aids understanding. A graphic representation of Simpson’s paradox was provided by Baker and Kramer in 2001. Consider the graphic representation of the results from this table shown in Figure 2. A solid diagonal line shows the average NAEP math score for various proportions of white examinees in Nebraska. At the extreme left, if no whites took the test, the mean score would be that for nonwhites, 236. At the extreme right is what the mean score would be if only whites took the test, 281. The large black dot labeled “277” represents the observed score for the mixture that includes 87% whites. A second solid line above the one for Nebraska shows the same thing for New Jersey; the large open dot

Figure 2: A Baker–Kramer plot of the New Jersey–Nebraska average 8th grade National Assessment of Educational Progress (NAEP) mathematics scores.



labeled “271” denotes the score for a mixture in which 66% of those tested were white.

We see that for any fixed percentage of whites on the horizontal axis, the advantage of New Jersey over Nebraska is the same, two NAEP points. But because Nebraska has a much larger proportion of higher scoring white examinees, its mean score is higher than that of New Jersey. The small vertical box marks the percentage mixture representing the United States as a whole, and hence, encloses the standardized values. The graph makes clear how and why standardization works; it uses the same location on the horizontal axis for all groups being compared.

Simpson’s paradox generally occurs when data are aggregated. If data are collapsed across a subclassification (such as grades, race, or age), the overall difference observed might not represent what is

Table 1: National Assessment of Educational Progress (NAEP) 1992 8th grade mathematics scores.

	State	White	Black	Other Non-White	Standardized
Nebraska	277	281	236	259	271
New Jersey	271	283	242	260	273
		% Population			
Nebraska		87%	5%	8%	
New Jersey		66%	15%	19%	
Nation		69%	16%	15%	

really occurring. Standardization can help correct this, but nothing will prevent the possibility of yet another subclassification, as yet unidentified, from changing things around. We believe, however, that knowing of the possibility helps contain the enthusiasm for what may be overly impulsive first inferences.[3]

Although Simpson's paradox has been known by this name only rather recently (as coined by Colin Blyth in 1972), the phenomenon has been recognized and discussed for well over a hundred years; in fact, it has a complete textbook development in Yule's *An Introduction to the Theory of Statistics*, first published in 1911.

In honor of Yule's early contribution, we sometimes see the title of the Yule–Simpson effect. But most often, Stigler's Law of Eponymy

[3] Fienberg (1988, p. 40) discusses an interesting example of Simpson's paradox as it occurred in a court case involving alleged racial employment discrimination in the receipt of promotions. In this instance, blacks were being "under-promoted" in virtually every pay grade, but because of the differing numbers of blacks and whites in the various grades, blacks appeared to be "over-promoted" in the aggregate. As always, before an overall conclusion is reached based on data that have been aggregated over a variable (such as pay grade), it is always wise to "look under the hood."

is operative (that is, "every scientific discovery is named after the last individual too ungenerous to give due credit to his predecessors."), and Simpson is given sole naming credit for the phenomenon.[4]

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

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 pun-

[4] To get a better sense of the ubiquity of Simpson's paradox in day-to-day reporting of economic statistics, see the article by Cari Tuna, *Wall Street Journal* (December 2, 2009), "When Combined Data Reveal the Flaw of Averages."

ishment. 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).

Although no majority opinion was actually written in the 5 to 4 decision in *Furman v. Georgia*, Justice Brennan writing separately in concurrence noted that

> There are, then, four principles by which we may determine whether a particular punishment is 'cruel and unusual' ... [the] essential predicate [is] that a punishment must not by its severity be degrading to human dignity ... a severe punishment that is obviously inflicted in wholly arbitrary fashion ... a severe punishment that is clearly and totally rejected throughout society ... a severe punishment that is patently unnecessary.

Brennan went on to write that he expected that no state would pass laws obviously violating any one of these principles; and that court decisions involving the Eighth Amendment would use a "cumulative" analysis of the implication of each of the four principles.

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, including 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.

There are many analyses done by Baldus et al. and others on the interrelation between the race of the victim and of the defendant and the imposition of the death penalty. Most do not show an explicit Simpson's paradox such as for the Radelet data of the last section, where a black defendant has a higher probability of receiving the death penalty compared to a white defendant. But universally, the race of the victim plays a crucial part in death penalty imposition – when the victim is white, the probability of receiving the death penalty is substantially higher than for black victims. The relative risks, for example, are all much greater than the value of 2.0 needed to legally assert specific causation.

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 below. 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."

2.1 United States Supreme Court, McCleskey v. Kemp (1987): Syllabus

In 1978, petitioner, a black man, was convicted in a Georgia trial court of armed robbery and murder, arising from the killing of a white police officer during the robbery of a store. Pursuant to Georgia statutes, the jury at the penalty hearing considered the mitigating and aggravating circumstances of petitioner's conduct, and recommended the death penalty on the murder charge. The trial court followed the recommendation, and the Georgia Supreme Court affirmed. After unsuccessfully seeking post-conviction relief in state courts, petitioner sought habeas corpus relief in Federal District Court. His petition included a claim that the Georgia capital sentencing process was administered in a racially discriminatory manner in violation of the Eighth and Fourteenth Amendments. In support of the claim, petitioner proffered a statistical study (the Baldus study) that purports to show a disparity in the imposition of the death sentence in Georgia based on the murder victim's race and, to a lesser extent, the defendant's race. The study is based on over 2,000 murder cases that occurred in Georgia during the 1970's, and involves data relating to the victim's race, the defendant's race, and the various combinations of such persons' races. The study indicates that black defendants who killed white victims have the greatest likelihood of receiving the death penalty. Rejecting petitioner's constitutional claims, the court denied his petition insofar as it was based on the Baldus study, and the Court of Appeals affirmed the District Court's decision on this issue. It assumed the validity of the Baldus study, but found the statistics insufficient to demonstrate unconstitutional discrimination in the Fourteenth Amendment context or to show irrationality, arbitrariness, and capriciousness under Eighth Amendment analysis.

Held:

1. The Baldus study does not establish that the administration of the Georgia capital punishment system violates the Equal Protection Clause.

(a) To prevail under that Clause, petitioner must prove that the decision makers in his case acted with discriminatory purpose. Petitioner offered no evidence specific to his own case that would support an inference that racial considerations played a part in his sentence, and the Baldus study is insufficient to support an inference that any of the decision makers in his case acted with discriminatory purpose. This Court has accepted statistics as proof of intent to discriminate in the context of a State's selection of the jury venire, and in the context of statutory violations under Title VII of the Civil Rights Act of 1964. However, the nature of the capital sentencing decision and the relationship of the statistics to that decision are fundamentally different from the corresponding elements in the venire selection or Title VII cases. Petitioner's statistical proffer must be viewed in the context of his challenge to decisions at the heart of the State's criminal justice system. Because discretion is essential to the criminal justice process, exceptionally clear proof is required before this Court will infer that the discretion has been abused.

(b) There is no merit to petitioner's argument that the Baldus study proves that the State has violated the Equal Protection Clause by adopting the capital punishment statute and allowing it to remain in force despite its allegedly discriminatory application. For this claim to prevail, petitioner would have to prove that the Georgia Legislature enacted or maintained the death penalty statute because of an anticipated racially discriminatory effect. There is no evidence that the legislature either enacted the statute to further a racially discriminatory purpose or maintained the statute because of the racially disproportionate impact suggested by the Baldus study.

2. Petitioner's argument that the Baldus study demonstrates that the Georgia capital sentencing system violates the Eighth Amendment's prohibition of cruel and unusual punishment must be analyzed in the light of this Court's prior decisions under that Amendment. Decisions since Furman v. Georgia, have identified a constitutionally permissible range of discretion in imposing the death penalty. First, there is a required threshold below which the death penalty cannot be imposed, and the State must establish rational criteria that narrow the decision-maker's judgment as to whether the circumstances of a particular defendant's case meet the threshold. Second, States cannot limit the sentencer's consideration of any relevant circumstance that could cause it to decline to impose the death penalty. In this respect, the State cannot channel the sentencer's discretion, but must allow it to consider any relevant information offered by the defendant.

3. The Baldus study does not demonstrate that the Georgia capital sentencing system violates the Eighth Amendment.

(a) Petitioner cannot successfully argue that the sentence in his case is disproportionate to the sentences in other murder cases. On the one hand, he cannot base a constitutional claim on an argument that his case differs from other cases in which defendants did receive the death penalty. The Georgia Supreme Court found that his death sentence was not disproportionate to other death sentences imposed in the State. On the other hand, absent a showing that the Georgia capital punishment system operates in an arbitrary and capricious manner, petitioner cannot prove a constitutional violation by demonstrating that other defendants who may be similarly situated did not receive the death penalty. The opportunities for discretionary leniency under state law do not render the capital sentences imposed arbitrary and capricious. Because petitioner's sentence was imposed under Georgia sentencing procedures that focus discretion "on the particularized nature of the crime and the particularized characteristics of the individual defendant," it may be presumed that his death sentence was not "wantonly and freakishly" imposed, and thus that the sentence is not disproportionate within any recognized meaning under the Eighth Amendment.

(b) There is no merit to the contention that the Baldus study shows that Georgia's capital punishment system is arbitrary and capricious in application. The statistics do not prove that race enters into any capital sentencing decisions or that race was a factor in petitioner's case. The likelihood of racial prejudice allegedly shown by the study does not constitute the constitutional measure of an unacceptable risk of racial prejudice. The inherent lack of predictability of jury decisions does not justify their condemnation. On the contrary, it is the jury's function to make the difficult and uniquely human judgments that defy codification and that build discretion, equity, and flexibility into the legal system.

(c) At most, the Baldus study indicates a discrepancy that appears to correlate with race, but this discrepancy does not constitute a major systemic defect. Any mode for determining guilt or punishment has its weaknesses and the potential for misuse. Despite such imperfections, constitutional guarantees are met when the mode for determining guilt or punishment has been surrounded with safeguards to make it as fair as possible.

4. Petitioner's claim, taken to its logical conclusion, throws into serious question the principles that underlie the entire criminal justice system. His claim easily could be extended to apply to other types of penalties and to claims based on unexplained discrepancies correlating to membership in other minority groups and even to gender. The Constitution does not require that a State eliminate any demonstrable disparity that correlates with a potentially irrelevant factor in order to operate a criminal justice system that includes capital punishment. Petitioner's arguments are best presented to the legislative bodies, not the courts.

We make a number of comments about the majority opinion in *McCleskey v. Kemp* just 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 in Module 1) to infer that a toxic agent was more likely than not the cause of a specific person's disease (and thus open to compensation).[5] 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

[5] 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."

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

The statement that "petitioner cannot successfully argue that the sentence in his case is disproportionate to the sentences in other murder cases" again assigns an impossible personal standard. It will always be impossible to define unequivocally what the "comparables" are that might be used in such comparisons. The operation of confirmation biases would soon overwhelm any attempt to define a set of comparables. Even realtors have huge difficulties in assigning comparable sales to a given property when deciding on an asking or selling price. Usually, realtors just fall back on a simple linear rule of dollars per square foot. But unfortunately, nothing so simple exists in defining comparables in imposing (or not) death sentences in Georgia.

If it can be shown that an enacted (legislative or legal) policy has the effect of denying constitutional rights for an identifiable group of individuals, then that policy should be declared discriminatory and changed. It should never be necessary to show that the enactors of such a policy consciously meant for that effect to occur—the law of unintended consequences again rears its ugly head—or that in one specific case it was operative. When policies must be carried out through human judgment, any number of subjective biases may be present at any given moment, and without any possibility of identifying which ones are at work and which ones are not.

In various places throughout the majority opinion, there appears to be argument by sheer assertion with no other supporting evidence at all. We all need to repeat to ourselves the admonition that just saying so doesn't necessarily make it so. Thus, we have the admission that there appears to be discriminatory effects correlated with race, with the empty assertion that "this discrepancy does not constitute a major systemic weakness" or "despite such imperfection, constitutional guarantees are met." To us, this seems like nonsense, pure and simple.

The final point in the syllabus is that "if the Petitioner's claim is taken to its logical conclusion, questions arise about the principles underlying the entire criminal justice system." Or in Justice Brennan's dissent, the majority opinion is worried about "too much justice." God forbid that other anomalies be identified that correlate with membership in other groups (for example, sex, age, other minorities) that would then have to be dealt with.

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.

We give redactions of the majority opinion and dissent in an appendix (by Justice Brennan) for *McCleskey v. Kemp*. It is a pity that Brennan's dissent did not form the majority opinion as it would have but for Justice Powell's vote that in hindsight he wished he could change. It also would have given greater legitimacy and importance to such landmark statistical studies as done by Baldus, et al. (1983). We will leave readers to peruse the majority and dissenting opinions and arrive at their own identification of outrageous argumentation on either side. In reading the majority and dissenting opinions, it is best to keep in mind the word "opinion." Such opinions include disregarding incontrovertible statistical evidence that something is amiss in the administration of the Georgia death penalty, wherever that may arise from. Although the cause may be ambiguous, there is no doubt that it results from all the various actors in the legal system who make the series of decisions necessary in determining who lives and who dies.

References

[1] Baker, S. G., & Kramer, B. S. (2001). Good for women, good for men, bad for people: Simpson's paradox and the importance of sex-specific analysis in observational studies. *Journal of Women's Health & Gender-Based Medicine*, *10*, 867–872.

[2] Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*, 398–404.

[3] Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*, 364–366.

[4] Fienberg, S. E. (Ed.) (1988). *The evolving role of statistical assessments as evidence in the courts*. New York: Springer-Verlag.

[5] Radelet, M. L. (1981). Racial characterisics and the compositon of the death penalty. *American Sociological Review*, *46*, 918–927.

[6] Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton, N.J.: Princeton University Press.

3 Appendix: United States Supreme Court, McCleskey v. Kemp (1987): Majority Opinion and Dissent

Justice Powell delivered the opinion of the Court.

This case presents the question whether a complex statistical study that indicates a risk that racial considerations enter into capital sentencing determinations proves that petitioner McCleskey's capital sentence is unconstitutional under the Eighth or Fourteenth Amendment.

...

McCleskey next filed a petition for a writ of habeas corpus in the Federal District Court for the Northern District of Georgia. His petition raised 18 claims, one of which was that the Georgia capital sentencing process is administered in a racially discriminatory manner in violation of the Eighth and Fourteenth Amendments to the United States Constitution. In support of his claim, McCleskey proffered a statistical study performed by Professors David C. Baldus, Charles Pulaski, and George Woodworth (the Baldus study) that purports to show a disparity in the imposition of the death sentence in Georgia based on the race of the murder victim and, to a lesser extent, the race of the defendant. The Baldus study is actually two sophisticated statistical studies that examine over 2,000 murder cases that occurred in Georgia during the 1970's. The raw numbers collected by Professor Baldus indicate that defendants charged with killing white persons received the death penalty in 11% of the cases, but defendants charged with killing blacks received the death penalty in only 1% of the cases. The raw numbers also indicate a reverse racial disparity according to the race of the defendant: 4% of the black defendants received the death penalty, as opposed to 7% of the white defendants.

Baldus also divided the cases according to the combination of the race of the defendant and the race of the victim. He found that the death penalty was assessed in 22% of the cases involving black defendants and white victims; 8% of the cases involving white defendants and white victims; 1% of the cases involving black defendants and black victims; and 3% of the cases involving white defendants and black victims. Similarly, Baldus found that prosecutors sought the death penalty in 70% of the cases involving black defendants and white victims; 32% of the cases involving white defendants and white victims; 15% of the cases involving black defendants and black victims; and 19% of the cases involving white defendants and black victims.

Baldus subjected his data to an extensive analysis, taking account of 230 variables that could have explained the disparities on nonracial grounds. One of his models concludes that, even after taking account of 39 nonracial variables, defendants charged with killing white victims were 4.3 times as likely to receive a death sentence as defendants charged with killing blacks. According to this model, black defendants were 1.1 times as likely to receive a death sentence as other defendants. Thus, the Baldus study indicates that black defendants, such as McCleskey, who kill white victims have the greatest likelihood of receiving the death penalty.

The District Court held an extensive evidentiary hearing on McCleskey's petition. ... It concluded that McCleskey's statistics do not demonstrate a prima facie case in support of the contention that the death penalty was imposed upon him because of his race, because of the race of the victim, or because of any Eighth Amendment concern.

As to McCleskey's Fourteenth Amendment claim, the court found that the methodology of the Baldus study was flawed in several respects. Because of these defects, the court held that the Baldus study "fail[ed] to contribute anything of value" to McCleskey's claim. Accordingly, the court denied the petition insofar as it was based upon the Baldus study.[6]

[6] Baldus, among other experts, testified at the evidentiary hearing. The District Court "was impressed with the learning of all of the experts." Nevertheless, the District Court noted that, in many respects, the data were incomplete. In its view, the questionnaires used to obtain the data failed to capture the full degree of the aggravating or mitigating circumstances. The court criticized the researcher's decisions regarding unknown variables. The researchers could not discover whether penalty trials were held in many of the cases, thus undercutting the value of the study's statistics as to prosecutorial decisions. In certain cases, the study lacked information on the race of the victim in cases involving multiple victims, on whether or not the prosecutor offered a plea bargain, and on credibility problems with witnesses. The court concluded that McCleskey had failed to establish by a preponderance of the

The Court of Appeals for the Eleventh Circuit, sitting en banc, carefully reviewed the District Court's decision on McCleskey's claim. It assumed the validity of the study itself, and addressed the merits of McCleskey's Eighth and Fourteenth Amendment claims. That is, the court assumed that the study showed that systematic and substantial disparities existed in the penalties imposed upon homicide defendants in Georgia based on race of the homicide victim, that the disparities existed at a less substantial rate in death sentencing based on race of defendants, and that the factors of race of the victim and defendant were at work in Fulton County.

Even assuming the study's validity, the Court of Appeals found the statis-

evidence that the data were trustworthy.

It is a major premise of a statistical case that the database numerically mirrors reality. If it does not in substantial degree mirror reality, any inferences empirically arrived at are untrustworthy.

The District Court noted other problems with Baldus' methodology. First, the researchers assumed that all of the information available from the questionnaires was available to the juries and prosecutors when the case was tried. The court found this assumption "questionable." Second, the court noted the instability of the various models. Even with the 230-variable model, consideration of 20 further variables caused a significant drop in the statistical significance of race. In the court's view, this undermined the persuasiveness of the model that showed the greatest racial disparity, the 39-variable model. Third, the court found that the high correlation between race and many of the nonracial variables diminished the weight to which the study was entitled.

Finally, the District Court noted the inability of any of the models to predict the outcome of actual cases. As the court explained, statisticians use a measure called an "r-squared" to measure what portion of the variance in the dependent variable (death sentencing rate, in this case) is accounted for by the independent variables of the model. A perfectly predictive model would have an r-squared value of 1.0. A model with no predictive power would have an r-squared value of 0. The r-squared value of Baldus' most complex model, the 230-variable model, was between .46 and .48. Thus, as the court explained, "the 230-variable model does not predict the outcome in half of the cases."

tics insufficient to demonstrate discriminatory intent or unconstitutional discrimination in the Fourteenth Amendment context, [and] insufficient to show irrationality, arbitrariness and capriciousness under any kind of Eighth Amendment analysis.

The court noted:

The very exercise of discretion means that persons exercising discretion may reach different results from exact duplicates. Assuming each result is within the range of discretion, all are correct in the eyes of the law. It would not make sense for the system to require the exercise of discretion in order to be facially constitutional, and at the same time hold a system unconstitutional in application where that discretion achieved different results for what appear to be exact duplicates, absent the state showing the reasons for the difference.

The Baldus approach ... would take the cases with different results on what are contended to be duplicate facts, where the differences could not be otherwise explained, and conclude that the different result was based on race alone. ... This approach ignores the realities. ... There are, in fact, no exact duplicates in capital crimes and capital defendants. The type of research submitted here tends to show which of the directed factors were effective, but is of restricted use in showing what undirected factors control the exercise of constitutionally required discretion.

The court concluded:

Viewed broadly, it would seem that the statistical evidence presented here, assuming its validity, confirms, rather than condemns, the system. ... The marginal disparity based on the race of the victim tends to support the state's contention that the system is working far differently from the one which Furman v. Georgia, condemned. In pre-Furman days, there was no rhyme or reason as to who got the death penalty and who did not. But now, in the vast majority of cases, the reasons for a difference are well documented. That they are not so clear in a small percentage of the cases is no reason to declare the entire system unconstitutional.

The Court of Appeals affirmed the denial by the District Court of McCleskey's petition for a writ of habeas corpus insofar as the petition was based upon the Baldus study, with three judges dissenting as to McCleskey's claims based on the Baldus study. We granted certiorari, and now affirm.

. . .

McCleskey's first claim is that the Georgia capital punishment statute violates the Equal Protection Clause of the Fourteenth Amendment.[7] He argues that race has infected the administration of Georgia's statute in two ways: persons who murder whites are more likely to be sentenced to death than persons who murder blacks, and black murderers are more likely to be sentenced to death than white murderers. As a black defendant who killed a white victim, McCleskey claims that the Baldus study demonstrates that he was discriminated against because of his race and because of the race of his victim. In its broadest form, McCleskey's claim of discrimination extends to every actor in the Georgia capital sentencing process, from the prosecutor who sought the death penalty and the jury that imposed the sentence to the State itself that enacted the capital punishment statute and allows it to remain in effect despite its allegedly discriminatory application. We agree with the Court of Appeals, and every other court that has considered such a challenge, that this claim must fail.

Our analysis begins with the basic principle that a defendant who alleges an equal protection violation has the burden of proving "the existence of purposeful discrimination." A corollary to this principle is that a criminal defendant must prove that the purposeful discrimination "had a discriminatory effect" on him. Thus, to prevail under the Equal Protection Clause, McCleskey must prove that the decision-makers in his case acted with dis-

[7] Although the District Court rejected the findings of the Baldus study as flawed, the Court of Appeals assumed that the study is valid, and reached the constitutional issues. Accordingly, those issues are before us. As did the Court of Appeals, we assume the study is valid statistically, without reviewing the factual findings of the District Court. Our assumption that the Baldus study is statistically valid does not include the assumption that the study shows that racial considerations actually enter into any sentencing decisions in Georgia. Even a sophisticated multiple-regression analysis such as the Baldus study can only demonstrate a risk that the factor of race entered into some capital sentencing decisions, and a necessarily lesser risk that race entered into any particular sentencing decision.

criminatory purpose. He offers no evidence specific to his own case that would support an inference that racial considerations played a part in his sentence. Instead, he relies solely on the Baldus study. McCleskey argues that the Baldus study compels an inference that his sentence rests on purposeful discrimination. McCleskey's claim that these statistics are sufficient proof of discrimination, without regard to the facts of a particular case, would extend to all capital cases in Georgia, at least where the victim was white and the defendant is black.

The Court has accepted statistics as proof of intent to discriminate in certain limited contexts. First, this Court has accepted statistical disparities as proof of an equal protection violation in the selection of the jury venire in a particular district. Although statistical proof normally must present a "stark" pattern to be accepted as the sole proof of discriminatory intent under the Constitution, [b]ecause of the nature of the jury-selection task, ... we have permitted a finding of constitutional violation even when the statistical pattern does not approach [such] extremes.

Second, this Court has accepted statistics in the form of multiple-regression analysis to prove statutory violations under Title VII of the Civil Rights Act of 1964.

But the nature of the capital sentencing decision, and the relationship of the statistics to that decision, are fundamentally different from the corresponding elements in the venire selection or Title VII cases. Most importantly, each particular decision to impose the death penalty is made by a petit jury selected from a properly constituted venire. Each jury is unique in its composition, and the Constitution requires that its decision rest on consideration of innumerable factors that vary according to the characteristics of the individual defendant and the facts of the particular capital offense. Thus, the application of an inference drawn from the general statistics to a specific decision in a trial and sentencing simply is not comparable to the application of an inference drawn from general statistics to a specific venire-selection or Title VII case. In those cases, the statistics relate to fewer entities, and fewer variables are relevant to the challenged decisions.

Another important difference between the cases in which we have accepted statistics as proof of discriminatory intent and this case is that, in the venire-selection and Title VII contexts, the decision-maker has an opportunity to explain the statistical disparity. Here, the State has no practical opportunity to rebut the Baldus study. "[C]ontrolling considerations of ... public policy," dictate that jurors "cannot be called ... to testify to the motives and influences that led to their verdict." Similarly, the policy considerations behind a prosecutor's traditionally "wide discretion" suggest the impropriety of our requiring prosecutors to defend their decisions to seek death penalties, "often years after they were made." Moreover, absent far stronger proof, it is unnecessary to seek such a rebuttal, because a legitimate and unchallenged explanation for the decision is apparent from the record: McCleskey committed an act for which the United States Constitution and Georgia laws permit imposition of the death penalty.

Finally, McCleskey's statistical proffer must be viewed in the context of his challenge. McCleskey challenges decisions at the heart of the State's criminal justice system.

[O]ne of society's most basic tasks is that of protecting the lives of its citizens, and one of the most basic ways in which it achieves the task is through criminal laws against murder.

Implementation of these laws necessarily requires discretionary judgments. Because discretion is essential to the criminal justice process, we would demand exceptionally clear proof before we would infer that the discretion has been abused. The unique nature of the decisions at issue in this case also counsels against adopting such an inference from the disparities indicated by the Baldus study. Accordingly, we hold that the Baldus study is clearly insufficient to support an inference that any of the decision-makers in McCleskey's case acted with discriminatory purpose.

...

McCleskey also suggests that the Baldus study proves that the State as a whole has acted with a discriminatory purpose. He appears to argue that the State has violated the Equal Protection Clause by adopting the capital punishment statute and allowing it to remain in force despite its allegedly discriminatory application. But "[d]iscriminatory purpose" ... implies more than intent as volition or intent as awareness of consequences. It implies that the decision-maker, in this case a state legislature, selected or reaffirmed a particular course of action at least in part “because of,” not merely “in spite of,” its adverse effects upon an identifiable group.

For this claim to prevail, McCleskey would have to prove that the Georgia Legislature enacted or maintained the death penalty statute because of an anticipated racially discriminatory effect. In Gregg v. Georgia, this Court found that the Georgia capital sentencing system could operate in a fair and neutral manner. There was no evidence then, and there is none now, that the Georgia Legislature enacted the capital punishment statute to further a racially discriminatory purpose. Nor has McCleskey demonstrated that the legislature maintains the capital punishment statute because of the racially disproportionate impact suggested by the Baldus study. As legislatures necessarily have wide discretion in the choice of criminal laws and penalties, and as there were legitimate reasons for the Georgia Legislature to adopt and maintain capital punishment, we will not infer a discriminatory purpose on the part of the State of Georgia. Accordingly, we reject McCleskey’s equal protection claims.

. . .

Although our decision in Gregg as to the facial validity of the Georgia capital punishment statute appears to foreclose McCleskey’s disproportionality argument, he further contends that the Georgia capital punishment system is arbitrary and capricious in application, and therefore his sentence is excessive, because racial considerations may influence capital sentencing decisions in Georgia. We now address this claim.

To evaluate McCleskey’s challenge, we must examine exactly what the Baldus study may show. Even Professor Baldus does not contend that his statistics prove that race enters into any capital sentencing decisions, or that race was a factor in McCleskey’s particular case. Statistics, at most, may show only a likelihood that a particular factor entered into some decisions. There is, of course, some risk of racial prejudice influencing a jury’s decision in a criminal case. There are similar risks that other kinds of prejudice will influence other criminal trials. The question “is at what point that risk becomes constitutionally unacceptable,” McCleskey asks us to accept the likelihood allegedly shown by the Baldus study as the constitutional measure of an unacceptable risk of racial prejudice influencing capital sentencing decisions. This we decline to do. Because of the risk that the factor of race may enter the criminal justice process, we have engaged in "unceasing efforts" to eradicate racial prejudice from our criminal justice system. Our efforts have been guided by our recognition that the inestimable privilege of trial by jury . . . is a vital principle, underlying the whole administration of criminal justice system. Thus, it is the jury that is a criminal defendant's fundamental "protection of life and liberty against race or color prejudice." Specifically, a capital sentencing jury representative of a criminal defendant's community assures a 'diffused impartiality,' in the jury's task of "express[ing] the conscience of the community on the ultimate question of life or death.

Individual jurors bring to their deliberations "qualities of human nature and varieties of human experience, the range of which is unknown and perhaps unknowable." The capital sentencing decision requires the individual jurors to focus their collective judgment on the unique characteristics of a particular criminal defendant. It is not surprising that such collective judgments often are difficult to explain. But the inherent lack of predictability of jury decisions does not justify their condemnation. On the contrary, it is the jury's function to make the difficult and uniquely human judgments that defy codification, and that "buil[d] discretion, equity, and flexibility into a legal system."

McCleskey's argument that the Constitution condemns the discretion allowed decision-makers in the Georgia capital sentencing system is antithetical to the fundamental role of discretion in our criminal justice system. Discretion in the criminal justice system offers substantial benefits to the criminal defendant. Not only can a jury decline to impose the death sentence, it can decline to convict or choose to convict of a lesser offense. Whereas decisions against a defendant's interest may be reversed by the trial judge or on appeal, these discretionary exercises of leniency are final and unreviewable. Similarly, the capacity of prosecutorial discretion to provide individualized justice is "only entrenched in American law." As we have noted, a prosecutor can decline to charge, offer a plea bargain, or decline to seek a death sentence in any particular case. Of course, "the power to be lenient [also] is the power to discriminate," but a capital punishment system that did not allow for discretionary acts of leniency "would be totally alien to our notions of criminal justice."

. . .

At most, the Baldus study indicates a discrepancy that appears to correlate with race. Apparent disparities in sentencing are an inevitable part of our criminal justice system. The discrepancy indicated by the Baldus study is "a far cry from the major systemic defects identified in Furman." As this Court has recognized, any mode for determining guilt or punishment "has its weaknesses and the potential for misuse." Specifically, "there can be 'no perfect procedure for deciding in which cases governmental authority should be used to impose death.'" Despite these imperfections, our consistent rule has been that constitutional guarantees are met when "the mode [for determining guilt or punishment] itself has been surrounded with safeguards to make it as fair as possible." Where the discretion that is fundamental to our criminal process is involved, we decline to assume that what is unexplained is invidious. In light of the safeguards designed to minimize racial bias in the process, the fundamental value of jury trial in our criminal justice system, and the benefits that discretion provides to criminal defendants, we hold that the Baldus study does not demonstrate a constitutionally significant risk of racial bias affecting the Georgia capital sentencing process.

. . .

Two additional concerns inform our decision in this case. First, McCleskey's claim, taken to its logical conclusion, throws into serious question the principles that underlie our entire criminal justice system. The Eighth Amendment is not limited in application to capital punishment, but applies to all penalties. Thus, if we accepted McCleskey's claim that racial bias has impermissibly tainted the capital sentencing decision, we could soon be faced with similar claims as to other types of penalty. Moreover, the claim that his sentence rests on the irrelevant factor of race easily could be extended to apply to claims based on unexplained discrepancies that correlate to membership in other minority groups, and even to gender. Similarly, since McCleskey's claim relates to the race of his victim, other claims could apply with equally logical force to statistical disparities that correlate with the race or sex of other actors in the criminal justice system, such as defense attorneys or judges. Also, there is no logical reason that such a claim need be limited to racial or sexual bias. If arbitrary and capricious punishment is the touchstone under the

Eighth Amendment, such a claim could—at least in theory—be based upon any arbitrary variable, such as the defendant's facial characteristics, or the physical attractiveness of the defendant or the victim, that some statistical study indicates may be influential in jury decision-making. As these examples illustrate, there is no limiting principle to the type of challenge brought by McCleskey. The Constitution does not require that a State eliminate any demonstrable disparity that correlates with a potentially irrelevant factor in order to operate a criminal justice system that includes capital punishment. As we have stated specifically in the context of capital punishment, the Constitution does not "plac[e] totally unrealistic conditions on its use."

Second, McCleskey's arguments are best presented to the legislative bodies. It is not the responsibility—or indeed even the right—of this Court to determine the appropriate punishment for particular crimes. It is the legislatures, the elected representatives of the people, that are "constituted to respond to the will and consequently the moral values of the people." Legislatures also are better qualified to weigh and evaluate the results of statistical studies in terms of their own local conditions and with a flexibility of approach that is not available to the courts,

Capital punishment is now the law in more than two-thirds of our States. It is the ultimate duty of courts to determine on a case-by-case basis whether these laws are applied consistently with the Constitution. Despite McCleskey's wide-ranging arguments that basically challenge the validity of capital punishment in our multiracial society, the only question before us is whether, in his case, the law of Georgia was properly applied. We agree with the District Court and the Court of Appeals for the Eleventh Circuit that this was carefully and correctly done in this case.

Accordingly, we affirm the judgment of the Court of Appeals for the Eleventh Circuit.

It is so ordered.

———

Justice Brennan, Dissenting Opinion

. . .

At some point in this case, Warren McCleskey doubtless asked his lawyer whether a jury was likely to sentence him to die. A candid reply to this question would have been disturbing. First, counsel would have to tell McCleskey that few of the details of the crime or of McCleskey's past criminal conduct were more important than the fact that his victim was white. Furthermore, counsel would feel bound to tell McCleskey that defendants charged with killing white victims in Georgia are 4.3 times as likely to be sentenced to death as defendants charged with killing blacks. In addition, frankness would compel the disclosure that it was more likely than not that the race of McCleskey's victim would determine whether he received a death sentence: 6 of every 11 defendants convicted of killing a white person would not have received the death penalty if their victims had been black, while, among defendants with aggravating and mitigating factors comparable to McCleskey's, 20 of every 34 would not have been sentenced to die if their victims had been black. Finally, the assessment would not be complete without the information that cases involving black defendants and white victims are more likely to result in a death sentence than cases featuring any other racial combination of defendant and victim. The story could be told in a variety of ways, but McCleskey could not fail to grasp its essential narrative line: there was a significant chance that race would play a prominent role in determining if he lived or died.

The Court today holds that Warren McCleskey's sentence was constitutionally imposed. It finds no fault in a system in which lawyers must tell their clients that race casts a large shadow on the capital sentencing process. The Court arrives at this conclusion by stating that the Baldus study cannot "prove that race enters into any capital sentencing decisions or that race was a factor in McCleskey's particular case." Since, according to Professor Baldus, we cannot say "to a moral certainty" that race influenced a decision, we can identify only "a likelihood that a particular factor entered into some decisions," and "a discrepancy that appears to correlate with race." This "likelihood" and "discrepancy," holds the Court, is insufficient to establish a constitutional violation. The Court reaches this conclusion by placing four factors on the scales opposite McCleskey's evidence: the desire to encourage sentencing discretion, the existence of "statutory safeguards" in the Georgia scheme, the fear of encouraging widespread challenges to other sentencing decisions, and the limits of the judicial role. The Court's evaluation of the significance of petitioner's evidence is fundamentally at odds with our consistent concern for rationality in capital sentencing, and the considerations that the majority invokes to discount that evidence cannot justify ignoring its force.

. . .

It is important to emphasize at the outset that the Court's observation that McCleskey cannot prove the influence of race on any particular sentencing decision is irrelevant in evaluating his Eighth Amendment claim. Since Furman v. Georgia, the Court has been concerned with the risk of the imposition of an arbitrary sentence, rather than the proven fact of one. Furman held that the death penalty may not be imposed under sentencing procedures that create a substantial risk that the punishment will be inflicted in an arbitrary and capricious manner.

As Justice O'Connor observed in Caldwell v. Mississippi, a death sentence must be struck down when the circumstances under which it has been imposed creat[e] an unacceptable risk that "the death penalty [may have been] meted out arbitrarily or capriciously," or through "whim or mistake." This emphasis on risk acknowledges the difficulty of divining the jury's motivation in an individual case. In addition, it reflects the fact that concern for arbitrariness focuses on the rationality of the system as a whole, and that a system that features a significant probability that sentencing decisions are influenced by impermissible considerations cannot be regarded as rational. As we said in Gregg v. Georgia, "the petitioner looks to the sentencing system as a whole (as the Court did in Furman and we do today)": a constitutional violation is established if a plaintiff demonstrates a "pattern of arbitrary and capricious sentencing."

As a result, our inquiry under the Eighth Amendment has not been directed to the validity of the individual sentences before us. In Godfrey, for instance, the Court struck down the petitioner's sentence because the vagueness of the statutory definition of heinous crimes created a risk that prejudice or other impermissible influences might have infected the sentencing decision. In vacating the sentence, we did not ask whether it was likely that Godfrey's own sentence reflected the operation of irrational considerations. Nor did we demand a demonstration that such considerations had actually entered into other sentencing decisions involving heinous crimes. Similarly, in Roberts v. Louisiana, and Woodson v. North Carolina, we struck down death sentences in part because mandatory imposition of the death penalty created the risk that a jury might rely on arbitrary considerations in deciding which persons should be convicted of capital crimes. Such a risk would arise, we said, because of the likelihood that jurors, reluctant to impose capital punishment on a particular defendant, would refuse to return a conviction, so that the effect of mandatory sentencing would be to recreate the unbounded sentencing discretion condemned in Furman. We did not ask whether the death sentences in the cases before us could have reflected the jury's rational consideration and rejection of mitigating factors. Nor did we require proof that juries had actually acted irrationally in other cases.

Defendants challenging their death sentences thus never have had to prove that impermissible considerations have actually infected sentencing decisions. We have required instead that they establish that the system under which they were sentenced posed a significant risk of such an occurrence. McCleskey's claim does differ, however, in one respect from these earlier cases: it is the first to base a challenge not on speculation about how a system might operate, but on empirical documentation of how it does operate.

The Court assumes the statistical validity of the Baldus study, and acknowledges that McCleskey has demonstrated a risk that racial prejudice plays a role in capital sentencing in Georgia. Nonetheless, it finds the probability of prejudice insufficient to create constitutional concern. Close analysis of the Baldus study, however, in light of both statistical principles and human experience, reveals that the risk that race influenced McCleskey's sentence is intolerable by any imaginable standard.

. . .

The Baldus study indicates that, after taking into account some 230 nonracial factors that might legitimately influence a sentencer, the jury more likely than not would have spared McCleskey's life had his victim been black. The study distinguishes between those cases in which (1) the jury exercises virtually no discretion because the strength or weakness of aggravating factors usually suggests that only one outcome is appropriate; and (2) cases reflecting an "intermediate" level of aggravation, in which the jury has considerable discretion in choosing a sentence. McCleskey's case falls into the intermediate range. In such cases, death is imposed in 34% of white-victim crimes and 14% of black-victim crimes, a difference of 139% in the rate of imposition of the death penalty. In other words, just under 59%—almost 6 in 10—defendants comparable to McCleskey would not have received the death penalty if their victims had been black.

Furthermore, even examination of the sentencing system as a whole, factoring in those cases in which the jury exercises little discretion, indicates the influence of race on capital sentencing. For the Georgia system as a whole, race accounts for a six percentage point difference in the rate at which capital punishment is imposed. Since death is imposed in 11% of all white-victim cases, the rate in comparably aggravated black-victim cases is 5%. The rate of capital sentencing in a white-victim case is thus 120% greater than the rate in a black-victim case. Put another way, over half—55%—of defendants in white-victim crimes in Georgia would not have been sentenced to die if their victims had been black. Of the more than 200 variables potentially relevant to a sentencing decision, race of the victim is a powerful explanation for variation in death sentence rates—as powerful as nonracial aggravating factors such as a prior murder conviction or acting as the principal planner of the homicide.

These adjusted figures are only the most conservative indication of the risk that race will influence the death sentences of defendants in Georgia. Data unadjusted for the mitigating or aggravating effect of other factors show an even more pronounced disparity by race. The capital sentencing rate for all white-victim cases was almost 11 times greater than the rate for black-victim cases. Furthermore, blacks who kill whites are sentenced to death at nearly 22 times the rate of blacks who kill blacks, and more than 7 times the rate of whites who kill blacks. In addition, prosecutors seek the death penalty for 70% of black defendants with white victims, but for only 15% of black defendants with black victims, and only 19% of white defendants with black victims. Since our decision upholding the Georgia capital sentencing system in Gregg, the State has executed seven persons. All of the seven were convicted of killing whites, and six of the seven executed were black. Such execution figures are especially striking in light of the fact that, during the period encompassed by the Baldus study, only 9.2% of Georgia homicides involved black defendants and white victims, while 60.7% involved black victims.

McCleskey's statistics have particular force because most of them are the product of sophisticated multiple-regression analysis. Such analysis is designed precisely to identify patterns in the aggregate, even though we may not be able to reconstitute with certainty any individual decision that goes to make up that pattern. Multiple-regression analysis is particularly well suited to identify the influence of impermissible considerations in sentencing, since it is able to control for permissible factors that may explain an apparent arbitrary pattern. While the decision-making process of a body such as a jury may be complex, the Baldus study provides a massive compilation of the details that are most relevant to that decision. As we held in the context of Title VII of the Civil Rights Act of 1964 last Term in Bazemore v. Friday, a multiple-regression analysis need not include every conceivable variable to establish a party's case, as long as it includes those variables that account for the major factors that are likely to influence decisions. In this case, Professor Baldus in fact conducted additional regression analyses in response to criticisms and suggestions by the District Court, all of which confirmed, and some of which even strengthened, the study's original conclusions.

The statistical evidence in this case thus relentlessly documents the risk that McCleskey's sentence was influenced by racial considerations. This evidence shows that there is a better than even chance in Georgia that race will influence the decision to impose the death penalty: a majority of defendants in white-victim crimes would not have been sentenced to die if their victims had been black. In determining whether this risk is acceptable, our judgment must be shaped by the awareness that [t]he risk of racial prejudice infecting a capital sentencing proceeding is especially serious in light of the complete finality of the death sentence, and that [i]t is of vital importance to the defendant and to the community that any decision to impose the death sentence be, and appear to be, based on reason rather than caprice or emotion. In determining the guilt of a defendant, a State must prove its case beyond a reasonable doubt. That is, we refuse to convict if the chance of error is simply less likely than not. Surely, we should not be willing to take a person's life if the chance that his death sentence was irrationally imposed is more likely than not. In light of the gravity of the interest at stake, petitioner's statistics, on their face, are a powerful demonstration of the type of risk that our Eighth Amendment jurisprudence has consistently condemned.

. . .

Evaluation of McCleskey's evidence cannot rest solely on the numbers themselves. We must also ask whether the conclusion suggested by those numbers is consonant with our understanding of history and human experience. Georgia's legacy of a race-conscious criminal justice system, as well as this Court's own recognition of the persistent danger that racial attitudes may affect criminal proceedings, indicates that McCleskey's claim is not a fanciful product of mere statistical artifice.

For many years, Georgia operated openly and formally precisely the type of dual system the evidence shows is still effectively in place. The criminal law expressly differentiated between crimes committed by and against blacks and whites, distinctions whose lineage traced back to the time of slavery. During the colonial period, black slaves who killed whites in Georgia, regardless of whether in self-defense or in defense of another, were automatically executed.

By the time of the Civil War, a dual system of crime and punishment was well established in Georgia. The state criminal code contained separate sections for "Slaves and Free Persons of Color," and for all other persons. The code provided, for instance, for an automatic death sentence for murder committed by blacks, but declared that anyone else convicted of murder might receive life imprisonment if the conviction were founded solely on circumstantial testimony or simply if the jury so recommended. The code established that the rape of a free white female by a black "shall be punishable by death. However, rape by anyone else of a free white female was punishable by a prison term not less than 2 nor more than 20 years. The rape of blacks was punishable "by fine and imprisonment, at the discretion of the court." A black convicted of assaulting a free white person with intent to murder could be put to death at the discretion of the court, but the same offense committed against a black, slave or free, was classified as a "minor" offense whose punishment lay in the discretion of the court, as long as such punishment did not "extend to life, limb, or health." Assault with intent to murder by a white person was punishable by a prison term of from 2 to 10 years. While sufficient provocation could reduce a charge of murder to manslaughter, the code provided that [o]bedience and submission being the duty of a slave, much greater provocation is necessary to reduce a homicide of a white person by him to voluntary manslaughter, than is prescribed for white persons.

In more recent times, some 40 years ago, Gunnar Myrdal's epochal study of American race relations produced findings mirroring McCleskey's evidence:

As long as only Negroes are concerned and no whites are disturbed, great leniency will be shown in most cases. . . . The sentences for even major crimes are ordinarily reduced when the victim is another Negro.

. . .

For offenses which involve any actual or potential danger to whites, however, Negroes are punished more severely than whites.

. . .

On the other hand, it is quite common for a white criminal to be set free if his crime was against a Negro.

. . .

This Court has invalidated portions of the Georgia capital sentencing system three times over the past 15 years. The specter of race discrimination was acknowledged by the Court in striking down the Georgia death penalty statute in Furman. Justice Douglas cited studies suggesting imposition of the death penalty in racially discriminatory fashion, and found the standard-less statutes before the Court "pregnant with discrimination." Justice Marshall pointed to statistics indicating that Negroes [have been] executed far more often than whites in proportion to their percentage of the population. Studies indicate that, while the higher rate of execution among Negroes is partially due to a higher rate of crime, there is evidence of racial discrimination. Although Justice Stewart declined to conclude that racial discrimination had been plainly proved, he stated that [m]y concurring Brothers have demonstrated that, if any basis can be discerned for the selection of these few to be sentenced to die, it is the constitutionally impermissible basis of race. In dissent, Chief Justice Burger acknowledged that statistics suggest, at least as a historical matter, that Negroes have been sentenced to death with greater frequency than whites in several States, particularly for the crime of interracial rape. Finally, also in dissent, Justice Powell intimated that an Equal Protection Clause argument would be available for a black who could demonstrate that members of his race were being singled out for more severe punishment than others charged with the same offense. He noted that, although the Eighth Circuit had rejected a claim of discrimination in Maxwell v. Bishop, vacated and remanded on other grounds, the statistical evidence in that case tend[ed] to show a pronounced disproportion in the number of Negroes receiving death sentences for rape in parts of Arkansas and elsewhere in the South. It is clear that the Court regarded the opportunity for the operation of racial prejudice a particularly troublesome aspect of the unbounded discretion afforded by the Georgia sentencing scheme. Five years later, the Court struck down the imposition of the death penalty in Georgia for the crime of rape. Although the Court did not explicitly mention race, the decision had to have been informed by the specific observations on rape by both the Chief Justice and Justice Powell in Furman. Furthermore, evidence submitted to the Court indicated that black men who committed rape, particularly of white women, were considerably more likely to be sentenced to death than white rapists. For instance, by 1977, Georgia had executed 62 men for rape since the Federal Government began compiling statistics in 1930. Of these men, 58 were black and 4 were white. Three years later, the Court in Godfrey found one of the State's statutory aggravating factors unconstitutionally vague, since it resulted in "standard-less and unchanneled imposition of death sentences in the uncontrolled discretion of a basically uninstructed jury. . . . " Justice Marshall, concurring in the judgment, noted that [t]he disgraceful distorting effects of racial discrimination and poverty continue to be painfully visible in the imposition of death sentences.

This historical review of Georgia criminal law is not intended as a bill of indictment calling the State to account for past transgressions. Citation of past practices does not justify the automatic condemnation of current ones. But it would be unrealistic to ignore the influence of history in assessing the plausible implications of McCleskey's evidence. [A]mericans share a historical experience that has resulted in individuals within the culture ubiquitously attaching a significance to race that is irrational and often outside their awareness. As we said in Rose v. Mitchell: [W]e . . . cannot deny that, 114 years after the close of the War Between the States and nearly 100 years after Strauder, racial and other forms of discrimination still remain a fact of life, in the administration of justice as in our society as a whole. Perhaps today that discrimination takes a form more subtle than before. But it is not less real or pernicious.

The ongoing influence of history is acknowledged, as the majority observes, by our "unceasing efforts to eradicate racial prejudice from our criminal justice system." These efforts, however, signify not the elimination of the problem, but its persistence. Our cases reflect a realization of the myriad of opportunities for racial considerations to influence criminal proceedings: in the exercise of peremptory challenges, in the selection of the grand jury, in the selection of the petit jury, in the exercise of prosecutorial discretion, in the conduct of argument, and in the conscious or unconscious bias of jurors.

The discretion afforded prosecutors and jurors in the Georgia capital sentencing system creates such opportunities. No guidelines govern prosecutorial decisions to seek the death penalty, and Georgia provides juries with no list of aggravating and mitigating factors, nor any standard for balancing them against one another. Once a jury identifies one aggravating factor, it has complete discretion in choosing life or death, and need not articulate its basis for selecting life imprisonment. The Georgia sentencing system therefore provides considerable opportunity for racial considerations, however subtle and unconscious, to influence charging and sentencing decisions.

History and its continuing legacy thus buttress the probative force of McCleskey's statistics. Formal dual criminal laws may no longer be in effect, and intentional discrimination may no longer be prominent. Nonetheless, as we acknowledged in Turner, "subtle, less consciously held racial attitudes" continue to be of concern, and the Georgia system gives such attitudes considerable room to operate. The conclusions drawn from McCleskey's statistical evidence are therefore consistent with the lessons of social experience.

The majority thus misreads our Eighth Amendment jurisprudence in concluding that McCleskey has not demonstrated a degree of risk sufficient to raise constitutional concern. The determination of the significance of his evidence is at its core an exercise in human moral judgment, not a mechanical statistical analysis. It must first and foremost be informed by awareness of the fact that death is irrevocable, and that, as a result, the qualitative difference of death from all other punishments requires a greater degree of scrutiny of the capital sentencing determination. 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. It is true that every nuance of decision cannot be statistically captured, nor can any individual judgment be plumbed with absolute certainty. Yet the fact that we must always act without the illumination of complete knowledge cannot induce paralysis when we confront what is literally an issue of life and death. Sentencing data, history, and experience all counsel that Georgia has provided insufficient assurance of the heightened rationality we have required in order to take a human life.

. . .

The Court cites four reasons for shrinking from the implications of McCleskey's evidence: the desirability of discretion for actors in the criminal justice system, the existence of statutory safeguards against abuse of that discretion, the potential consequences for broader challenges to criminal sentencing, and an understanding of the contours of the judicial role. While these concerns underscore the need for sober deliberation, they do not justify rejecting evidence as convincing as McCleskey has presented.

The Court maintains that petitioner's claim "is antithetical to the fundamental role of discretion in our criminal justice system." It states that "[w]here the discretion that is fundamental to our criminal process is involved, we decline to assume that what is unexplained is invidious."

Reliance on race in imposing capital punishment, however, is antithetical to the very rationale for granting sentencing discretion. Discretion is a means, not an end. It is bestowed in order to permit the sentencer to "trea[t] each defendant in a capital case with that degree of respect due the uniqueness of the individual." The decision to impose the punishment of death must be based on a "particularized consideration of relevant aspects of the character and record of each convicted defendant." Failure to conduct such an individualized moral inquiry treats all persons convicted of a designated offense not as unique individual human beings, but as members of a faceless, undifferentiated mass to be subjected to the blind infliction of the penalty of death.

Considering the race of a defendant or victim in deciding if the death penalty should be imposed is completely at odds with this concern that an individual be evaluated as a unique human being. Decisions influenced by race rest in part on a categorical assessment of the worth of human beings according to color, insensitive to whatever qualities the individuals in question may possess. Enhanced willingness to impose the death sentence on black defendants, or diminished willingness to render such a sentence when blacks are victims, reflects a devaluation of the lives of black persons. When confronted with evidence that race more likely than not plays such a role in a capital sentencing system, it is plainly insufficient to say that the importance of discretion demands that the risk be higher before we will act—for, in such a case, the very end that discretion is designed to serve is being undermined.

Our desire for individualized moral judgments may lead us to accept some inconsistencies in sentencing outcomes. Since such decisions are not reducible to mathematical formulae, we are willing to assume that a certain degree of variation reflects the fact that no two defendants are completely alike. There is thus a presumption that actors in the criminal justice system exercise their discretion in responsible fashion, and we do not automatically infer that sentencing patterns that do not comport with ideal rationality are suspect.

As we made clear in Batson v. Kentucky, however, that presumption is rebuttable. Batson dealt with another arena in which considerable discretion traditionally has been afforded, the exercise of peremptory challenges. Those challenges are normally exercised without any indication whatsoever of the grounds for doing so. The rationale for this deference has been a belief that the unique characteristics of particular prospective jurors may raise concern on the part of the prosecution or defense, despite the fact that counsel may not be able to articulate that concern in a manner sufficient to support exclusion for cause. As with sentencing, therefore, peremptory challenges are justified as an occasion for particularized determinations related to specific individuals, and, as with sentencing, we presume that such challenges normally are not made on the basis of a factor such as race. As we said in Batson, however, such features do not justify imposing a "crippling burden of proof," in order to rebut that presumption. The Court in this case apparently seeks to do just that. On the basis of the need for individualized decisions, it rejects evidence, drawn from the most sophisticated capital sentencing analysis ever performed, that reveals that race more likely than not infects capital sentencing decisions. The Court's position converts a rebuttable presumption into a virtually conclusive one.

The Court also declines to find McCleskey's evidence sufficient in view of "the safeguards designed to minimize racial bias in the [capital sentencing] process." Gregg v. Georgia, upheld the Georgia capital sentencing statute against a facial challenge which Justice White described in his concurring opinion as based on "simply an assertion of lack of faith" that the system could operate in a fair manner (opinion concurring in judgment). Justice White observed that the claim that prosecutors might act in an arbitrary fashion was "unsupported by any facts," and that prosecutors must be assumed to exercise their charging duties properly "[a]bsent facts to the contrary." It is clear that Gregg bestowed no permanent approval on the Georgia system. It simply held that the State's statutory safeguards were assumed sufficient to channel discretion without evidence otherwise.

It has now been over 13 years since Georgia adopted the provisions upheld in Gregg. Professor Baldus and his colleagues have compiled data on almost 2,500 homicides committed during the period 1973-1979. They have taken into account the influence of 230 nonracial variables, using a multitude of data from the State itself, and have produced striking evidence that the odds of being sentenced to death are significantly greater than average if a defendant is black or his or her victim is white. The challenge to the Georgia system is not speculative or theoretical; it is empirical. As a result, the Court cannot rely on the statutory safeguards in discounting McCleskey's evidence, for it is the very effectiveness of those safeguards that such evidence calls into question. While we may hope that a model of procedural fairness will curb the influence of race on sentencing, "we cannot simply assume that the model works as intended; we must critique its performance in terms of its results."

The Court next states that its unwillingness to regard petitioner's evidence as sufficient is based in part on the fear that recognition of McCleskey's claim would open the door to widespread challenges to all aspects of criminal sentencing. Taken on its face, such a statement seems to suggest a fear of too much justice. Yet surely the majority would acknowledge that, if striking evidence indicated that other minority groups, or women, or even persons with blond hair, were disproportionately sentenced to death, such a state of affairs would be repugnant to deeply rooted conceptions of fairness. The prospect that there may be more widespread abuse than McCleskey documents may be dismaying, but it does not justify complete abdication of our judicial role. The Constitution was framed fundamentally as a bulwark against governmental power, and preventing the arbitrary administration of punishment is a basic ideal of any society that purports to be governed by the rule of law.

In fairness, the Court's fear that McCleskey's claim is an invitation to descend a slippery slope also rests on the realization that any humanly imposed system of penalties will exhibit some imperfection. Yet to reject McCleskey's powerful evidence on this basis is to ignore both the qualitatively different character of the death penalty and the particular repugnance of racial discrimination, considerations which may properly be taken into account in determining whether various punishments are "cruel and unusual." Furthermore, it fails to take account of the unprecedented refinement and strength of the Baldus study.

It hardly needs reiteration that this Court has consistently acknowledged the uniqueness of the punishment of death. Death, in its finality, differs more from life imprisonment than a 100-year prison term differs from one of only a year or two. Because of that qualitative difference, there is a corresponding difference in the need for reliability in the determination that death is the appropriate punishment. Furthermore, the relative interests of the state and the defendant differ dramatically in the death penalty context. The marginal benefits accruing to the state from obtaining the death penalty, rather than life imprisonment, are considerably less than the marginal difference to the defendant between death and life in prison. Such a disparity is an additional reason for tolerating scant arbitrariness in capital sentencing. Even those who believe that society can impose the death penalty in a manner sufficiently rational to justify its continuation must acknowledge that the level of rationality that is considered satisfactory must be uniquely high. As a result, the degree of arbitrariness that may be adequate to render the death penalty "cruel and unusual" punishment may not be adequate to invalidate lesser penalties. What these relative degrees of arbitrariness might be in other cases need not concern us here; the point is that the majority's fear of wholesale invalidation of criminal sentences is unfounded.

The Court also maintains that accepting McCleskey's claim would pose a threat to all sentencing because of the prospect that a correlation might be demonstrated between sentencing outcomes and other personal characteristics. Again, such a view is indifferent to the considerations that enter into a determination whether punishment is "cruel and unusual." Race is a consideration whose influence is expressly constitutionally proscribed. We have expressed a moral commitment, as embodied in our fundamental law, that this specific characteristic should not be the basis for allotting burdens and benefits. Three constitutional amendments, and numerous statutes, have been prompted specifically by the desire to address the effects of racism.

Over the years, this Court has consistently repudiated "[d]istinctions between citizens solely because of their ancestry" as being "odious to a free people whose institutions are founded upon the doctrine of equality."

Furthermore, we have explicitly acknowledged the illegitimacy of race as a consideration in capital sentencing. That a decision to impose the death penalty could be influenced by race is thus a particularly repugnant prospect, and evidence that race may play even a modest role in levying a death sentence should be enough to characterize that sentence as "cruel and unusual."

Certainly, a factor that we would regard as morally irrelevant, such as hair color, at least theoretically could be associated with sentencing results to such an extent that we would regard as arbitrary a system in which that factor played a significant role. As I have said above, however, the evaluation of evidence suggesting such a correlation must be informed not merely by statistics, but by history and experience. One could hardly contend that this Nation has, on the basis of hair color, inflicted upon persons deprivation comparable to that imposed on the basis of race. Recognition of this fact would necessarily influence the evaluation of data suggesting the influence of hair color on sentencing, and would require evidence of statistical correlation even more powerful than that presented by the Baldus study.

Furthermore, the Court's fear of the expansive ramifications of a holding for McCleskey in this case is unfounded, because it fails to recognize the uniquely sophisticated nature of the Baldus study. McCleskey presents evidence that is far and away the most refined data ever assembled on any system of punishment, data not readily replicated through casual effort. Moreover, that evidence depicts not merely arguable tendencies, but striking correlations, all the more powerful because nonracial explanations have been eliminated. Acceptance of petitioner's evidence would therefore establish a remarkably stringent standard of statistical evidence unlikely to be satisfied with any frequency.

The Court's projection of apocalyptic consequences for criminal sentencing is thus greatly exaggerated. The Court can indulge in such speculation only by ignoring its own jurisprudence demanding the highest scrutiny on issues of death and race. As a result, it fails to do justice to a claim in which both those elements are intertwined—an occasion calling for the most sensitive inquiry a court can conduct. Despite its acceptance of the validity of Warren McCleskey's evidence, the Court is willing to let his death sentence stand because it fears that we cannot successfully define a different standard for lesser punishments. This fear is baseless.

Finally, the Court justifies its rejection of McCleskey's claim by cautioning against usurpation of the legislatures' role in devising and monitoring criminal punishment. The Court is, of course, correct to emphasize the gravity of constitutional intervention, and the importance that it be sparingly employed. The fact that "[c]apital punishment is now the law in more than two thirds of our States," however, does not diminish the fact that capital punishment is the most awesome act that a State can perform. The judiciary's role in this society counts for little if the use of governmental power to extinguish life does not elicit close scrutiny. It is true that society has a legitimate interest in punishment. Yet, as Alexander Bickel wrote:

It is a premise we deduce not merely from the fact of a written constitution but from the history of the race, and ultimately as a moral judgment of the good society, that government should serve not only what we conceive from time to time to be our immediate material needs, but also certain enduring values. This in part is what is meant by government under law.

Our commitment to these values requires fidelity to them even when there is temptation to ignore them. Such temptation is especially apt to arise in criminal matters, for those granted constitutional protection in this context are those whom society finds most menacing and opprobrious. Even less sympathetic are those we consider for the sentence of death, for execution "is a way of saying, 'You are not fit for this world, take your chance elsewhere.' " For these reasons, [t]he methods we employ in the enforcement of our criminal law have aptly been called the measures by which the quality of our civilization may be judged. 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. The Court thus fulfills, rather than disrupts, the scheme of separation of powers by closely scrutinizing the imposition of the death penalty, for no decision of a society is more deserving of "sober second thought."

...

At the time our Constitution was framed 200 years ago this year, blacks had for more than a century before been regarded as beings of an inferior order, and altogether unfit to associate with the white race, either in social or political relations; and so far inferior that they had no rights which the white man was bound to respect. Only 130 years ago, this Court relied on these observations to deny American citizenship to blacks. A mere three generations ago, this Court sanctioned racial segregation, stating that "[i]f one race be inferior to the other socially, the Constitution of the United States cannot put them upon the same plane."

In more recent times, we have sought to free ourselves from the burden of this history. Yet it has been scarcely a generation since this Court's first decision striking down racial segregation, and barely two decades since the legislative prohibition of racial discrimination in major domains of national life. These have been honorable steps, but we cannot pretend that, in three decades, we have completely escaped the grip of a historical legacy spanning centuries. Warren McCleskey's evidence confronts us with the subtle and persistent influence of the past. His message is a disturbing one to a society that has formally repudiated racism, and a frustrating one to a Nation accustomed to regarding its destiny as the product of its own will. Nonetheless, we ignore him at our peril, for we remain imprisoned by the past as long as we deny its influence in the present.

It is tempting to pretend that minorities on death row share a fate in no way connected to our own, that our treatment of them sounds no echoes beyond the chambers in which they die. Such an illusion is ultimately corrosive, for the reverberations of injustice are not so easily confined. "The destinies of the two races in this country are indissolubly linked together," and the way in which we choose those who will die reveals the depth of moral commitment among the living.

The Court's decision today will not change what attorneys in Georgia tell other Warren McCleskeys about their chances of execution. Nothing will soften the harsh message they must convey, nor alter the prospect that race undoubtedly will continue to be a topic of discussion. McCleskey's evidence will not have obtained judicial acceptance, but that will not affect what is said on death row. However many criticisms of today's decision may be rendered, these painful conversations will serve as the most eloquent dissents of all.

Module 4: Probabilistic Reasoning and Diagnostic Testing

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

Abstract: Two main questions are discussed that relate to diagnostic testing. First, when does prediction using simple base rate information outperform prediction with an actual diagnostic test?; and second, how should the performance of a diagnostic test be evaluated in general? Module 2 on the (un)reliability of clinical and actuarial prediction introduced the Meehl and Rosen (1955) notion of “clinical efficiency,” which is a phrase applied to a diagnostic test when it outperforms base rate predictions. In the first section to follow, three equivalent conditions are given for when “clinical efficiency” holds; these conditions are attributed to Meehl and Rosen (1955), Dawes (1962), and Bokhari and Hubert (2015). The second main section of this module introduces the Receiver Operating Characteristic (ROC) curve, and contrasts the use of a common measure of test performance, the “area under the curve” (AUC), with possibly more appropriate performance measures that take base rates into consideration. A final section of the module discusses several issues that must be faced when implementing screening programs: the evidence for the (in)effectiveness of cancer screening for breast (through mammography) and prostate (through the prostate-specific antigen (PSA) test); premarital screening debacles; prenatal screening; the cost of screening versus effectiveness; the ineffectiveness of airport behavioral detection programs implemented by the Transportation

Security Administration (TSA); informed consent and screening; the social pressure to screen.

Contents

1 Clinical Efficiency

We begin by (re)introducing a 2×2 contingency table cross-classifying n individuals by events A and $\bar{A}$ and B and $\bar{B}$ but now with terminology attuned to a diagnostic testing context. The events B (positive) or $\bar{B}$ (negative) occur when the test says the person has "it" or doesn't have "it," respectively, whatever "it" may be. The events A (positive) or $\bar{A}$ (negative) occur when the "state of nature" is such that the person has "it" or doesn't have "it," respectively:

		state of nature		
		A (positive)	$\bar{A}$ (negative)	row sums
diagnostic	B (positive)	n_{BA}	$n_{B\bar{A}}$	n_B
test result	$\bar{B}$ (negative)	$n_{\bar{B}A}$	$n_{\bar{B}\bar{A}}$	$n_{\bar{B}}$
	column sums	n_A	$n_{\bar{A}}$	n

As in the introductory Module 1, a physical "urn" model is tacitly assumed that will generate a probability distribution according to the frequency distribution just given. There are n such balls in the urn with each ball labeled B or $\bar{B}$ and A or $\bar{A}$. There are n_{BA} balls with the labels B and A; $n_{B\bar{A}}$ balls with the labels B and $\bar{A}$; $n_{\bar{B}A}$ balls with the labels $\bar{B}$ and A; $n_{\bar{B}\bar{A}}$ balls with the labels $\bar{B}$ and $\bar{A}$. When a single ball is chosen from the urn "at random" and the two labels observed, a number of different event probabilities (and conditional probabilities) can be defined. For example, $P(B) = n_B/n$; $P(A) = n_A/n$; $P(A \text{ and } B) = n_{BA}/n$; $P(A|B) = n_{BA}/n_B$; and so on.

Using the urn model and conditionalizing on the state of nature, a number of common terms can be defined that are relevant to a diagnostic testing context:

		state of nature	
		A (pos)	$\bar{A}$ (neg)
diagnostic	B (pos)	$P(B\|A) = n_{BA}/n_A$ (sensitivity)	$P(B\|\bar{A}) = n_{B\bar{A}}/n_{\bar{A}}$ (false positive)
test result	$\bar{B}$ (neg)	$P(\bar{B}\|A) = n_{\bar{B}A}/n_A$ (false negative)	$P(\bar{B}\|\bar{A}) = n_{\bar{B}\bar{A}}/n_{\bar{A}}$ (specificity)
	column sums	$\frac{n_{BA}+n_{\bar{B}A}}{n_A} = 1.0$	$\frac{n_{B\bar{A}}+n_{\bar{B}\bar{A}}}{n_{\bar{A}}} = 1.0$

To give words to the two important concepts of test sensitivity and specificity, we have:
sensitivity = $P(B|A)$ = the probability that the test is positive if the person has "it";
specificity = $P(\bar{B}|\bar{A})$ = the probability that the test is negative if the person doesn't have "it."

Using the urn model and conditionalizing on the diagnostic test results, several additional terms relevant to a diagnostic testing context can be defined::

		state of nature		
		A (pos)	$\bar{A}$ (neg)	row sums
diagnostic	B (pos)	$P(A\|B) = n_{BA}/n_B$ (positive predictive value)	$P(\bar{A}\|B) = n_{B\bar{A}}/n_B$	$\frac{n_{BA}+n_{B\bar{A}}}{n_B} = 1.0$
test result	$\bar{B}$ (neg)	$P(A\|\bar{B}) = n_{\bar{B}A}/n_{\bar{B}}$	$P(\bar{A}\|\bar{B}) = n_{\bar{B}\bar{A}}/n_{\bar{B}}$ (negative predictive) value)	$\frac{n_{\bar{B}A}+n_{\bar{B}\bar{A}}}{n_{\bar{B}}} = 1.0$

Again, to give words to the two important concepts of the positive and negative predictive values, we have:

positive predictive value $= P(A|B) =$ the probability that the person has "it" if the test says the person has "it";
negative predictive value $= P(\bar{A}|\bar{B}) =$ the probability that the person doesn't have "it" if the test says the person doesn't have "it."

Assuming that $P(A) \leq 1/2$ (this, by the way, can always be done without loss of any generality because the roles of A and $\bar{A}$ can be interchanged), prediction according to base rates would be to consistently say that a person doesn't have "it" (because $P(\bar{A}) \geq P(A)$). The probability of being correct in this prediction is $P(\bar{A})$ (which is greater than or equal to 1/2). Prediction according to the test would be to say the person has "it" if the test is positive, and doesn't have "it" if the test is negative. Thus, the probability of a correct diagnosis according to the test (called the "hit rate" or "accuracy") is:

$$P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) =$$

$$(\frac{n_{BA}}{n_A})(\frac{n_A}{n}) + (\frac{n_{\bar{B}\bar{A}}}{n_{\bar{A}}})(\frac{n_{\bar{A}}}{n}) = \frac{n_{BA} + n_{\bar{B}\bar{A}}}{n} \ ,$$

which is just the sum of main diagonal frequencies in the 2 × 2 contingency table divided by the total sample size n.

A general condition can be given for when prediction by a test will be better than prediction by base rates (again, assuming that $P(A) \leq 1/2$). It is for the accuracy to be strictly greater than $P(\bar{A})$:

$$P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) > P(\bar{A}).$$

Based on this first general condition, we give three equivalent conditions for clinical efficiency to hold that we attribute to Meehl and

Rosen (1955), Dawes (1962), and Bokhari and Hubert (2015). This last reference provides a formal proof of equivalence.

Meehl-Rosen condition: assuming that $P(A) \leq 1/2$, it is best to use the test (over base rates) if and only if

$$P(A) > \frac{1 - P(\bar{B}|\bar{A})}{P(B|A) + (1 - P(\bar{B}|\bar{A}))} = \frac{1 - \text{specificity}}{\text{sensitivity} + (1 - \text{specificity})}.$$

Dawes condition: assuming that $P(A) \leq 1/2$, it is better to use the test (over base rates) if and only if $P(\bar{A}|B) < 1/2$ (or, equivalently, when $P(A|B) > 1/2$; that is, when the positive predictive value is greater than 1/2).

Bokhari-Hubert condition: assuming that $P(A) \leq 1/2$, it is better to use the test (over base rates) if and only if differential prediction holds between the row entries in the frequency table: $n_{BA} > n_{B\bar{A}}$ but $n_{\bar{B}A} < n_{\bar{B}\bar{A}}$. In words, given the B (positive) row, the frequency of positive states of nature, n_{BA}, is greater than or equal to the frequency of negative states of nature, $n_{B\bar{A}}$; the opposite occurs within the $\bar{B}$ (negative) row.

To give a numerical example of these conditions, the COVR 2×2 contingency table from Module 2 is used. Recall that this table reports a cross-validation of an instrument for the diagnostic assessment of violence risk (B: positive (risk present); $\bar{B}$: negative (risk absent)) in relation to the occurrence of followup violence (A: positive (violence present); $\bar{A}$: negative (violence absent)):

		state of nature		
		A (positive)	$\bar{A}$ (negative)	row sums
prediction	B (positive)	19	36	55
	$\bar{B}$ (negative)	9	93	102
	column sums	28	129	157

To summarize what this table shows, we first note that 2 out of 3 predictions of "dangerous" are wrong (.65 = 36/55, to be precise); 1 out of 11 predictions of "not dangerous" are wrong (.09 = 9/102, to be precise). The accuracy or "hit-rate" is .71 (= (10 + 93)/157). If everyone was predicted to be "not dangerous", we would be correct 129 out of 157 times, the base rate for $\bar{A}$: $P(\bar{A}) = 129/157 = .82$. Because this is better than the accuracy of .71, all three conditions will fail for when the test would do better than the base rates:

Meehl-Rosen condition: for a specificity = 93/129 = .72, sensitivity = 19/28 = .68, and $P(A) = 28/157 = .18$,

$$P(A) \not> \frac{1 - \text{specificity}}{\text{sensitivity} + (1 - \text{specificity})}$$

$$.18 \not> \frac{1 - .72}{.68 + (1 - .72)} = .29$$

Dawes condition: the positive predictive value of .35 = 19/55 is not greater than 1/2.

Bokhari-Hubert condition: there is no differential prediction because the row entries in the frequency table are ordered in the same direction.

1.1 Measuring the Degree of Clinical Efficiency

The Dawes condition described in the previous section shows the importance of clinical efficiency in the bottom-line justification for the use of a diagnostic instrument. When you can do better with base rates than with a diagnostic test, the Dawes condition implies that the positive predictive value is less than 1/2. In other words, it is more likely that a person doesn't have "it" than they do, even though the test says the person has "it." This anomalous circumstance has been called the "false positive paradox."

For base rates to be worse than the test, the Bokhari-Hubert condition requires differential prediction to exist; explicitly, within those predicted to be dangerous, the number who were dangerous (n_{BA}) must be greater than the number who were not dangerous ($n_{B\bar{A}}$); conversely, within those predicted to be not dangerous, the number who were not dangerous ($n_{\bar{B}\bar{A}}$) must be greater than those who were dangerous ($n_{\bar{B}A}$).

As a way of assessing the degree of clinical efficiency, the Goodman-Kruskal (λ) Index of Prediction Association can be adopted. The lambda coefficient is a proportional reduction in error measure for predicting a column event (A or $\bar{A}$) from knowledge of a row event (B or $\bar{B}$) over a naive prediction based just on marginal column frequencies. For the 2×2 contingency table of frequencies, it is defined as:

$$\lambda_{\text{column}|\text{row}} = \frac{\max\{n_{BA}, n_{B\bar{A}}\} + \max\{n_{\bar{B}A}, n_{\bar{B}\bar{A}}\} - \max\{n_A, n_{\bar{A}}\}}{n - \max\{n_A, n_{\bar{A}}\}}$$

If $\lambda_{\text{column}|\text{row}}$ is zero, the maximum of the column marginal frequencies is the same as the sum of the maximum frequencies within rows. In other words, no differential prediction of a column event is made based on knowledge of what particular row an object belongs to. A non-zero $\lambda_{\text{column}|\text{row}}$ is just another way of specifying the Bokhari-Hubert differential prediction condition. The upper limit for $\lambda_{\text{column}|\text{row}}$ is 1.0, which corresponds to perfect prediction with the diagnostic test, and where test accuracy is 1.0.

To justify $\lambda_{\text{column}|\text{row}}$ as an index of clinical efficiency through a "proportional reduction in error measure," suppose the Bokhari-Hubert condition holds for the 2×2 contingency table and assume that $P(A) \leq 1/2$. Now, consider a ball picked randomly from the urn, and that we are asked to predict the "state of nature" in the absence of any information about the diagnostic test result; we would predict $\bar{A}$ (negative) and be wrong with probability $n_A/n = P(A)$. If asked to predict the "state of nature" but were told there is a diagnostic test result of B (positive) for this randomly selected ball, we would predict A (positive) and be wrong $n_{B\bar{A}}/n_B = P(\bar{A}|B)$. If the test result is $\bar{B}$ (negative), we would predict $\bar{A}$ (negative) and be wrong with probability $n_{\bar{B}A}/n_B = P(A|\bar{B})$. Thus, incorporating the probability of picking a ball from B or $\bar{B}$, the probability of error when given the diagnostic test result must be $P(\bar{A}|B)P(B)+P(A|\bar{B})P(\bar{B})$. Recalling that the probability of error when not knowing the diagnostic test result is $P(A)$, consider the proportional reduction in error measure defined by

$$\frac{P(A) - [P(\bar{A}|B)P(B) + P(A|\bar{B})P(\bar{B})]}{P(A)} .$$

After some simple algebra, this reduces to $\lambda_{\text{column}|\text{row}}$.

It might be noted in passing that significance testing in a 2 × 2 table with the usual chi-squared test of association tells us nothing about differential prediction. For example, the chi-squared test could show a significant relation between the A and $\bar{A}$, and the B and $\bar{B}$ events, but if $\lambda_{\text{column}|\text{row}}$ is zero, there is no differential prediction, and therefore base rates will outperform the use of a diagnostic test. More generally, when attempting to predict an event having a low base rate (for example, "dangerous") by using a "test" possessing less than ideal sensitivity and specificity values, it is common to be more accurate in prediction merely by using the larger base rate (for example, "not dangerous") rather than the diagnostic test.

One might conclude that it is ethically questionable to use a clinically inefficient test. If you can't do better than just predicting with base rates, what is the point of using the diagnostic instrument in the first place. The only mechanism that we know of that might justify the use of a clinically inefficient instrument would be to adopt severe unequal costs in the misclassification of individuals (that is, the cost of predicting "dangerous" when the "state of nature" is "not dangerous," and in predicting "not dangerous" when the "state of nature" is "dangerous").[1]

The Bokhari and Hubert paper (2015) that discusses the three equivalent statements for clinical efficiency, also gives a generalized clinical efficiency condition (a generalized Bokhari-Hubert condition [GBH]) that allows for the assignment of unequal costs to the false

[1] But here we would soon have to acknowledge Sir William Blackstone's dictum (1765): "It is better that ten guilty escape than one innocent suffer."

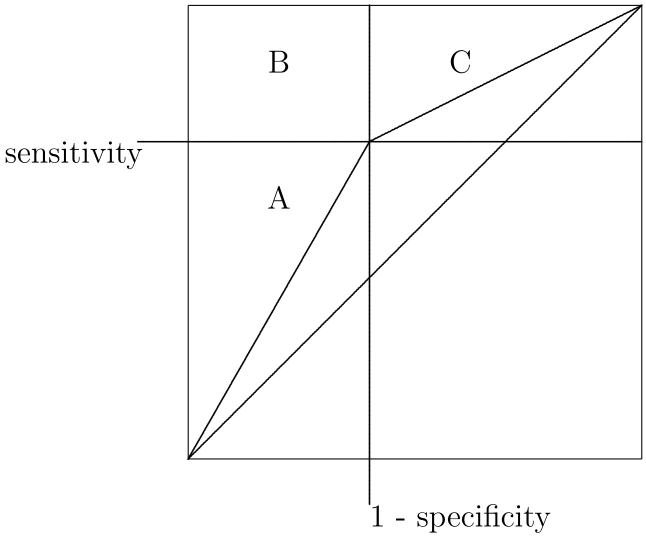
positives and false negatives. Depending on how the costs of misclassification are assigned, a determination can be made as to when generalized clinical efficiency holds; that is, when is the total costs of using a test less than the total costs obtained by just classifying through base rates? Further, depending on the specific data available in the 2×2 contingency table (such as that for the COVR instrument given earlier in this section), statements such as the following can be made based on explicit bounds given in Bokhari and Hubert (2015): for generalized clinical efficiency to hold, false negatives (releasing a dangerous person) cannot be considered more than 10.3 times more costly than false positiveS (detaining a non-dangerous person); also, one needs to have false negatives be more than twice as costly as false positives. So, in summary, false negatives must be at least twice as costly as false positives but no more than about ten times as costly.

When interests center on the prediction of a very infrequent event (such as the commission of suicide) and the cost of a false negative (releasing a suicidal patient) is greater than the cost of a false positive (detaining a non-suicidal patient), there still may be such a large number of false positives that implementing and acting on such a prediction system would be infeasible. An older discussion of this conundrum is by Albert Rosen, "Detection of Suicidal Patients: An Example of Some Limitations in the Prediction of Infrequent Events," *Journal of Consulting Psychology* (*18*, 1954, 397–403).

2 Diagnostic Test Evaluation

The Receiver Operating Characteristic (ROC) curve of a diagnostic test is a plot of test sensitivity (the probability of a "true" posi-

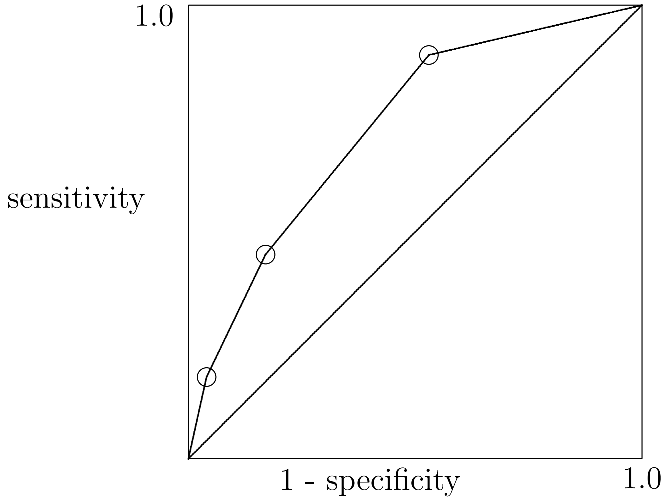
Figure 1: An ROC curve for a diagnostic test having just one cutscore.



tive) against 1.0 minus test specificity (the probability of a "false" positive). As shown in Figure 1, when there is a single 2 × 2 contingency table, the ROC plot would be based on a single point. In some cases, however, a diagnostic test might provide more than a simple dichotomy (for example, more than a value of 0 or 1, denoting a negative or a positive decision, respectively), and instead gives a numerical range (for example, integer scores from 0 to 20, as in the illustration to follow on the Psychopathy Checklist, Screening Version (PCL:SV)). In these latter cases, different possible "cutscores" might be used to reflect differing thresholds for a negative or a positive decision. Figure 2 gives the ROC plot for the PCL:SV discussed below using three possible cutscores.

The ROC curve is embedded in a box having unit-length sides. It begins at the origin defined by a sensitivity of 0.0 and a specificity of 1.0, and ends at a sensitivity of 1.0 and a specificity of 0.0. Along the way, the ROC curve goes through the various sensitivity and 1.0 −

Figure 2: An ROC curve for the PCL:SV having three cutscores.



specificity values attached to the possible cutscores. The diagonals in both Figures 1 and 2 represent lines of "no discrimination" where sensitivity values are equal to 1.0 minus specificity values. Restating, we have $P(B|A) = 1 - P(\bar{B}|\bar{A})$, and finally, $P(B|A) = P(B|\bar{A})$. This last equivalence provides an interpretation for the "no discrimination" phrase: irrespective of the "state of nature" (A or $\bar{A}$), the probability of a "yes" prediction remains the same.

For an ROC curve to represent a diagnostic test that is performing better than "chance," it has to lie above the "no discrimination" line where the probabilities of "true" positives exceed the probabilities of "false" positives (or equivalently, where sensitivities are greater than 1.0 minus the specificities). The characteristic of good diagnostic tests is the degree to which the ROC curve "gets close to hugging" the left and top line of the unit-area box and where the sensitivities are much bigger than 1.0 minus specificities. The most common summary measure of diagnostic test performance is the "area under the curve" (AUC), which ranges from an effective lower value of .5

(for the line of "no discrimination") to 1.0 for a perfect diagnostic test with sensitivity and specificity values both equal to 1.0. So, as an operational comparison of diagnostic test performances, those with bigger AUCs are better.

2.1 An Example Using the Psychopathy Checklist, Screening Version (PCL:SV): Data From the MacArthur Risk Assessment Study

The Psychopathy Checklist, Screening Version (PCL:SV) is the single best variable for the prediction of violence based on the data from the MacArthur Risk Assessment Study. It consists of twelve items, with each item being scored 0, 1, or 2 during the course of a structured interview. The items are identified below by short labels:

1) Superficial; 2) Grandiose; 3) Deceitful; 4) Lacks Remorse; 5) Lacks Empathy; 6) Doesn't Accept Responsibility; 7) Impulsive; 8) Poor Behavioral Controls; 9) Lacks Goals; 10) Irresponsible; 11) Adolescent Antisocial Behavior; 12) Adult Antisocial Behavior

The total score on the PCL:SV ranges from 0 to 24, with higher scores supposedly more predictive of dangerousness and/or violence.

Based on the MacArthur Risk Assessment Study data of Table 1, the three cutscores of 6, 12, and 18 were used to predict violence at followup (that is, when above or at a specific cutscore, predict "violence"; when below the cutscore, predict "nonviolence"). The basic statistics for the various diagnostic test results are given below:

Cutscore of 6:

Table 1: Data from the MacArthur Risk Assessment Study on the Psychopathy Checklist, Screening Version.

PCL-SV	block	violence at followup		block	totals
Score	yes	yes	no	no	
0	18	0	34	328	34
1		1	45		46
2		1	54		55
3		6	48		54
4		1	57		58
5		4	41		45
6		5	49		54
7	69	8	51	254	59
8		10	57		67
9		13	38		51
10		9	40		49
11		16	31		47
12		13	37		50
13	43	12	19	93	31
14		9	14		23
15		7	26		33
16		3	13		16
17		7	10		17
18		5	11		16
19	29	10	10	26	20
20		5	6		11
21		4	1		5
22		5	5		10
23		0	2		2
24		5	2		7
totals		159	701		860

		violence		
		Yes (A)	No ($\bar{A}$)	row sums
prediction	Yes (B)	141	373	414
	No ($\bar{B}$)	18	328	446
	column sums	159	701	860

accuracy: $(141 + 328)/860 = .55$
base rate: $(373 + 328)/860 = 701/860 = .815 \approx .82$
sensitivity: $141/159 = .89$
specificity: $328/701 = .47$
positive predictive value: $141/414 = .34$
negative predictive value: $328/446 = .74$

Cutscore of 12:

		violence		
		Yes (A)	No ($\bar{A}$)	row sums
prediction	Yes (B)	72	119	191
	No ($\bar{B}$)	87	582	669
	column sums	159	701	860

accuracy: $(72 + 582)/860 = .76$
base rate: $701/860 = .815 \approx .82$
sensitivity: $72/159 = .45$
specificity: $582/701 = .83$
positive predictive value: $72/191 = .38$
negative predictive value: $582/669 = .87$

Cutscore of 18:

		violence		
		Yes (A)	No ($\bar{A}$)	row sums
prediction	Yes (B)	29	26	55
	No ($\bar{B}$)	130	675	805
	column sums	159	701	860

accuracy: (29 +675)/860 = 704/860 = .819 ≈ .82 (which is slightly better than using base rates)
base rate: 701/860 = .815 ≈ .82
sensitivity: 29/159 = .18
specificity: 675/701 = .96
positive predictive value: 29/55 = .53
negative predictive value: 675/805 = .84

As noted earlier, a common measure of diagnostic adequacy is the area under the ROC curve (or AUC). Figure 2 gives the ROC plot for the PCL:SV data based on the following sensitivity and 1.0 − specificity values:

cutscore	sensitivity	specificity	1 - specificity
6	.89	.47	.53
12	.45	.83	.17
18	.18	.96	.04

The AUC in this case has a value of .73, as computed in the section to follow. Only the cutpoint of 18 gives a better accuracy than using base rates, and even here, the accuracy is only minimally better than with the use of base rates: $704/860 = .819 > 701/860 = .815$. Also, the area under the ROC curve is not necessarily a good measure of clinical efficiency because it does not incorporate base rates. It is only a function of the test itself and not of its use on a sample of individuals.

Figure 1 helps show the independence of base rates for the AUC; the AUC is simply the average of sensitivity and specificity when only one cutscore is considered, and neither sensitivity or specificity is a function of base rates:

A = (1 - sens)(1 - spec)
B = (1/2)(1 - spec)(sens)
C = (1/2)(1 - sens)(spec)
AUC = 1.0 - (A + B + C) = (1/2)(sensitivity + specificity)

We can also see explicitly how different normalizations (using base rates) are used in calculating an AUC or accuracy:

$P(B|A) = n_{BA}/n_A =$ sensitivity
$P(\bar{B}|\bar{A}) = n_{\bar{B}\bar{A}}/n_{\bar{A}} =$ specificity
AUC $= ((n_{BA}/n_A) + (n_{\bar{B}\bar{A}})/n_{\bar{A}})/2$
accuracy $= (n_{BA} + n_{\bar{B}\bar{A}})/n$ $(= P(A|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B}))$

Note that only when $n_A = n_{\bar{A}}$ (that is, when the base rates are equal), are accuracy and the AUC identical. In instances of unequal base rates, the AUC can be a poor measure of diagnostic test usage in a particular sample. We will come back to this issue shortly and suggest several alternative measures to the AUC that do take base rates into consideration when evaluating the use of diagnostic tests in populations where one of the base rates may be small, such as in the prediction of "dangerous" behavior.

2.2 The Wilcoxon Test Statistic Interpretation of the AUC

As developed in detail by Hanley and McNeil (1982), it is possible to calculate numerically the AUC for an ROC curve that is constructed for multiple cutscores by first computing a well-known two-sample Wilcoxon test statistic. Given two groups of individuals each with a score on some test, the Wilcoxon test statistic can be interpreted as follows: choose a pair of individuals at random (and with replacement) from the two groups (labeled A and $\bar{A}$, say, in anticipation of usage to follow), and assess whether the group A score is greater than the group $\bar{A}$ score. If this process is continued and the proportion of group A scores greater than those from group $\bar{A}$ is computed, this later value will converge to the proportion of all possible pairs constructed from the groups A and $\bar{A}$ in which the value for the A group member is greater than or equal to that for the $\bar{A}$ group member. In particular, we ask for the probability that in a randomly selected pair of people, where one committed violence and the other did not, the psychopathy score for the person committing violence is greater than that for the person not committing violence. This is the same as the two-sample Wilcoxon statistic (with a caveat that we will need to have a way of dealing with ties); it is also an interpretation for the AUC.

What follows is an example of the Wilcoxon test statistic calculation that relates directly back to the PCL:SV results of Table 1 and the computation of the AUC for Figure 2. Specifically, we compute the Wilcoxon statistic for a variable with four ordinal levels (I, II, III, and IV, with the IV level being the highest, as it is in the PCL:SV example):

	Violence	Present
	Yes (A)	No ($\bar{A}$)
I	m_{11}	m_{12}
II	m_{21}	m_{22}
III	m_{31}	m_{32}
IV	m_{41}	m_{42}
totals	n_A	$n_{\bar{A}}$

There is a total of $n_A n_{\bar{A}}$ pairs that can be formed from groups A and $\bar{A}$. The number of pairs for which the group A score is strictly greater than the group $\bar{A}$ score is:

$$\begin{gathered}\{m_{12}(m_{21}+m_{31}+m_{41})\}+\\ \{m_{22}(m_{31}+m_{41})\}+\\ \{m_{32}(m_{41})\}\end{gathered}$$

The number of pairs for which there is a tie on the ordinal variable is:

$$(m_{11}m_{12})+(m_{21}m_{22})+(m_{31}m_{32})+(m_{41}m_{42})$$

By convention, the Wilcoxon test statistic is the number of "strictly greater" pairs plus one-half of the "tied" pairs, all divided by the total number of pairs:

$$\begin{gathered}[\{m_{12}(m_{21}+m_{31}+m_{41})+(1/2)(m_{11}m_{12})\}+\\ \{m_{22}(m_{31}+m_{41})+(1/2)(m_{21}m_{22})\}+\\ \{m_{32}(m_{41})+(1/2)(m_{31}m_{32})\}+\{(1/2)(m_{41}m_{42})\}]/[n_A n_{\bar{A}}]\end{gathered}$$

For the PCL:SV results of Table 1:

	Violence	Present	
	Yes (A)	No ($\bar{A}$)	row totals
I	18	328	346
II	69	254	323
III	43	93	136
IV	29	26	55
column totals	159	701	860

the Wilcoxon test statistic = 81,701.5/111,459.0 = .73 = AUC.

Using only the cutscore of 18:

	Violence	Present
	Yes(A)	No($\bar{A}$)
(No) (I + II + III)	130	675
(Yes) (IV)	29	26
column totals	159	701

the Wilcoxon statistic =

$$[(675)(29) + (1/2)(675)(130) + (1/2)(26)(29)]/[(159)(701)] = .57 \; ;$$

here, the AUC is merely defined by the average of sensitivity and specificity: (.18 + .96)/2 = .57

The relation just shown numerically can also be given in the notation used for the general Wilcoxon test:

sensitivity = m_{21}/n_A

specificity = $m_{12}/n_{\bar{A}}$

So, the average of sensitivity and specificity $((1/2)((m_{21}/n_A)+(m_{12}/n_{\bar{A}})))$ is equal to (after some algebra) the Wilcoxon statistic $(m_{12}m_{21} + (1/2)m_{22}m_{21} + (1/2)m_{11}m_{12})$.

2.3 A Modest Proposal for Evaluating a Diagnostic Test When Different Cutscores Can Be Set

One suggestion for evaluating a diagnostic test when different cutscores are possible is to set a cutscore so that the proportion of positive predictions is "close" to the prior probability of a positive "state of nature" — and to then look at the consistency of subject classifications by A and B and by $\bar{A}$ and $\bar{B}$. To give an example, we use the PCL:SV data and a cutscore of 13:

		violence		
		Yes (A)	No ($\bar{A}$)	row sums
prediction	Yes (B)	60	100	160
	No ($\bar{B}$)	99	601	700
	column sums	159	701	860

accuracy: $(60 + 601)/860 = .77$
base rate: $701/860 = .815 \approx .82$
sensitivity: $60/159 = .38$
specificity: $601/701 = .86$
positive predictive value: $60/160 = .38$
negative predictive value: $601/700 = .86$

Here, $P(A \cap B | A \cup B)$ = the proportion of positive classifications (by A or B) that are consistent $= 60/(60 + 100 + 99) = 60/259 = .23$; so, only 1/4 of the time are the positive classifications consistent; $P(\bar{A} \cap \bar{B} | \bar{A} \cup \bar{B})$ = the proportion of negative classifications (by $\bar{A}$ or $\bar{B}$) that are consistent $= 601/(601 + 100 + 99) = 601/800 = .75$; so, 3/4 of the time the negative classifications are consistent.[2]

[2] These two types of consistency index just presented may be of particular value when

Note that from Bayes' theorem, we have the two statements:

$$P(A|B) = P(B|A)(\frac{P(A)}{P(B)}) \ ,$$

and

$$P(\bar{A}|\bar{B}) = P(\bar{B}|\bar{A})(\frac{P(\bar{A})}{P(\bar{B})}) \ .$$

If $P(A) = P(B)$ (and thus, $P(\bar{A}) = P(\bar{B})$), $P(A|B) = P(B|A)$ and $P(\bar{A}|\bar{B}) = P(\bar{B}|\bar{A})$. Or, in words, the positive predictive value is equal to the sensitivity, and the negative predictive value is equal to the specificity. This is seen numerically in the example given above where $P(A)$ and $P(B)$ are very close (that is, $P(A) = .185; P(B) = .186$).

Possibly the use of these measures will eliminate the terminological confusion about what a "false positive" means; one usual interpretation is 1 - specificity (which does not take base rates into account): the probability that the test is positive given that the person doesn't have "it"; the other is 1 - the negative predictive value (which does take base rates into account): the probability that the person has "it" given that the test is negative. Also, for a "false negative," the usual interpretation is 1 - sensitivity (which does not take base rates into account): the probability that the test is negative given that the person has "it"; the other is 1 - positive predictive value (which does take base rates into account): the probability that the person doesn't have "it" given that the test is positive. By equating $P(A)$

two distinct diagnostic tests are to be compared. Here, no explicit "state of nature" pair of events (A and $\bar{A}$) would be available, but one of the diagnostic tests would serve the same purpose.

and $P(B)$, the confusions about the meaning of a "false positive" and a "false negative" can be finessed because different interpretations can be given as to what is "false" and what is "positive" or "negative."

Because of the equivalence of sensitivity and the positive predictive value and of specificity and the negative predictive value when the base rates $P(A)$ and $P(B)$ are equal, another measure of diagnostic accuracy but one that does take base rates into account would be the simple average of the positive and negative predictive values. This would correspond to an AUC measure for the single cutpoint that equalizes the base rates $P(A)$ and $P(B)$; that AUC measure would be, as usual, the simple average of specificity and sensitivity.

3 Summary Comments

The answer we have for the general question of "how should a diagnostic test be evaluated?" is in contrast to current widespread practice. Whenever the base rate for the condition being assessed is relatively low (for example, for "dangerous" behavior), the area under the ROC curve (AUC) is not necessarily a good measure for conveying the adequacy of the actual predictions made from a diagnostic test. The AUC does not incorporate information about base rates. It only evaluates the test itself and not how the test actually performs when used on a specific population with differing base rates for the presence or absence of the condition being assessed.

The use of AUC as a measure of diagnostic value can be very misleading in assessing conditions with unequal base rates, such as being "dangerous." This misinformation is further compounded when

AUC measures become the basic data subjected to a meta-analysis. Our general suggestion is to rely on some function of the positive and negative predictive values to evaluate a diagnostic test. These measures incorporate both specificity and sensitivity as well as the base rates in the sample for the presence or absence of the condition under study.

A simple condition given in an earlier section of this module (and attributed to Robyn Dawes) points to a minimal condition that a diagnostic test should probably satisfy (and which leads to prediction with the test being better than just prediction according to base rates): the positive predictive value must be greater than 1/2. If this minimal condition does not hold, it will be more likely that a person doesn't have "it" than they do, even where the test says the person has "it." As noted earlier, this situation is so unusual that it has been referred to as the "false positive paradox."

As an another measure of diagnostic accuracy we might consider a weighted function of the positive and negative predictive values, such as the simple proportion of correct decisions. When the positive and negative predictive values are each weighted by the probabilities that the diagnostic test is positive or negative, and these values then summed, the simple measure of accuracy (defined as the proportion of correct decisions) is obtained.

Just saying that a measure is "good" because it is independent of base rates doesn't make it "good" for the use to which it is being put (or, in the jargon of computer science, a "bug" doesn't suddenly become a "feature" by bald face assertion). As an example from the MacArthur data given in Module 2 on the cross-validation of an actuarial model of violence risk assessment, the AUC would be given as the simple average of sensitivity and specificity (AUC = (.68 + .72)/2 = .70). This number tells us precious little of importance in how the diagnostic test is doing with the cross-validation sample. The (very low) accuracy or "hit-rate" measure is .71, which is worse than just using the base rate (.82) and predicting that everyone will be "not dangerous." Using the test, 2 out of 3 predictions of dangerousness are wrong; 1 out of 11 predictions of "not dangerous" are wrong. It is morally questionable to have one's liberty jeopardized by an assessment of being " dangerous" that is wrong 2 out of 3 times (or, in some Texas cases, one's life, such as in Barefoot v. Estelle (1983) discussed at length in Module 2).

In contrast to some incorrect understandings in the literature about the invariance of specificity and sensitivity across samples, sizable subgroup variation can be present in the sensitivity and specificity values for a diagnostic test; this is called "spectrum bias" and is discussed thoroughly by Ransohoff and Feinstein (1978). Also, sensitivities and specificities are subject to a variety of other biases that have been known for some time (for example, see Begg, 1971). In short, because ROC measures are generally *not* invariant across subgroups, however formed, we do not agree with the sentiment expressed in the otherwise informative review article by John A. Swets, Robyn M. Dawes, and John Monahan, "Psychological Science Can Improve Diagnostic Decisions," *Psychological Science in the Public Interest* (2000, *1*, 1–26). We quote:

> These two probabilities [sensitivity and specificity] are independent of the prior probabilities (by virtue of using the priors in the denominators of their defining ratios). The significance of this fact is that ROC measures do not

depend on the proportions of positive and negative instances in any test sample, and hence, generalize across samples made up of different proportions. All other existing measures of accuracy vary with the test sample's proportions and are specific to the proportions of the sample from which they are taken.

A particularly pointed critique of the sole reliance on specificity and sensitivity (and thus on the AUC) is given in an article by Karel Moons and Frank Harrell (*Academic Radiology*, *10*, 2003, 670–672), entitled "Sensitivity and Specificity Should Be De-emphasized in Diagnostic Accuracy Studies." We give several telling paragraphs from this article below:

> ... a single test's sensitivity and specificity are of limited value to clinical practice, for several reasons. The first reason is obvious. They are reverse probabilities, with no direct diagnostic meaning. They reflect the probability that a particular test result is positive or negative given the presence (sensitivity) or absence (specificity) of the disease. In practice, of course, patients do not enter a physician's examining room asking about their probability of having a particular test result given that they have or do not have a particular disease; rather, they ask about their probability of having a particular disease given the test result. The predictive value of test results reflects this probability of disease, which might better be called "posttest probability."
>
> It is well known that posttest probabilities depend on disease prevalence and therefore vary across populations and across subgroups within a particular population, whereas sensitivity and specificity do not depend on the prevalence of the disease. Accordingly, the latter are commonly considered characteristics or constants of a test. Unfortunately, it is often not realized that this is a misconception.
>
> Various studies in the past have empirically shown that sensitivity, specificity, and likelihood ratio vary not only across different populations but also across different subgroups within particular populations.
>
> ...

> Since sensitivity and specificity have no direct diagnostic meaning and vary across patient populations and subgroups within populations, as do posttest probabilities, there is no advantage for researchers in pursuing estimates of a test's sensitivity and specificity rather than posttest probabilities. As the latter directly reflect and serve the aim of diagnostic practice, researchers instead should focus on and report the prevalence (probability) of a disease given a test's result – or even better, the prevalence of a disease given combinations of test results.

Finally, because sensitivity and specificity are calculated from frequencies present in a 2×2 contingency table, it is always best to remember the operation of Berkson's fallacy—the relationship that may be present between two dichotomous variables in one population may change dramatically for a selected sample based on some other variable or condition, for example, hospitalization, being a volunteer, age, and so on.

4 Issues in Medical Screening

It might be an obvious statement to make, but in our individual dealings with doctors and the medical establishment generally, it is important for all to understand the positive predictive values (PPVs) for whatever screening tests we now seem to be constantly subjected to, and thus, the number, $(1 - \text{PPV})$, referring to the false positives; that is, if a patient tests positive, what is the probability that "it" is not actually present. It is a simple task to plot PPV against $P(A)$ from 0 to 1 for any given pair of sensitivity and specificity values. Such a plot can show dramatically the need for highly reliable tests in the presence of low base rate values for $P(A)$ to attain even mediocre PPV values.

Besides a better understanding of how PPVs are determined, there is a need to recognize that even when a true positive exists, not every disease needs to be treated. In the case of another personal favorite of ours, prostate cancer screening, its low accuracy makes mammograms look good, where the worst danger is one of overdiagnosis and overtreatment, leading to more harm than good (see, for example, Gina Kolata, "Studies Show Prostate Test Save Few Lives," *New York Times*, March 19, 2009). Armed with this information, we no longer give blood for a PSA screening test. When we so informed our doctors as to our wishes, they agreed completely. The only reason such tests were done routinely was to practice "defensive medicine" on behalf of their clinics, and to prevent possible lawsuits arising from such screening tests not being administered routinely. In other words, clinics get sued for underdiagnosis but not for overdiagnosis and overtreatment.[3]

[3] We list several additional items that are relevant to screening: an article by Sandra G. Boodman for the *AARP Bulletin* (Januaary 1, 2010) summarizes well what its title offers: "Experts Debate the Risks and Benefits of Cancer Screening." A cautionary example of breast cancer screening that tries to use dismal specificity and sensitivity values for detecting the HER2 protein, is by Gina Kolata, "Cancer Fight: Unclear Tests for New Drug," *New York Times*, April 19, 2010). The reasons behind proposing cancer screening guidelines and the contemporary emphasis on evidence-based medicine is discussed by Gina Kolata in "Behind Cancer Guidelines, Quest for Data" (*New York Times*, November 22, 2009). Other articles that involve screening discuss how a fallible test for ovarian cancer (based on the CA-125 protein) might be improved using a particular algorithm to monitor CA-125 fluctuations more precisely (Tom Randall, *Bloomberg Businessweek*, May 21, 2010, "Blood Test for Early Ovarian Cancer May Be Recommended for All"); three items by Gina Kolata concern food allergies (or nonallergies, as the case may be) and a promising screening test for Alzheimer's: "Doubt Is Cast on Many Reports of Food Allergies" (*New York Times*, May 11, 2010); and "I Can't Eat That. I'm Allergic" (*New York Times*, May 15, 2010); "Promise Seen for Detection of Alzheimer's" (*New York Times*, June 23, 2010); a final item to mention discusses a promising alternative to mammogram screening: "Breast Screening Tool Finds Many Missed Cancers" (Janet Raloff, *ScienceNews*, July 1, 2010).

A good way to conclude this discussion of issues involving (cancer) screening is to refer the reader to three items from the *New York Times*: an OpEd article ("The Great Prostate Mistake," March 9, 2010) by Richard J. Ablin, a recent piece by Gina Kolata summarizing a large longitudinal randomized controlled Canadian study on the value of mammograms ("Vast Study Casts Doubt On Value of Mammograms"; February 11, 2014), and a second article by Gina Kolata on the severe overdiagnosis of thyroid cancer in South Korea.

Dr. Ablin is a research professor of immunobiology and pathology at the University of Arizona College of Medicine, and President of the Robert Benjamin Ablin Foundation for Cancer Research. Most importantly for our purposes, he is the individual who in 1970 discovered the PSA test for detecting prostate cancer; his perspective on the issues is therefore unique:[4]

[4] To show the ubiquity of screening appeals, we reproduce a solicitation letter to LH from Life Line Screening suggesting that for only $139, he could get four unnecessary screenings right in Champaign, Illinois, at the Temple Baptist Church:

Dear Lawrence,

Temple Baptist Church in Champaign may not be the location that you typically think of for administering lifesaving screenings. However, on Tuesday, September 22, 2009, the nation's leader in community-based preventive health screenings will be coming to your neighborhood.

Over 5 million people have participated in Life Line Screening's ultrasound screenings that can determine your risk for stroke caused by carotid artery diseases, abdominal aortic aneurysms and other vascular diseases. Cardiovascular disease is the #1 killer in the United States of both men and women—and a leading cause of permanent disability.

Please read the enclosed information about these painless lifesaving screenings. A package of four painless Stroke, Vascular Disease & Heart Rhythm screenings costs only $139. Socks and shoes are the only clothes that will be removed and your screenings will be completed in a little more than an hour.

You may think that your physician would order these screenings if they were necessary. However, insurance companies typically will not pay for screenings unless there are symptoms. Unfortunately, 4 out of 5 people that suffer a stroke have no apparent symptoms or warning signs. That is why having a Life Line Screening is so important to keep you and

I never dreamed that my discovery four decades ago would lead to such a profit-driven public health disaster. The medical community must confront reality and stop the inappropriate use of P.S.A. screening. Doing so would save billions of dollars and rescue millions of men from unnecessary, debilitating treatments.

Several excerpts are provided below from the Gina Kolata article on the Canadian mammogram study:

One of the largest and most meticulous studies of mammography ever done, involving 90,000 women and lasting a quarter-century, has added powerful new doubts about the value of the screening test for women of any age.

It found that the death rates from breast cancer and from all causes were the same in women who got mammograms and those who did not. And the screening had harms: One in five cancers found with mammography and treated was not a threat to the woman's health and did not need treatment such as chemotherapy, surgery or radiation.

The study, published Tuesday in *The British Medical Journal*, is one of the few rigorous evaluations of mammograms conducted in the modern era of more effective breast cancer treatments. It randomly assigned Canadian

your loved ones healthy and independent.

"These screenings can help you avoid the terrible consequences of stroke and other vascular diseases. I've seen firsthand what the devastating effects of stroke, abdominal aortic aneurysms and other vascular diseases can have on people and I feel it is important that everyone be made aware of how easily they can be avoided through simple, painless screenings." — Andrew Monganaro, MD, FACS, FACC (Board Certified Cardiothoracic and Vascular Surgeon)

I encourage you to talk to your physician about Life Line Screening. I am confident that he or she will agree with the hundreds of hospitals that have partnered with us and suggest that you participate in this health event.

We are coming to Champaign for one day only and appointments are limited, so call 1-800-395-1801 now.

Wishing you the best of health,
Karen R. Law, RDMS, RDCS, RVT
Director of Clinical Operations

> women to have regular mammograms and breast exams by trained nurses or to have breast exams alone.
>
> Researchers sought to determine whether there was any advantage to finding breast cancers when they were too small to feel. The answer is no, the researchers report.
>
> ...
>
> Dr. Kalager, whose editorial accompanying the study was titled "Too Much Mammography," compared mammography to prostate-specific antigen screening for prostate cancer, using data from pooled analyses of clinical trials. It turned out that the two screening tests were almost identical in their overdiagnosis rate and had almost the same slight reduction in breast or prostate deaths.
>
> "I was very surprised," Dr. Kalager said. She had assumed that the evidence for mammography must be stronger since most countries support mammography screening and most discourage PSA screening.

Finally, and as noted above, a recent example of a medical screening fiasco and the resulting overdiagnoses and overtreatments, involves thyroid cancer, and the detection of tiny and harmless tumors that are better left undisturbed. The situation is particularly serious in South Korea, as pointed out by the excerpts given below from an article by Gina Kolata ("Study Warns Against Overdiagnosis of Thyroid Cancer," *New York Times*, November 5, 2014):

> To the shock of many cancer experts, the most common cancer in South Korea is not lung or breast or colon or prostate. It is now thyroid cancer, whose incidence has increased fifteenfold in the past two decades. "A tsunami of thyroid cancer," as one researcher puts it.
>
> Similar upward trends for thyroid cancer are found in the United States and Europe, although not to the same degree. The thyroid cancer rate in the United States has more than doubled since 1994.
>
> Cancer experts agree that the reason for the situation in South Korea and elsewhere is not a real increase in the disease. Instead, it is down to screening,

which is finding tiny and harmless tumors that are better left undisturbed, but that are being treated aggressively.

South Koreans embraced screening about 15 years ago when the government started a national program for a variety of cancers – breast, cervix, colon, stomach and liver. Doctors and hospitals often included ultrasound scans for thyroid cancer for an additional fee of $30 to $50.

Since South Korea adopted widespread cancer screening in 1999, thyroid cancer has become the most diagnosed cancer in the country. But if this early detection were saving lives, the already-low death rate from thyroid cancer should have fallen, not remained steady.

In the United States and Europe, where there are no formal, widespread screening programs for thyroid cancer, scans for other conditions, like ultrasound exams of the carotid artery in the neck or CT scans of the chest, are finding tiny thyroid tumors.

Although more and more small thyroid cancers are being found, however, the death rate has remained rock steady, and low. If early detection were saving lives, death rates should have come down.

That pattern – more cancers detected and treated but no change in the death rate – tells researchers that many of the cancers they are finding and treating were not dangerous. It is a phenomenon that researchers call overdiagnosis, finding cancers that did not need treatment because they were growing very slowly or not at all. Left alone, they would probably never cause problems. Overdiagnosis is difficult to combat. Pathologists cannot tell which small tumors are dangerous, and most people hear the word "cancer" and do not want to take a chance. They want the cancer gone.

But cancer experts said the situation in South Korea should be a message to the rest of the world about the serious consequences that large-scale screening of healthy people can have.

"It's a warning to us in the U.S. that we need to be very careful in our advocacy of screening," said Dr. Otis W. Brawley, chief medical officer at the American Cancer Society. "We need to be very specific about where we have good data that it saves lives."

Colon cancer screening wins Dr. Brawley's unqualified endorsement. Breast cancer screening saves lives, he said, and he advocates doing it, but he said it could also result in overdiagnosis. Even lung cancer screening can be susceptible to overdiagnosis, with as many as 18 percent of patients treated when they did not need to be, Dr. Brawley said.

The soaring increase in thyroid cancers in South Korea is documented in a paper published on Thursday in the New England Journal of Medicine. The authors report not only that the number of diagnoses escalated as screening became popular, but also that the newly detected cancers were almost all very tiny ones. These tiny cancers, called papillary thyroid cancers, are the most common kind and are the sort typically found with screening. They are known to be the least aggressive.

The epidemic was not caused by an environmental toxin or infectious agent, said Dr. H. Gilbert Welch of Dartmouth, an author of the paper. “An epidemic of real disease would be expected to produce a dramatic rise in the number of deaths from disease,” he said. “Instead we see an epidemic of diagnosis, a dramatic rise in diagnosis and no change in death.”

Cancer experts stress that some thyroid cancers are deadly – usually they are the larger ones. And, they say, if a person notices symptoms like a lump on the neck or hoarseness, they should not be ignored.

“But there is a real difference between not ignoring something obvious and telling the population to try really hard to find something wrong,” Dr. Welch said.

Thyroid cancer tends to be particularly indolent. On autopsy, as many as a third of people have tiny thyroid cancers that went undetected in their lifetime. Once a cancer is found, though, treatment is onerous and involves removing the thyroid. Patients must then take thyroid hormones for the rest of their lives. For some, Dr. Brawley said, the replacement hormones are not completely effective, and they end up with chronically low thyroid hormone levels, feeling depressed and sluggish as a result.

In a small percentage of those having thyroid surgery, surgeons accidentally damage the nearby vocal cords – that happened to the 2 percent of South Korean patients who ended up with vocal cord paralysis. Or they damage the parathyroid glands, tiny yellow glands just behind the thyroid that control calcium levels in the body. When the parathyroids are damaged, as happened in 11 percent of patients in South Korea, patients get hypoparathyroidism, a difficult condition to treat.

In South Korea, some doctors, including Dr. Hyeong Sik Ahn of the College of Medicine at Korea University in Seoul, the first author of the new paper, have called for thyroid cancer screening to be banned. But their calls were mostly ignored, Dr. Ahn explained in an email. "Most thyroid doctors, especially surgeons, deny or minimize harms."

Thyroid experts in the United States are calling for restraint in diagnosing and treating tiny tumors. A few places, like Memorial Sloan-Kettering Cancer Center in Manhattan, offer patients with small tumors the option of simply waiting and having regular scans to see if the tumor grows. But few patients have joined the program.

"Once we have made a diagnosis of cancer it is difficult to say, 'Don't do anything,' " said Dr. Ashok R. Shaha, a thyroid cancer surgeon at Memorial Sloan-Kettering who is concerned about the zeal to diagnose and treat tiny tumors. Doctors as well as patients can be wary, he said. "In the U.S. we have a fear that if we miss a cancer the patient will sue."

Dr. R. Michael Tuttle, who runs the wait-and-see program at Memorial-Sloan Kettering, said the best way to encourage observation of very low-risk thyroid cancer instead of aggressive treatment was to "stop the diagnosis." That means, he said, "decrease screening and decrease F.N.A.," meaning fine needle aspiration, which is used to examine thyroid lumps noticed coincidentally.

And the lesson from South Korea should be heeded, said Dr. Barnett S. Kramer, director of the division of cancer prevention at the National Cancer Institute.

"The message for so long is that early detection is always good for you," he said. But this stark tale of screening gone wrong "should acutely raise awareness of the consequences of acting on the intuition that all screening must be of benefit and all diagnoses at an early stage are of benefit."

Before we leave the topic of medical screening completely, there are several additional issues having possible ethical and probabilistic implications that should at least be raised, if only briefly:[5]

[5] Besides miscarriages of justice that result from confusions involving probabilities, others

Premarital screening: From the early part of the 20th century, it has been standard practice for states to require a test for syphilis before a marriage license was issued. The rationale for this requirement was so the disease was not passed on to a newborn in the birth canal, with the typical result of blindness, or to an unaffected partner. Besides requiring a test for syphilis, many states in the late 1980s considered mandatory HIV evaluations before marriage licenses were issued. Illinois passed such a law in 1987 that took effect on January 1, 1988, and continued through August of 1989. It was a public health disaster. In the first six months after enactment, the number of marriage licenses issued in Illinois dropped by 22.5%; and of the 70,846 licenses issued during this period, only eight applicants tested positive with a cost of $312,000 per seropositive identified individual. Even for the eight

have suffered because of failures to clearly understand the fallibility of diagnostic testing. Probably the most famous example of this is the disappearance of Azaria Chamberlain, a nine-week-old Australian baby who disappeared on the night of August 17, 1980, while on a camping trip to Ayers Rock. The parents, Lindy and Michael Chamberlain, contended that Azaria had been taken from their tent by a dingo. After several inquests, some broadcast live on Australian television, Lindy Chamberlain was tried and convicted of murder, and sentenced to life imprisonment. A later chance finding of a piece of Azaria's clothing in an area with many dingo lairs, lead to Lindy Chamberlain's release from prison and eventual exoneration of all charges.

The conviction of Lindy Chamberlain for the alleged cutting of Azaria's throat in the front seat of the family car rested on evidence of fetal hemoglobin stains on the seat. Fetal hemoglobin is present in infants who are six months old or younger—Azaria Chamberlain was only nine weeks old when she disappeared. As it happens, the diagnostic test for fetal hemoglobin is very unreliable, and many other organic compounds can produce similar results, such as nose mucus and chocolate milkshakes, both of which were present in the vehicle (in other words, the specificity of the test was terrible). It was also shown that a "sound deadener" sprayed on the car during its production produced almost identical results for the fetal hemoglobin test.

The Chamberlain case was the most publicized in Australian history (and on a par with the O.J. Simpson trial in the United States). Because most of the evidence against Lindy Chamberlain was later rejected, it is a good illustration of how media hype and bias can distort a trial.

identified as positive, the number of false positives was unknown; the more definitive follow-up Western blot test was not available at that time. This particular episode was the most expensive public health initiative ever for Illinois; the understated conclusion from this experience is that mandatory premarital testing is not a cost-effective method for the control of human immunodeficiency virus infection. For a further discussion of the Illinois experience in mandatory HIV premarital testing, see Turnock and Kelly (1989).

Prenatal screening: The area of prenatal screening inevitably raises ethical issues. Some screening could be labeled quickly as unethical, for example, when selective abortions occur as the result of an ultrasound to determine the sex of a fetus. In other cases, the issues are murkier.[6] For instance, in screening for Down's syndrome because of a mother's age, acting solely on the use of noninvasive biomedical markers with poor selectivity and sensitivity values is questionable; the further screening with more invasive methods, such as amniocentesis, may be justifiable even when considering an accompanying one to two percent chance of the invasive test inducing a miscarriage. At least in the case of screening for Down's syndrome, these trade-offs between invasive screening and the risk of spontaneous miscarriage may no longer exist given a new noninvasive DNA blood test announced in the *British Medical Journal* in January 2011, "Noninvasive Prenatal Assessment of Trisomy 21 by Multiplexed Maternal

[6] There is also the fear that increasingly sophisticated prenatal genetic testing will enable people to engineer "designer babies," where parents screen for specific traits and not for birth defects per se. The question about perfection in babies being an entitlement is basically an ethical one; should otherwise healthy fetuses be aborted if they do not conform to parental wishes? To an extent, some of this selection is done indirectly and crudely already when choices are made from a sperm bank according to desired donor characteristics.

Plasma DNA Sequencing: Large Scale Validity Study." The article abstract follows:

Objectives: To validate the clinical efficacy and practical feasibility of massively parallel maternal plasma DNA sequencing to screen for fetal trisomy 21 among high risk pregnancies clinically indicated for amniocentesis or chorionic villus sampling.

Design: Diagnostic accuracy validated against full karyotyping, using prospectively collected or archived maternal plasma samples.

Setting: Prenatal diagnostic units in Hong Kong, United Kingdom, and the Netherlands.

Participants: 753 pregnant women at high risk for fetal trisomy 21 who underwent definitive diagnosis by full karyotyping, of whom 86 had a fetus with trisomy 21.

Intervention: Multiplexed massively parallel sequencing of DNA molecules in maternal plasma according to two protocols with different levels of sample throughput: 2-plex and 8-plex sequencing.

Main outcome measures: Proportion of DNA molecules that originated from chromosome 21. A trisomy 21 fetus was diagnosed when the z-score for the proportion of chromosome 21 DNA molecules was greater than 3. Diagnostic sensitivity, specificity, positive predictive value, and negative predictive value were calculated for trisomy 21 detection.

Results: Results were available from 753 pregnancies with the 8-plex sequencing protocol and from 314 pregnancies with the 2-plex protocol. The performance of the 2-plex protocol was superior to that of the 8-plex protocol. With the 2-plex protocol, trisomy 21 fetuses were detected at 100% sensitivity and 97.9% specificity, which resulted in a positive predictive value of 96.6% and negative predictive value of 100%. The 8-plex protocol detected 79.1% of the trisomy 21 fetuses and 98.9% specificity, giving a positive predictive value of 91.9% and negative predictive value of 96.9%.

Conclusion: Multiplexed maternal plasma DNA sequencing analysis could be used to rule out fetal trisomy 21 among high risk pregnancies. If referrals for amniocentesis or chorionic villus sampling were based on the sequencing test results, about 98% of the invasive diagnostic procedures could be avoided.

Costs of screening: All screening procedures have costs attached, if only for the laboratory fees associated with carrying out the diagnostic test. When implemented on a more widespread public health basis, however, screenings may soon become cost-prohibitive for the results obtained. The short-lived premarital HIV screening in Illinois is one example, but new diagnostic screening methods seem to be reported routinely in the medical literature. These then get picked up in the more popular media, possibly with some recommendation for further broad implementation. A societal reluctance to engage in such a process may soon elicit a label of "medical rationing" (possibly, with some further allusion to socialized medicine, or what one can expect under "Obama-care").[7]

One recent example of a hugely expensive but (mostly) futile screening effort is by the Transportation Security Administration (TSA) and its airport passenger screening program. We give excerpts from three reports that appeared in the *New York Times* in 2013 and 2014:

"Report Says TSA Screening Is Not Objective" (Michael S. Schmidt, June 4, 2013) –

> The Transportation Security Administration has little evidence that an airport passenger screening program, which some employees believe is a magnet for racial profiling and has cost taxpayers nearly one billion dollars, screens passengers objectively, according to a report by the inspector general for the Homeland Security Department.

[7] One possible mechanism that may be a viable strategy for keeping the cost of screenings under some control is through a clever use of statistics. Depending on what is being assessed (for example, in blood, soil, air), it may be possible to test a "pooled" sample; only when that sample turns out to be "positive" would the individual tests on each of the constituents need to be carried out.

The T.S.A.'s "behavioral detection program" is supposed to rely on security officers who pull aside passengers who exhibit what are considered telltale signs of terrorists for additional screening and questioning. It is illegal to screen passengers because of their nationality, race, ethnicity or religion.

According to the report, the T.S.A. has not assessed the effectiveness of the program, which has 2,800 employees and does not have a comprehensive training program. The T.S.A. cannot "show that the program is cost-effective, or reasonably justify the program's expansion," the report said.

As a result of the T.S.A.'s ineffective oversight of the program, it "cannot ensure that passengers at U.S. airports are screened objectively," the report said.

...

In August, *The Times* reported that more than 30 officers at Logan International Airport in Boston had said that the program was being used to profile passengers like Hispanics traveling to Florida or blacks wearing baseball caps backward.

The officers said that such passengers were being profiled by the officers in response to demands from managers who believed that stopping and questioning them would turn up drugs, outstanding arrest warrants or immigration problems.

The managers wanted to generate arrests so they could justify the program, the officers said, adding that officers who made arrests were more likely to be promoted. The Homeland Security Department said then that its inspector general was investigating the matter, although the coming report does not address the program at Logan Airport.

In a written statement, Representative Bennie Thompson, Democrat of Mississippi, the ranking member on the House Homeland Security Committee, said that the report "deals yet another blow to T.S.A.'s efforts to implement a behavioral detection screening program."

Mr. Thompson added that he would be offering an amendment to the Homeland Security appropriations bill this week that would "prevent any more taxpayer dollars from being spent on this failed and misguided effort."

"At Airports, A Misplaced Faith in Body Language" (John Tierney, March 23, 2012) –

Like the rest of us, airport security screeners like to think they can read body language. The Transportation Security Administration has spent some $1 billion training thousands of "behavior detection officers" to look for facial expressions and other nonverbal clues that would identify terrorists.

But critics say there's no evidence that these efforts have stopped a single terrorist or accomplished much beyond inconveniencing tens of thousands of passengers a year. The T.S.A. seems to have fallen for a classic form of self-deception: the belief that you can read liars' minds by watching their bodies.

Most people think liars give themselves away by averting their eyes or making nervous gestures, and many law-enforcement officers have been trained to look for specific tics, like gazing upward in a certain manner. But in scientific experiments, people do a lousy job of spotting liars. Law-enforcement officers and other presumed experts are not consistently better at it than ordinary people even though they're more confident in their abilities.

"Theres an illusion of insight that comes from looking at a person's body," says Nicholas Epley, a professor of behavioral science at the University of Chicago. "Body language speaks to us, but only in whispers."

...

"The common-sense notion that liars betray themselves through body language appears to be little more than a cultural fiction," says Maria Hartwig, a psychologist at John Jay College of Criminal Justice in New York City. Researchers have found that the best clues to deceit are verbal – liars tend to be less forthcoming and tell less compelling stories – but even these differences are usually too subtle to be discerned reliably.

One technique that has been taught to law-enforcement officers is to watch the upward eye movements of people as they talk. This is based on a theory from believers in "neuro-linguistic programming" that people tend to glance upward to their right when lying, and upward to the left when telling the truth.

But this theory didn't hold up when it was tested by a team of British and North American psychologists. They found no pattern in the upward eye movements of liars and truth tellers, whether they were observed in the laboratory or during real-life news conferences. The researchers also found that people who were trained to look for these eye movements did not do any better than a control group at detecting liars.

“Behavior Detection Isn’t Paying Off” (The Editorial Board, April 6, 2014) –

A multiyear experiment in behavior detection is only worsening the Transportation Security Administration’s reputation for wastefulness. Since 2007, the T.S.A. has trained officers to identify high-risk passengers on the basis of mostly nonverbal signs, like fidgeting or sweating, which may indicate stress or fear. The total price tag: nearly $1 billion.

In theory we’re all for the T.S.A. devoting resources to human intelligence, but this particular investment does not appear to be paying off.

As John Tierney wrote in *The Times* on March 25, the T.S.A. “seems to have fallen for a classic form of self-deception: the belief that you can read liars’ minds by watching their bodies.” He cited experiments showing that people are terrible at spotting liars. One survey of more than 200 studies found that “people correctly identified liars only 47 percent of the time, less than chance.”

The T.S.A.’s behavior-detection officers are no better. The Government Accountability Office told Congress in November that T.S.A. employees could not reliably single out dangerous passengers and that the program was ineffective.

In its review of 49 airports in 2011 and 2012, the G.A.O. calculated that behavior-detection officers designated passengers for additional screening on 61,000 occasions. From that group, 8,700, or 14 percent, were referred to law enforcement. Only 4 percent of the 8,700, or 0.6 percent of the total, were arrested – none for suspected terrorism. (The T.S.A. said the Federal Air Marshal Service earmarked certain cases for further investigation, but could not provide the G.A.O. with details.) The G.A.O. attributed these poor results to a general “absence of scientifically validated evidence” for training T.S.A. employees in the dark art of behavior detection, and urged Congress to limit future funding.

> The union representing T.S.A. officers has defended the program, which costs roughly $200 million a year, arguing that an "imperfect deterrent to terrorist attacks is better than no deterrent at all." But behavior detection is far from the country's only shield, and "imperfect" is an understatement. Congress should take the G.A.O.'s advice.

Besides initial screening costs and those involved in dealing with follow-up procedures for all the false positives identified, there may also be costs involved in the particular choice among alternatives for a diagnostic procedure. If one strategy has demonstrable advantages but increased costs over another, based on an evidence-based assessment it still may be cost-effective to choose the higher-priced alternative. But if the evidence does not document such an advantage, it would seem fiscally prudent in controlling the increasing societal health-care costs to not choose the more expensive option as the default, irrespective of what professional pressure groups may want and who would profit the most from the specific choices made. A case in point is the use of colonoscopy in preference to sigmoidoscopy. We quote from a short letter to the editor of the *New York Times* by John Abramson (February 22, 2011) entitled "The Price of Colonoscopy":

> Colon cancer screening with colonoscopy—viewing the entire colon—has almost completely replaced more limited sigmoidoscopy, which costs as little as one-tenth as much. Yet studies have repeatedly failed to show that colonoscopy reduces the risk of death from colon cancer more effectively than sigmoidoscopy.

A recent example of a breakthrough in medical screening for lung cancer that may end up being very cost-ineffective was reported in a *News of the Week* article by Eliot Marshall, appearing in *Sci-*

ence (2010), entitled "The Promise and Pitfalls of a Cancer Breakthrough." It reviews the results of a $250 million study sponsored by the National Cancer Institute (NCI) named the National Lung Screening Trial (NLST). The diagnostic test evaluated was a three-dimensional low-dose helical computed tomography (CT) scan of an individual's lung. Although Harold Varmus commented that he saw "a potential for saving many lives," others saw some of the possible downsides of widespread CT screening, including costs. For example, note the comments from the NCI Deputy Director, Douglas Lowy (we quote from the *Science* news item):[8]

[8] Continued from the main text:

In NLST (National Lung Screening Trial), about 25% of those screened with CT got a positive result requiring followup. Some researchers have seen higher rates. Radiologist Stephen Swensen of the Mayo Clinic in Rochester, Minnesota, says that a nonrandomized study he led in 2005 gave positive results for 69% of the screens. One difference between the Mayo and NLST studies, Swensen says, is that Mayo tracked nodules as small as 1 to 3 millimeters whereas NLST, which began in 2002, cut off positive findings below 4 mm.

One negative consequence of CT screening, Lowy said at the teleconference, is that it triggers follow-up scans, each of which increases radiation exposure. Even low-dose CT scans deliver a "significantly greater" exposure than conventional chest x-rays, said Lowy, noting that, "It remains to be determined how, or if, the radiation doses from screening ... may have increased the risks for cancer during the remaining lifetime" of those screened. Clinical followup may also include biopsy and surgery, Lowy said, "potentially risky procedures that can cause a host of complications."

G. Scott Gazelle, a radiologist and director of the Institute for Technology Assessment at Massachusetts General Hospital in Boston, has been analyzing the likely impacts of lung cancer screening for a decade. He agrees that people are going to demand it—and that "there are going to be a huge number of false positives." He was not surprised at NLST's finding of a lifesaving benefit of 20%. His group's prediction of mortality reduction through CT scans, based on "micromodeling" of actual cancers and data from previous studies, was 18% to 25%, right on target. But Gazelle says this analysis, now under review, still suggests that a national program of CT screening for lung cancer "would not be cost effective." Indeed, the costs seem likely to be three to four times those of breast cancer screening, with similar benefits.

Advocates of screening, in contrast, see the NLST results as vindicating a campaign to put advanced computer technology to work on lung cancer. The detailed images of early

> Lowy, also speaking at the teleconference, ticked off some "disadvantages" of CT screening. One is cost. The price of a scan, estimated at about $300 to $500 per screening, is the least of it. Big expenses ensue, Lowy said, from the high ratio of people who get positive test results but do not have lung cancer. Even if you focus strictly on those with the highest risk—this trial screened smokers and ex-smokers who had used a pack of cigarettes a day for 30 years—"20% to 50%" of the CT scans "will show abnormalities" according to recent studies, said Lowy. According to NCI, about 96% to 98% are false positives. (p. 900)

Besides controlling health-case expenditures by considering the cost-effectiveness of tests, there are other choices involved in who should get screened and at what age. In an article by Gina Kolata in the *New York Times* (April 11, 2011), "Screening Prostates at Any Age," a study is discussed that found men 80 to 85 years old are being screened (using the PSA test) as often as men 30 years younger. Both the American Cancer Society and the American Urological Society discourage screenings for men whose life expectancy is ten years or less; prostate cancer is typically so slow-growing that it would take that long for any benefits of screening to appear. In addition, the United States Preventative Services Task Force recommends that screening should stop at 75. Given the observations we made about prostate screening in the previous section and the OpEd article by Richard Ablin, it appears we have an instance, not

tumors in CT scans are "exquisite," says James Mulshine, vice president for research at Rush University Medical Center in Chicago, Illinois, and an adviser to the pro-screening advocacy group, the Lung Cancer Alliance in Washington, D.C. He thinks it should be straightforward to reduce the number of biopsies and surgeries resulting from false positives by monitoring small tumors for a time before intervening. There are 45 million smokers in the United States who might benefit from CT screening, says Mulshine. He asks: Do we provide it, or "Do we tell them, 'Tough luck'?"

of practicing "evidence-based medicine," but a more likely one of "(Medicare) greed-induced medicine."

Informed consent and screening: Before participation in a screening program, patients must give informed consent, with an emphasize on the word "informed." Thus, the various diagnostic properties of the test should be clearly communicated, possibly with the use of Gigerenzer's "natural frequencies"; the risk of "false positives" must be clearly understood, as well as the risks associated with any follow-up invasive procedures. All efforts must be made to avoid the type of cautionary tale reported in Gigerenzer et al. 2007: at a conference on AIDS held in 1987, the former senator from Florida, Lawton Childs, reported that of twenty-two (obviously misinformed about false positives) blood donors in Florida who had been notified they had tested HIV-positive, seven committed suicide.

To inform patients properly about screening risks and benefits, the medical professionals doing the informing must be knowledgeable themselves. Unfortunately, as pointed out in detail by Gigerenzer et al. 2007, there is now ample evidence that many in the medical sciences are profoundly confused. An excellent model for the type of informed dialogue that should be possible is given by John Lee in a short "sounding board" article in the *New England Journal of Medicine* (1993, *328*, 438–440), "Screening and Informed Consent." This particular article is concerned with mammograms for detecting breast cancer but the model can be easily extended to other diagnostic situations where informed consent is required. Finally, to show that the type of exemplar dialogue that Lee models is not now widespread, we refer the reader to an editorial by Gerd Gigerenzer in *Maturitas* (2010, *67*, 5–6) entitled "Women's Perception of the Benefit of Breast Cancer Screening." The gist of the evidence given in the editorial should be clear from its concluding two sentences: "Misleading women, whether intentionally or unintentionally, about the benefit of mammography screening is a serious issue. All of those in the business of informing women about screening should recall that medical systems are for patients, not the other way around" (p. 6).

The (social) pressure to screen: Irrespective of the evidence for the value for a diagnostic screen, there are usually strong social pressures for us to engage in this behavior. These urgings may comes from medical associations devoted to lobbying some topic, from private groups formed to advocate for some position, or from our own doctors and clinics not wishing to be sued for underdiagnosis. The decision to partake or not in some screening process, should depend on the data-driven evidence of its value, or on the other side, of the potential for harm. On the other hand, there are many instances where the evidence is present for the value of some ongoing screening procedure. One of the current authors (LH) takes several medications, all to control surrogate endpoints (or test levels), with the promise of keeping one in a reasonable healthy state. Eye drops are used to control eye pressure (and to forestall glaucoma); lisinopril and amlodipine to keep blood pressure under control (and prevent heart attacks); and a statin to keep cholesterol levels down (and again, to avoid heart problems).

In addition to contending with social pressures to screen wherever those pressures may come from, there is now what seems to be a never-ending stream of media reports about new screening devices to consider or updated guidelines to follow about who should be screened, when to screen, and how often. There is now, for example, the possibility of genomic scans for a variety of mutations that might increase the risk of breast or ovarian cancer, or of the use of newer three-dimensional and hopefully more sensitive mammography. For the later we give several paragraphs from a Denise Grady article from the *New York Times* (June, 24, 2014), entitled "3-D Mammography Test Appears to Improve Breast Cancer Detection Rate":

> Adding a newer test to digital mammograms can increase the detection rate for breast cancer and decrease nerve-racking false alarms, in which suspicious findings lead women to get extra scans that turn out normal, a study found.
>
> Millions of women will get the newer test, tomosynthesis, this year. The procedure is nearly identical to a routine mammogram, except that in mammography the machine is stationary, while in tomosynthesis it moves around the breast. Sometimes called 3-D mammography, the test takes many X-rays at different angles to create a three-dimensional image of the breast. It was approved in the United States in 2011.
>
> The verdict is still out on the long-term worth of this new technology. The new results are promising but not definitive, according to experts not associated with the study, published Tuesday in The Journal of the American Medical Association. Tomosynthesis has not been around long enough to determine whether it saves lives or misses tumors.
>
> Even so, more and more mammography centers are buying the equipment, which is far more costly than a standard mammography unit, and marketing the test to patients as a more sensitive and accurate type of screening. It has come on the scene at a time when the value of breast cancer screening and the rising costs of health care are increasingly debated.

A variety of medically-related agencies issue guidelines periodically that concern general health practice. Unfortunately, some of these may be conflicting depending on the agencies involved and who they represent. As a good controversial case in point, there is the ongoing debate about the wisdom of annual pelvic exams for women. An editorial given below from the *New York Times* (authored by "The Editorial Board"; July 2, 2014), and entitled "The Dispute Over Annual Pelvic Exams," illustrates well the type of confusion that might be present among "dueling" recommendations:

> Two major medical groups have taken opposing positions on whether healthy, low-risk women with no symptoms should have an annual pelvic exam. The American College of Physicians, the largest organization of physicians who practice internal medicine, strongly advised against the exams, which many women find distasteful or painful. The American College of Obstetricians and Gynecologists, the leading group of specialists providing health care for women, immediately reiterated its support for yearly pelvic exams for asymptomatic adult women.
>
> The exams at issue are not the Pap smears used to detect cervical cancers. Those are still recommended although there is disagreement on how often they should be done. The new dispute involves the "bimanual examination," in which a doctor inserts two gloved fingers into a woman's vagina and presses down on her abdomen with the other hand to check from both sides the shape and size of her uterus, ovaries and fallopian tubes. It also involves procedures that use a speculum to open the vagina for examination.
>
> Oddly enough, both professional groups agree there is no credible scientific evidence that the annual pelvic examinations save lives. They simply disagree over whether that lack of evidence matters much.
>
> The College of Physicians thinks it does. In a review of published scientific studies from 1946 through January 2014, it found no evidence that the pelvic exams provide any benefit in asymptomatic, nonpregnant adult women and significant evidence of harm, such as unnecessary surgeries, fear, anxiety and pain. The exams drive some women to avoid the doctors and can be traumatic for rape victims. The physicians organization estimated the annual cost of the exams at $2.6 billion. Unnecessary follow-up tests drive the cost even

higher.

By contrast, the gynecologists group argues that the "clinical experiences" of gynecologists, while not "evidence-based," demonstrate that annual pelvic exams are useful in detecting problems like incontinence and sexual dysfunction and in establishing a dialogue with patients about a wide range of health issues.

In recent years, medical groups and researchers have issued changing and sometimes conflicting recommendations on how often women should get a routine mammogram, how often to get pap smears, and now, whether to get an annual pelvic exam. Women will need to make their own judgments about procedures that many of them, and their doctors, may have used for years as a matter of standard practice.

The decision to institute or encourage widespread diagnostic screening should be based on evidence that shows effectiveness in relation to all the costs incurred. Part of the national discussion in the United States of evidence-based medical decision making is now taking place for the common screening targets of cervical, prostate, and breast cancer. Until recently it was considered an inappropriate question to ask whether it might be best if we didn't screen and identify a nonlethal cancer, and thus avoid debilitating and unnecessary treatment. A recent survey article by Gina Kolata makes these points well: "Considering When it Might Be Best Not to Know About Cancer" (*New York Times*, October 29, 2011). The United Kingdom is somewhat more advanced than the United States with respect to guidelines when screening programs should be implemented. The British National Health Service has issued useful "appraisal criteria" to guide the adoption of a screening program. The appendix to follow reproduces these criteria.

4.1 Appendix: U.K. National Screening Committee Programme Appraisal Criteria

Criteria for appraising the viability, effectiveness and appropriateness of a screening programme —

Ideally all the following criteria should be met before screening for a condition is initiated:

The Condition:

1. The condition should be an important health problem.

2. The epidemiology and natural history of the condition, including development from latent to declared disease, should be adequately understood and there should be a detectable risk factor, disease marker, latent period or early symptomatic stage.

3. All the cost-effective primary prevention interventions should have been implemented as far as practicable.

4. If the carriers of a mutation are identified as a result of screening, the natural history of people with this status should be understood, including the psychological implications.

The Test:

5. There should be a simple, safe, precise and validated screening test.

6. The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.

7. The test should be acceptable to the population.

8. There should be an agreed policy on the further diagnostic investigation of individuals with a positive test result and on the choices available to those individuals.

9. If the test is for mutations, the criteria used to select the subset of mutations to be covered by screening, if all possible mutations are not being tested, should be clearly set out.

The Treatment:

10. There should be an effective treatment or intervention for patients identified through early detection, with evidence of early treatment leading to better outcomes than late treatment.

11. There should be agreed evidence-based policies covering which individ-

uals should be offered treatment and the appropriate treatment to be offered.

12. Clinical management of the condition and patient outcomes should be optimised in all health care providers prior to participation in a screening programme.

The Screening Programme:

13. There should be evidence from high quality Randomised Controlled Trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an "informed choice" (e.g., Down's syndrome, cystic fibrosis carrier screening), there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.

14. There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/intervention) is clinically, socially and ethically acceptable to health professionals and the public.

15. The benefit from the screening programme should outweigh the physical and psychological harm (caused by the test, diagnostic procedures and treatment).

16. The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (i.e., value for money). Assessment against this criteria should have regard to evidence from cost benefit and/or cost effectiveness analyses and have regard to the effective use of available resources.

17. All other options for managing the condition should have been considered (e.g., improving treatment, providing other services), to ensure that no more cost effective intervention could be introduced or current interventions increased within the resources available.

18. There should be a plan for managing and monitoring the screening programme and an agreed set of quality assurance standards.

19. Adequate staffing and facilities for testing, diagnosis, treatment and programme management should be available prior to the commencement of the screening programme.

20. Evidence-based information, explaining the consequences of testing, investigation and treatment, should be made available to potential participants to assist them in making an informed choice.

21. Public pressure for widening the eligibility criteria for reducing the screening interval, and for increasing the sensitivity of the testing process, should be anticipated. Decisions about these parameters should be scientifically justifiable to the public.

22. If screening is for a mutation, the programme should be acceptable to people identified as carriers and to other family members.

References

[1] Begg, C. B. (1987). Bias in the assessment of diagnostic tests. *Statistics in Medicine*, *6*, 411–423.

[2] Bokhari, E., & Hubert, L. (2015). A new condition for assessing the clinical efficiency of a diagnostic test. *Psychological Assessment*, *xx*, xxx–xxx.

[3] Dawes, R. M. (1962). A note on base rates and psychometric efficiency. *Journal of Consulting Psychology*, *26*, 422–424.

[4] Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

[5] Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology*, *143*, 29–36.

[6] Meehl, P., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194–215.

[7] Ransohoff, D. F., & Feinstein, R. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, *299*, 926–930.

[8] Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.

[9] Turnock, B. J., & Kelly, C. J. (1989). Mandatory premarital testing for human immunodeficiency virus: The Illinois experience. *Journal of the American Medical Association*, *261*, 3415–3418.

Module 5: Probabilistic Reasoning in the Service of Gambling

All life is six to five against.
– Damon Runyon

Abstract: Probabilistic reasoning is applied to several topics in gambling. We begin with the Chevalier de Méré asking the mathematician Blaise Pascal in the early 17th century for help with his gambling interests. Pascal in a series of letters with another mathematician, Pierre de Fermat, laid out what was to be the foundations for a modern theory of probability. Some of this formalization is briefly reviewed; also, to give several numerical examples, the Pascal-Fermat framework is applied to the type of gambles the Chevelier engaged in. Several other gambling related topics are discussed at some length: spread betting, parimutuel betting, and the psychological considerations behind gambling studied by Tversky, Kahneman, and others concerned with the psychology of choice and decision making.

Contents

1 Betting, Gaming, and Risk

Antoine Gombaud, better known as the Chevalier de Méré, was a French writer and amateur mathematician from the early 17th century. He is important to the development of probability theory because of one specific thing; he asked a mathematician, Blaise Pascal, about a gambling problem dating from the Middle Ages, named "the problem of points." The question was one of fairly dividing the stakes among individuals who had agreed to play a certain number of games, but for whatever reason had to stop before they were finished. Pascal in a series of letters with Pierre de Fermat, solved this equitable division task, and in the process laid out the foundations for a modern theory of probability. Pascal and Fermat also provided the Chevalier with a solution to a vexing problem he was having in his own personal gambling. Apparently, the Chevalier had been very successful in making even money bets that a six would be rolled at least once in four throws of a single die. But when he tried a similar bet based on tossing two dice 24 times and looking for a double-six to occur, he was singularly unsuccessful in making any money. The reason for this difference between the Chevalier's two wagers was clarified by the formalization developed by Pascal and Fermat for such games of chance. This formalization is briefly reviewed below, and then used to discuss the Chevalier's two gambles as well as those occurring in various other casino-type games.

We begin by defining several useful concepts: a simple experiment, sample space, sample point, event, elementary event:

A *simple experiment* is some process that we engage in that leads to one single outcome from a set of possible outcomes that could occur. For example, a simple experiment could consist of rolling a single die once, where the set of possible outcomes is $\{1, 2, 3, 4, 5, 6\}$ (note that curly braces will be used consistently to denote a set). Or, two dice could be tossed and the number of spots occurring on each die noted; here, the possible outcomes are integer number pairs: $\{(a, b) \mid 1 \leq a \leq 6; 1 \leq b \leq 6\}$. Flipping a single coin would give the set of outcomes, $\{H, T\}$, with "H" for "heads" and "T" for "tails"; picking a card from a normal deck could give a set of outcomes containing 52 objects, or if we were only interested in the particular suit for a card chosen, the possible outcomes could be $\{H, D, C, S\}$, corresponding to heart, diamond, club, and spade, respectively.

The set of possible outcomes for a simple experiment is the *sample space* (which we denote by the script letter $\mathcal{S}$). An object in a sample space is a *sample point*. An *event* is defined as a subset of the sample space, and an event containing just a single sample point is an *elementary event*. A particular event is said to occur when the outcome of the simple experiment is a sample point belonging to the defining subset for that event.

As a simple example, consider the toss of a single die, where $\mathcal{S}$ $= \{1, 2, 3, 4, 5, 6\}$. The event of obtaining an even number is the subset $\{2, 4, 6\}$; the event of obtaining an odd number is $\{1, 3, 5\}$; the (elementary) event of tossing a 5 is a subset with a single sample point, $\{5\}$, and so on.

For a sample space containing K sample points, there are 2^K possible events (that is, there are 2^K possible subsets of the sample space). This includes the "impossible event" (usually denoted by

$\emptyset$), characterized as that subset of $\mathcal{S}$ containing no sample points and which therefore can never occur; and the "sure event," defined as that subset of $\mathcal{S}$ containing all sample points (that is, $\mathcal{S}$ itself), which therefore must always occur. In our single die example, there are $2^6 = 64$ possible events, including $\emptyset$ and $\mathcal{S}$.

The motivation for introducing the idea of a simple experiment and sundry concepts is to use this structure as an intuitively reasonable mechanism for assigning probabilities to the occurrence of events. These probabilities are usually assigned through an assumption that sample points are equally likely to occur, assuming we have characterized appropriately what is to be in $\mathcal{S}$. Generally, only the probabilities are needed for the K elementary events containing single sample points. The probability for any other event is merely the sum of the probabilities for all those elementary events defined by the sample points making up that particular event. This last fact is due to the disjoint set property of probability introduced in the first module. In the specific instance in which the sample points are equally likely to occur, the probability assigned to any event is merely the number of sample points defining the event divided by K. As special cases, we obtain a probability of 0 for the impossible event, and 1 for the sure event.

The use of the word *appropriately* in characterizing a sample space is important to keep in mind whenever we wish to use the idea of being equally likely to generate the probabilities for all the various events. For example, in throwing two dice and letting the sample space be $\mathcal{S} = \{(a, b) \mid 1 \leq a \leq 6; 1 \leq b \leq 6\}$, it makes sense, assuming that the dice are not "loaded," to consider the 36 integer number pairs to be equally likely. When the conception of what is being observed changes, however, the equally-likely notion may no longer be "appropriate." For example, suppose our interest is only in the sum of spots on the two dice being tossed, and let our sample space be $\mathcal{S} = \{2, 3, \ldots, 12\}$. The eleven integer sample points in this sample space are not equally likely; in fact, it is a common exercise in an elementary statistics course to derive the probability distribution for the objects in this latter sample space based on the idea that the underlying 36 integer number pairs are equally likely. To illustrate, suppose our interest is in the probability that a "sum of seven" appears on the dice. At the level of the sample space containing the 36 integer number pairs, a "sum of seven" corresponds to the event $\{(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)\}$. Thus, the probability of a "sum of seven" is $6/36$; there are six equally-likely sample points making up the event and there are 36 equally-likely integer pairs in the sample space. Although probably apocryphal, it has been said that many would-be probabilists hired by gambling patrons in the 17th century, came to grief when they believed that every stated sample space had objects that could be considered equally likely, and communicated this fact to their employers as an aid in betting.

One particularly helpful use of the sample space/event concepts is when a simple experiment is carried out multiple times (for, say, N replications), and the outcomes defining the sample space are the ordered N-tuples formed from the results obtained for the individual simple experiments. The Chevalier who rolls a single die four times, generates the sample space

$$\{(D_1, D_2, D_3, D_4) \mid 1 \leq D_i \leq 6, 1 \leq i \leq 4\} ,$$

that is, all 4-tuples containing the integers from 1 to 6. Generally, in a replicated simple experiment with K possible outcomes on each trial, the number of different N-tuples is K^N (using a well-known arithmetic multiplication rule). Thus, for the Chevalier example, there are $6^4 = 1296$ possible 4-tuples, and each such 4-tuple should be equally likely to occur (given the "fairness" of the die being used; so, no "loaded" dice are allowed). To define the event of "no sixes rolled in four replications," we would use the subset (event)

$$\{(D_1, D_2, D_3, D_4) \mid 1 \le D_i \le 5, 1 \le i \le 4\} \ ,$$

containing $5^4 = 625$ sample points. Thus, the probability of "no sixes rolled in four replications" is $625/1296 = .4822$. As we will see formally below, the fact that this latter probability is strictly less than 1/2 gives the Chevalier a distinct advantage in playing an even money game defined by his being able to roll at least one six in four tosses of a die.

The other game that was not as successful for the Chevalier, was tossing two dice 24 times and betting on obtaining a double-six somewhere in the sequence. The sample space here is

$$\{(P_1, P_2, \ldots, P_{24})\}, \text{ where } P_i = \{(a_i, b_i) \mid 1 \le a_i \le 6; 1 \le b_i \le 6\},$$

and has 36^{24} possible sample points. The event of "not obtaining a double-six somewhere in the sequence" would look like the sample space just defined except that the $(6, 6)$ pair would be excluded from each P_i. Thus, there are 35^{24} members in this event. The probability of "not obtaining a double-six somewhere in the sequence" is

$$\frac{35^{24}}{36^{24}} = (\frac{35}{36})^{24} = .5086 \ .$$

Because this latter value is greater than $1/2$ (in contrast to the previous gamble), the Chevalier would now be at a disadvantage making an even money bet.

The best way to evaluate the perils or benefits present in a wager is through the device of a discrete random variable. Suppose X denotes the outcome of some bet; and let $a_1, \ldots, a_T$ represent the T possible payoffs from one wager, where positive values reflect gain and negative values reflect loss. In addition, we know the probability distribution for X; that is, $P(X = a_t)$ for $1 \leq t \leq T$. What one expects to realize from one observation on X (or from one play of the game) is its expected value,

$$E(X) = \sum_{t=1}^{T} a_t P(X = a_t).$$

If $E(X)$ is negative, we would expect to lose this much on each bet; if positive, this is the expected gain on each bet. When $E(X)$ is 0, the term "fair game" is applied to the gamble, implying that one neither expects to win or lose anything on each trial; one expects to "break even." When $E(X) \neq 0$, the game is "unfair" but it could be unfair in your favor ($E(X) > 0$), or unfair against you ($E(X) < 0$).

To evaluate the Chevalier's two games, suppose X takes on the values of $+1$ and -1 (the winning or losing of one dollar, say). For the single die rolled four times, $E(X) = (+1)(.5178) + (-1)(.4822) = .0356 \approx .04$. Thus, the game is unfair in the Chevalier's favor because he expects to win a little less than four cents on each wager. For the 24 tosses of two dice, $E(X) = (+1)(.4914) + (-1)(.5086) = -.0172 \approx -.02$. Here, the Chevalier is at a disadvantage. The game is unfair against him, and he expects to lose about two cents on each play of the game.

Besides using the expectation of X as an indication of whether a game is fair or not, and in whose favor, the variance of X is an important additional characteristic of any gamble. The larger the variance, the more one would expect a "boom or bust" scenario to take over, with the possibility of wild swings in the sizes of the gains or losses. But if one cannot play a game having a large variance multiple times, then it doesn't make much difference if one has a slight positive favorable expectation. There is another story, probably again apocryphal, of a man with a suitcase of money who for whatever reason needed twice this amount or it really didn't matter if he lost it all. He goes into a casino and bets it all at once at a roulette table—on red. He either gets twice his money on this one play or loses it all; in the latter case as we noted, maybe it doesn't matter; for example, because he previously borrowed money, the mob will place a "hit" on him if he can't come up with twice the amount that he had to begin with. Or recently, consider the hugely successful negative bets that Goldman Sachs and related traders (such as John Paulson) made on the toxic derivatives they had themselves created (in the jargon, they held a "short position" where one expects the price to fall and to thereby make money in the process).

A quotation from the author of the 1995 novel *Casino*, Nicholas Pileggi, states the issue well for casinos and the usual games of chance where skill is irrelevant (for example, roulette, slots, craps, keno, lotto, or blackjack [without card counting]); all are unfair and in the house's favor:

A casino is a mathematics palace set up to separate players from their money. Every bet in a casino has been calibrated within a fraction of its life to maximize profit while still giving the players the illusion they have a chance.

The negative expectations may not be big in any absolute sense, but given the enormous number of plays made, and the convergent effects of the law of large numbers (to be discussed in a later chapter), casinos don't lose money, period. The next time an acquaintance brags about what a killing he or she made in the casino on a game involving no skill, you can just comment that the game must not have been played long enough.[1]

[1] We give two short anecdotes that may be helpful in motivating the material in this section:

Charles Marie de La Condamine (1701–1774) is best known for answering the question as to whether the earth was flat or round. He based his answer (which was "round") on extensive measurements taken at the equator in Ecuador and in Lapland. For our purposes, however, he will be best known for giving the French philosopher Voltaire a gambling tip that allowed him to win 500,000 francs in a lottery. Condamine noted to Voltaire that through a miscalculation, the sum of all the ticket prices for the lottery was far less than the prize. Voltaire bought all the tickets and won.

Joseph Jagger (1830–1892) is known as "the man who broke the bank at Monte Carlo." In reality, he was a British engineer working in the Yorkshire cotton manufacturing industry, and very knowledgeable about spindles that were "untrue." Jagger speculated that a roulette wheel did not necessarily "turn true," and the outcomes not purely random but biased toward particular outcomes. We quote a brief part of the Wikipedia entry on Joseph Jagger that tells the story:

Jagger was born in September 1829 in the village of Shelf near Halifax, Yorkshire. Jagger gained his practical experience of mechanics working in Yorkshire's cotton manufacturing industry. He extended his experience to the behaviour of a roulette wheel, speculating that its outcomes were not purely random sequences but that mechanical imbalances might result in biases toward particular outcomes.

In 1873, Jagger hired six clerks to clandestinely record the outcomes of the six roulette wheels at the Beaux-Arts Casino at Monte Carlo, Monaco. He discovered that one of the six wheels showed a clear bias, in that nine of the numbers (7, 8, 9, 17, 18, 19, 22, 28 and 29) occurred more frequently than the others. He therefore placed his first bets on 7 July 1875

1.1 Spread Betting

The type of wagering that occurs in roulette or craps is often referred to as fixed-odds betting; you know your chances of winning when you place your bet. A different type of wager is spread betting, invented by a mathematics teacher from Connecticut, Charles McNeil, who became a Chicago bookmaker in the 1940s. Here, a payoff is based on the wager's accuracy; it is no longer a simple "win or lose" situation. Generally, a spread is a range of outcomes, and the bet itself is on whether the outcome will be above or below the spread. In common sports betting (for example, NCAA college basketball), a "point spread" for some contest is typically advertised by a bookmaker. If the gambler chooses to bet on the "underdog," he is said to "take the points" and will win if the underdog's score plus the point spread is greater than that of the favored team; conversely, if the gambler bets on the favorite, he "gives the points" and wins only if the favorite's score minus the point spread is greater than the underdog's score. In general, the announcement of a point spread is an attempt to even out the market for the bookmaker, and to generate an equal amount of money bet on each side. The commission that a bookmaker charges will ensure a livelihood, and thus, the bookmaker

and quickly won a considerable amount of money, £14,000 (equivalent to around 50 times that amount in 2005, or £700,000, adjusted for inflation). Over the next three days, Jagger amassed £60,000 in earnings with other gamblers in tow emulating his bets. In response, the casino rearranged the wheels, which threw Jagger into confusion. After a losing streak, Jagger finally recalled that a scratch he noted on the biased wheel wasn't present. Looking for this telltale mark, Jagger was able to locate his preferred wheel and resumed winning. Counterattacking again, the casino moved the frets, metal dividers between numbers, around daily. Over the next two days Jagger lost and gave up, but he took his remaining earnings, two million francs, then about £65,000 (around £3,250,000 in 2005), and left Monte Carlo never to return.

can be unconcerned about the actual outcome.

Several of the more notorious sports scandals in United States history have involved a practice of "point shaving," where the perpetrators of such a scheme try to prevent a favored team from "covering" a published point spread. This usually involves a sports gambler and one or more players on the favored team. They are compensated when their team fails to "cover the spread"; and those individuals who have bet on the underdog, win. Two famous examples of this practice in college basketball are the Boston College point shaving scandal of 1978/9, engineered by the gangsters Henry Hill and Jimmy Burke, and the CCNY scandal of 1950/1 involving organized crime and 33 players from some seven schools (CCNY, Manhattan College, NYU, Long Island University, Bradley University (Peoria), University of Kentucky, and the University of Toledo). More recently, there is the related 2007 NBA betting scandal surrounding a referee, Tim Donaghy.[2]

[2] When this section on point shaving was being written in June of 2014, an obituary for Gene Melchiorre appeared in the *New York Times* (June 26, 2014), with the title "For Gene Melchiorre, a Regretful Turn Brought a Unique N.B.A. Distinction." Several paragraphs are given below that shed some personal light on the point-shaving scandal of 1951 mentioned in the text:

> At the dead end of a private, wooded road about 20 miles north of Chicago sits a two-story house belonging to Gene Melchiorre, a short, pigeon-toed grandfather of 15 known by his many friends as Squeaky. Family photos decorate his office, but one artifact is unlike the others: a 63-year-old comic book drawing of a giant, youthful Melchiorre wearing a No. 23 basketball jersey, a superhero in short shorts.
>
> Melchiorre, 86, a former two-time all-American at Bradley once called the "greatest little man in basketball," was the first overall pick in the 1951 N.B.A. draft. But he holds an unusual distinction: He is the only No. 1 pick in N.B.A. history to never play in the league.
>
> There have been plenty of top draft picks who have flamed out, sometimes in spectacular fashion. But there has never been a draft pick like Squeaky Melchiorre. After being chosen first by the Baltimore Bullets, Melchiorre was barred for life from the N.B.A. for his role in

In an attempt to identify widespread corruption in college basketball, Justin Wolfers investigated the apparent tendency for favored NCAA teams nationally not to “cover the spread.” His article in the *American Economic Review* (2006, *96*, 279–283) is provoca-

the point-shaving scandal of 1951. He and more than 30 other players from seven universities were arrested in the scandal.

The trouble began in 1949, while Melchiorre’s team was in New York for the National Invitation Tournament. A gambler from Brooklyn named Nick Englisis (widely known as Nick the Greek) intentionally “bumped into” a player inside the team’s hotel, according to an account Melchiorre gave to Look Magazine in 1953. Soon, Melchiorre and two teammates were in a room with three gamblers, who “told us the colleges were getting rich on basketball and we ought to be getting something for it.”

The conversation changed Melchiorre’s life dramatically. He could have been an N.B.A. legend – “Melchiorre knows every trick that can shake a man loose,” Kentucky Coach Adolph Rupp declared in 1951. But that never happened.

...

When asked about the scandal today, Melchiorre falls silent, then changes the subject. But in a 1953 article in Look titled “How I Fell for the Basketball Bribers,” Melchiorre described his downfall.

The gamblers he met in the hotel room told him that point-shaving was widespread and had been going on for years. Players were using the money to start businesses after graduation. “It’s not as if you’re throwing a game,” a gambler said. “All you have to do is win by more points or fewer points than the bookmakers think you’re supposed to.”

They assured the players there was no chance of getting caught.

Melchiorre admitted in the article to accepting money during his career. But he denied ever altering his play to manipulate the point spread.

“Why did we do it?” Melchiorre said in the 1953 article. “Well, none of us had any money. We justified ourselves, I guess, by saying the colleges were making plenty out of us. We argued to ourselves that what we were doing was wrong, but not too wrong, because we weren’t going to throw any games.”

A Suspended Sentence

In February and March 1951, the Manhattan district attorney’s office arrested several players from City College and Long Island University on bribery charges. In July, Melchiorre and several other Bradley players were arrested.

Melchiorre eventually pleaded guilty to a misdemeanor and received a suspended sentence. The scandal ended the careers of two N.B.A. All-Stars and the nation’s leading scorer, Sherman White, who served nine months on Rikers Island. As for Melchiorre, the N.B.A. barred him for life.

tively entitled “Point Shaving: Corruption in NCAA Basketball.” We quote the discussion section of this article to give a sense of what Wolfers claims he found in the data:

> These data suggest that point shaving *may* be quite widespread, with an indicative, albeit rough, estimate suggesting that around 6 percent of strong favorites have been willing to manipulate their performance. Given that around one-fifth of all games involve a team favored to win by at least 12 points, this suggests that around 1 percent of all games (or nearly 500 games through my 16-year sample) involve gambling related corruption. This estimate derives from analyzing the extent to which observed patterns in the data are consistent with the incentives for corruption derived from spread betting; other forms of manipulation may not leave this particular set of footprints in the data, and so this is a lower bound estimate of the extent of corruption. Equally, the economic model suggests a range of other testable implications, which are the focus of ongoing research.
>
> My estimate of rates of corruption receives some rough corroboration in anonymous self-reports. Eight of 388 Men’s Division I basketball players surveyed by the NCAA reported either having taken money for playing poorly or having knowledge of teammates who had done so.
>
> A shortcoming of the economic approach to identifying corruption is that it relies on recognizing systematic patterns emerging over large samples, making it difficult to pinpoint specific culprits. Indeed, while the discussion so far has proceeded as if point shaving reflected a conspiracy between players and gamblers, these results might equally reflect selective manipulation by coaches of playing time for star players. Further, there need not be any shadowy gamblers offering bribes, as the players can presumably place bets themselves, rendering a coconspirator an unnecessary added expense.
>
> The advantage of the economic approach is that it yields a clear understanding of the incentives driving corrupt behavior, allowing policy conclusions that extend beyond the usual platitudes that “increased education, prevention, and awareness programs” are required. The key incentive driving point shaving is that bet pay-offs are discontinuous at a point—the spread—that is (or should be) essentially irrelevant to the players. Were gamblers

> restricted to bets for which the pay-off was a linear function of the winning margin, their incentive to offer bribes would be sharply reduced. Similarly, restricting wagers to betting on which team wins the game sharply reduces the incentive of basketball players to accept any such bribes. This conclusion largely repeats a finding that is now quite well understood in the labor literature and extends across a range of contexts—that highly nonlinear pay-off structures can yield rather perverse incentives and, hence, undesirable behaviors. (p. 283)

Another more recent article on this same topic is by Dan Bernhardt and Steven Heston (*Economic Inquiry*, 2010, *48*, 14–25) entitled "Point Shaving in College Basketball: A Cautionary Tale for Forensic Economics." As this title might suggest, an alarmist position about the rampant corruption present in NCAA basketball is not justified. An alternative explanation for the manifest "point shaving" is the use of strategic end-game efforts by a basketball team trying to maximize its probability of winning (for example, when a favored team is ahead late in the game, the play may move from a pure scoring emphasis to one that looks to "wind down the clock"). The first paragraph of the conclusion section of the Bernhardt and Heston article follows:

> Economists must often resort to indirect methods and inference to uncover the level of illegal activity in the economy. Methodologically, our article highlights the care with which one must design indirect methods in order to distinguish legal from illegal behavior. We first show how a widely reported interpretation of the patterns in winning margins in college basketball can lead a researcher to conclude erroneously that there is an epidemic of gambling-related corruption. We uncover decisive evidence that this conclusion is misplaced and that the patterns in winning margins are driven by factors intrinsic to the game of basketball itself. (p. 24)

The use of spreads in betting has moved somewhat dramatically to the world financial markets, particularly in the United Kingdom.

We suggest the reader view an article from the *Times (London)* (April 10, 2009) by David Budworth entitled "Spread-Betting Fails Investors in Trouble." Even though it emphasizes what is occurring in the United Kingdom, it still provides a cautionary tale for the United States as well. The moral might be that just because someone can create something to bet on (think CDOs [Collateralized Debt Obligations] and Goldman Sachs) doesn't mean that it is necessarily a good idea to do so.

1.2 Parimutuel Betting

The term *parimutuel betting* (based on the French for "mutual betting") characterizes the type of wagering system used in horse racing, dog tracks, jai alai, and similar contests where the participants end up in a rank order. It was devised in 1867 by Joseph Oller, a Catalan impresario (he was also a bookmaker and founder of the Paris Moulin Rouge in 1889). Very simply, all bets of a particular type are first pooled together; the house then takes its commission and the taxes it has to pay from this aggregate; finally, the payoff odds are calculated by sharing the residual pool among the winning bets. To explain using some notation, suppose there are T contestants and bets are made of $W_1, W_2, \ldots, W_T$ on an outright "win." The total pool is $T_{pool} = \sum_{t=1}^{T} W_t$. If the commission and tax rate is a proportion, R, the residual pool, R_{pool}, to be allocated among the winning bettors is $R_{pool} = T_{pool}(1 - R)$. If the winner is denoted by $t*$, and the money bet on the winner is W_{t*}, the payoff per dollar for a successful bet is R_{pool}/W_{t*}. We refer to the odds on outcome $t*$ as

$$(\frac{R_{pool}}{W_{t*}} - 1) \text{ to } 1 \ .$$

For example, if $\frac{R_{pool}}{W_{t*}}$ had a value of 9.0, the odds would be 8 to 1: you get 8 dollars back for every dollar bet plus the original dollar.

Because of the extensive calculations involved in a parimutuel system, a specialized mechanical calculating machine, named a totalizator, was invented by the mechanical engineer George Julius, and first installed at Ellerslie Race Track in New Zealand in 1913. In the 1930s, totalizators were installed at many of the race tracks in the United States (for example, Hialeah Park in Florida and Arlington Race Track and Sportsman's Park in Illinois). All totalizators came with "tote" boards giving the running payoffs for each horse based on the money bet up to a given time. After the pools for the various categories of bets were closed, the final payoffs (and odds) were then determined for all winning bets.

In comparison with casino gambling, parimutuel betting pits one gambler against other gamblers, and not against the house. Also, the odds are not fixed but calculated only after the betting pools have closed (thus, odds cannot be turned into real probabilities legitimately; they are empirically generated based on the amounts of money bet). A skilled horse player (or "handicapper") can make a steady income, particularly in the newer Internet "rebate" shops that return to the bettor some percentage of every bet made. Because of lower overhead, these latter Internet gaming concerns can reduce their "take" considerably (from, say, 15% to 2%), making a good handicapper an even better living than before.

1.3 Psychological Considerations in Gambling

As shown in the work of Tversky and Kahneman (for example, Tversky & Kahneman, 1981), the psychology of choice is dictated to a great extent by the framing of a decision problem; that is, the context into which a particular decision problem is placed. The power of framing in how decision situations are assessed, can be illustrated well though an example and the associated discussion provided by Tversky and Kahneman (1981, p. 453):

Problem 1 [$N = 152$]: Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
If Program A is adopted, 200 people will be saved. [72 percent]
If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved. [28 percent]
Which of the two programs would you favor?
The majority choice in this problem is risk averse: the prospect of certainly saving 200 lives is more attractive than a risky prospect of equal expected value, that is, a one-in-three chance of saving 600 lives.

A second group of respondents was given the cover story of problem 1 with a different formulation of the alternative programs, as follows:

Problem 2 [$N = 155$]:
If Program C is adopted, 400 people will die. [22 percent]
If Program D is adopted, there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. [78 percent]

Which of the two programs would you favor?

The majority choice in problem 2 is risk taking: the certain death of 400 people is less acceptable than the two-in-three chance that 600 will die. The preferences in problems 1 and 2 illustrate a common pattern: choices involving gains are often risk averse and choices involving losses are often risk

> taking. However, it is easy to see that the two problems are effectively identical. The only difference between them is that the outcomes are described in problem 1 by the number of lives saved and in problem 2 by the number of lives lost. The change is accompanied by a pronounced shift from risk aversion to risk taking. (p. 453)

The effects of framing can be very subtle when certain conscious or unconscious (coded) words are used to provide a salient context that influences decision processes. A recent demonstration of this in the framework of our ongoing climate-change debate is given by Hardisty, Johnson, and Weber (2010) in *Psychological Science*. The article has the interesting title, "A Dirty Word or a Dirty World? Attribute Framing, Political Affiliation, and Query Theory." The abstract follows:

> We explored the effect of attribute framing on choice, labeling charges for environmental costs as either an earmarked tax or an offset. Eight hundred ninety-eight Americans chose between otherwise identical products or services, where one option included a surcharge for emitted carbon dioxide. The cost framing changed preferences for self-identified Republicans and Independents, but did not affect Democrats' preferences. We explain this interaction by means of query theory and show that attribute framing can change the order in which internal queries supporting one or another option are posed. The effect of attribute labeling on query order is shown to depend on the representations of either taxes or offsets held by people with different political affiliations. (p. 86)

Besides emphasizing the importance of framing in making decisions, Tversky and Kahneman developed a theory of decision making, called prospect theory, to model peoples' real-life choices, which are not necessarily the optimal ones (Kahneman & Tversky, 1979). Prospect theory describes decisions between risky alternatives with uncertain outcomes when the probabilities are generally known. One particular phenomenon discussed at length in prospect theory is loss aversion, or the tendency to strongly avoid loss as opposed to acquiring gains. In turn, loss aversion leads to risk aversion, or the reluctance of people to choose gambles with an uncertain payoff rather than another with a more certain but possibly lower expected payoff. For example, an investor who is risk averse might choose to put money into a fixed-interest bank account or a certificate-of-deposit rather than into some stock with the potential of high returns but also with a chance of becoming worthless.

The notion of risk aversion has been around since antiquity. Consider the legend of Scylla and Charybdis, two sea monsters of Greek mythology situated on opposite sides of the Strait of Messina in Italy, between Calabria and Sicily. They were placed close enough to each other that they posed an inescapable threat to passing ships, so avoiding Scylla meant passing too close to Charybdis and conversely. In Homer's *Odyssey*, Odysseus is advised by Circe to follow the risk-adverse strategy of sailing closer to Scylla and losing a few men rather than sailing closer to the whirlpools created by Charybdis that could sink his ship. Odysseus sailed successfully past Scylla and Charybdis, losing six sailors to Scylla —

> they writhed
> gasping as Scylla swung them up her cliff and there
> at her cavern's mouth she bolted them down raw —
> screaming out, flinging their arms toward me,
> lost in that mortal struggle.

The phrase of being "between a rock and a hard place" is a more modern version of being "between Scylla and Charybdis."

The most relevant aspect of any decision-making proposition involving risky alternatives is the information one has, both on the probabilities that might be associated with the gambles and what the payoffs might be. In the 1987 movie, *Wall Street*, the character playing Gordon Gekko states: "The most valuable commodity I know of is information." The value that information has is reflected in a great many ways: by laws against "insider trading" (think Martha Stewart); the mandatory injury reports and the not-likely-to-play announcements by the sports leagues before games are played; the importance of counting cards in blackjack to obtain some idea of the number of high cards remaining in the deck (and to make blackjack an unfair game in your favor); massive speed-trading on Wall Street designed to obtain a slight edge in terms of what the market is doing currently (and to thereby "beat out" one's competitors with this questionably obtained edge); the importance of correct assessments by the credit rating agencies (think of all the triple-A assessments for the Goldman Sachs toxic collateralized debt obligations and what that meant to the buyers of these synthetic financial instruments); and finally, in the case against Goldman Sachs, the bank supposedly knew about the toxicity of what it sold to their clients and then made a huge profit betting against what they sold (the proverbial "short position"). A movie quotation from *Dirty Harry* illustrates the crucial importance of who has information and who doesn't – "I know what you're thinkin'. 'Did he fire six shots or only five?' Well, to tell you the truth, in all this excitement I kind of lost track myself." At the end of this Harry Callahan statement to the bank robber as to whether he felt lucky, the bank robber says: "I gots to know!" Harry puts the .44 Magnum to the robber's head and pulls the trigger; Harry knew that he had fired six shots and not five.

The availability of good information is critical in all the decisions we make under uncertainty and risk, both financially and in terms of our health. When buying insurance, for example, we knowingly engage in loss-adverse behavior. The information we have on the possible downside of not having insurance usually outweighs any consideration that insurance companies have an unfair game going in their favor. When deciding to take new drugs or undergo various medical procedures, information is again crucial in weighing risks and possible benefits—ask your doctor if he or she has some information that is right for you—and coming to a decision that is "best" for us (consider, for example, the previous discussion about undergoing screenings for various kinds of cancer).

At the same time that we value good information, it is important to recognize when available "information" really isn't of much value and might actually be counterproductive, for example, when we act because of what is most likely just randomness or "noise" in a system. An article by Jeff Sommer in the *New York Times* (March 13, 2010) has the intriguing title, "How Men's Overconfidence Hurts Them as Investors." Apparently, men are generally more prone to act (trade) on short-term financial news that is often only meaningless "noise." Men are also more confident in their abilities to make good decisions, and are more likely to make many more high-risk gambles.

For many decades, the financial markets have relied on rating agencies, such as Moody's, Standard & Poor's, and Fitch, to provide impeccable information to guide wise investing, and for assessing realistically the risk being incurred. We are now learning that we can no longer be secure in the data the rating agencies produce. Because rating agencies have made public the computer programs and algorithms they use, banks have learned how to "reverse-engineer" the process to see how the top ratings might be obtained (or better, scammed). In the Goldman Sachs case, for example, the firm profited from the misery it helped create through the inappropriate high ratings given to its toxic CDOs. As Carl Levin noted as Chair of the Senate Permanent Subcommittee on Investigations: "A conveyor belt of high-risk securities, backed by toxic mortgages, got AAA ratings that turned out not to be worth the paper they were printed on." The rating agencies have been in the position of the "fox guarding the hen house." The reader is referred to an informative editorial that appeared in the *New York Times* ("What About the Raters?", May 1, 2010) dealing with rating agencies and the information they provide.

By itself, the notion of "insurance" is psychologically interesting; the person buying insurance is willingly giving away a specific amount of money to avoid a more catastrophic event that might happen even though the probability of it occurring might be very small. Thus, we have a bookie "laying off" bets made with him or her to some third party; a blackjack player buying insurance on the dealer having a "blackjack" when the dealer has an ace showing (it is generally a bad idea for a player to buy insurance); or individuals purchasing catastrophic health insurance but paying the smaller day-to-day medical costs themselves. Competing forces are always at work between the insurer and the insured. The insurer wishes his "pool" to be as large as possible (so the central limit theorem discussed later can operate), and relatively “safe”; thus, the push to exclude high-risk individuals is the norm, and insuring someone with pre-existing conditions is always problematic. The insured, on the other hand, wants to give away the least money to buy the wanted protection. As one final item to keep in mind, we should remember that insurance needs to be purchased before and not after the catastrophic event occurs. In late 2010, there was a national cable news story about the person whose house burned down as the county firetrucks stood by. The person felt very put upon and did not understand why they just let his house burn down; he had offered to pay the $75 fire protection fee (but only after the house stated to burn). The cable news agencies declared a “duty to rescue,” and the failure of the fire trucks to act was “manifestly immoral.” Well, we doubt it because no life was lost, only the property, and all because of a failure to pay the small insurance premium “up front.” For a discussion of this incident, see the article by Robert Mackey, “Tennessee Firefighters Watch Home Burn” (*New York Times*, October 6, 2010)

A second aspect of insurance purchase with psychological interest is how to estimate the probability of some catastrophic event. Insurers commonly have a database giving an estimated value over those individuals they may consider insuring. This is where the actuaries and statisticians make their worth known; how much should the insurance companies charge for a policy so the company would continue to make money. The person to be insured has no easy access to any comparable database and merely guesses a value or more usually, acts on some vague “gut feeling” as to what one should be willing to pay to avoid the catastrophic downside. The person being insured has no personal relative frequency estimate on which to rely.

Assessing risks when no database is available to an insuring body is more problematic. If every one were honest about these situations, it might be labeled as subjectively obtained, or more straightforwardly, a "guess." This may be "gussied up" slightly with the phrase "engineering judgment," but at its basis it is still a guess. Richard Feynman, in his role on the Rogers Commission investigating the Challenger accident of 1986, commented that "engineering judgment" was making up numbers according to the hallowed tradition of the "dry lab." Here, one makes up data as opposed to observation and experimentation. You work backwards to the beginning from the result you want to obtain at the end. For shuttle risk, the management started with a level of risk that was acceptable and worked backwards until they got the probability estimate that gave this final "acceptable" risk level.

References

[1] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 203–291.

[2] Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Module 6: Probabilistic Reasoning Through the Basic Sampling Model

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'Law of Frequency of Error.' The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

– Sir Francis Galton (*Natural Inheritance*, 1889)

Abstract: One mechanism for assisting in various tasks encountered in probabilistic reasoning is to adopt a simple sampling model. A population of interest is first posited, characterized by some random variable, say X. This random variable has a population distribution (often assumed to be normal), characterized by (unknown) parameters. The sampling model posits n independent observations on X, denoted by $X_1, \ldots, X_n$, and which constitutes the sample. Various functions of the sample can then be constructed (that is, various statistics can be computed such as the sample mean and sample variance); in turn, statistics have their own sampling distributions. The general problem of statistical inference is to ask what sample statistics tell us about their population counterparts; for example, how can we construct a confidence interval for a population parameter such as the population mean from the sampling distribution for the sample mean.

Under the framework of a basic sampling model, a number of topics are discussed: confidence interval construction for a population mean where the length of the interval is determined by the square root of the sample size; the Central Limit Theorem and the Law of Large Numbers; the influence that sample size and variability have on our probabilistic reasoning skills; the massive fraud case involving the Dutch social psychologist, Diederik Stapel, and the role that lack of variability played in his exposure; the ubiquitous phenomenon of regression toward the mean and the importance it has for many of our probabilistic misunderstandings; how reliability corrections can be incorporated into prediction; the dichotomy and controversy encountered every ten years about complete enumeration versus sampling (to correct for, say, an undercount) in the United States Census.

Contents

1 The Basic Sampling Model and Associated Topics

We begin by refreshing our memories about the distinctions between *population* and *sample*, *parameters* and *statistics*, and *population distributions* and *sampling distributions*. Someone who has successfully completed a first course in statistics should know these distinctions well. Here, only a simple univariate framework is considered explicitly, but an obvious and straightforward generalization exists for the multivariate context as well.

A *population* of interest is posited, and operationalized by some random variable, say X. In this *Theory World* framework, X is characterized by *parameters*, such as the expectation of X, $\mu = \mathrm{E}(X)$, or its variance, $\sigma^2 = \mathrm{V}(X)$. The random variable X has a *(population) distribution*, which is often assumed normal. A *sample* is generated by taking observations on X, say, $X_1, \ldots, X_n$, considered independent and identically distributed as X; that is, they are exact copies of X. In this *Data World* context, statistics are functions of the sample and therefore characterize the sample: the sample mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$; the sample variance, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$, with some possible variation in dividing by $n - 1$ to generate an unbiased estimator for σ^2. The statistics, $\hat{\mu}$ and $\hat{\sigma}^2$, are *point estimators* of μ and σ^2. They are random variables by themselves, so they have distributions referred to as *sampling distributions*. The general problem of statistical inference is to ask what sample statistics, such as $\hat{\mu}$ and $\hat{\sigma}^2$, tell us about their population counterparts,

μ and σ^2. In other words, can we obtain a measure of accuracy for estimation from the sampling distributions through, for example, confidence intervals?

Assuming that the population distribution is normally distributed, the sampling distribution of $\hat{\mu}$ is itself normal with expectation μ and variance σ^2/n. Based on this result, an approximate 95% confidence interval for the unknown parameter μ can be given by

$$\hat{\mu} \ \pm \ 2.0\frac{\hat{\sigma}}{\sqrt{n}} \ .$$

Note that it is the square root of the sample size that determines the length of the interval (and not the sample size per se). This is both good news and bad. Bad, because if you want to double precision, you need a fourfold increase in sample size; good, because sample size can be cut by four with only a halving of precision.

Even when the population distribution is not originally normally distributed, the central limit theorem (CLT) (that is, the "Law of Frequency of Error," as noted by the opening epigram for this module) says that $\hat{\mu}$ is approximately normal in form and becomes exactly so as n goes to infinity. Thus, the approximate confidence interval statement remains valid even when the underlying distribution is not normal. Such a result is the basis for many claims of robustness; that is, when a procedure remains valid even if the assumptions under which it was derived may not be true, as long as some particular condition is satisfied; here, the condition is that the sample size be reasonably large.

Besides the robustness of the confidence interval calculations for μ, the CLT also encompasses the law of large numbers (LLN). As the sample size increases, the estimator, $\hat{\mu}$, gets closer to μ, and converges to μ at the limit as n goes to infinity. This is seen most directly in the variance of the sampling distribution for $\hat{\mu}$, which becomes smaller as the sample size gets larger.

The basic results obtainable from the CLT and LLN that averages are both less variable and more normal in distribution than individual observations, and that averages based on larger sample sizes will show less variability than those based on smaller sample sizes, have far-ranging and sometimes subtle influences on our probabilistic reasoning skills. For example, suppose we would like to study organizations, such as schools, health care units, or governmental agencies, and have a measure of performance for the individuals in the units, and the average for each unit. To identify those units exhibiting best performance (or, in the current jargon, "best practice"), the top 10%, say, of units in terms of performance are identified; a determination is then made of what common factors might characterize these top-performing units. We are pleased when we are able to isolate one very salient feature that most units in this top tier are small. We proceed on this observation and advise the breaking up of larger units. Is such a policy really justified based on these data? Probably not, if one also observes that the bottom 10% are also small units. That smaller entities tend to be more variable than the larger entities seems to vitiate a recommendation of breaking up the larger units for performance improvement. Evidence that the now-defunct "small schools movement," funded heavily by the Gates Foundation, was a victim of the "square root of n law" was presented by Wainer (2009, pp. 11–14).

Sports is an area in which there is a great misunderstanding and lack of appreciation for the effects of randomness. A reasonable model for sports performance is one of "observed performance" being the sum of "intrinsic ability" (or true performance) and "error," leading to a natural variability in outcome either at the individual or the team level. Somehow it appears necessary for sports writers, announcers, and other pundits to give reasons for what is most likely just random variability. We hear of team "chemistry," good or bad, being present or not; individuals having a "hot hand" (or a "cold hand," for that matter); someone needing to "pull out of a slump"; why there might be many .400 hitters early in the season but not later; a player being "due" for a hit; free-throw failure because of "pressure"; and so on. Making decisions based on natural variation being somehow "predictive" or "descriptive" of the truth, is not very smart, to say the least. But it is done all the time—sports managers are fired and CEOs replaced for what may be just the traces of natural variability.

People who are asked to generate random sequences of numbers tend to underestimate the amount of variation that should be present; for example, there are not enough longer runs and a tendency to produce too many short alternations. In a similar way, we do not see the naturalness in regression toward the mean (discussed in the next section of this module), where extremes are followed by less extreme observations just because of fallibility in observed performance. Again, causes are sought. We hear about multi-round golf tournaments where a good performance on the first day is followed by a less adequate score the second (due probably to "pressure"); or a bad performance on the first day followed by an improved performance the next (the golfer must have been able to "play loose"). Or in baseball, at the start of a season an underperforming Derek Jeter might be under "pressure" or too much "media scrutiny," or subject to the difficulties of performing in a "New York market." When individuals start off well but then appear to fade, it must be because people are trying to stop them ("gunning" for someone is a common expression). One should always remember that in estimating intrinsic ability, individuals are unlikely to be as good (or as bad) as the pace they are on. It is always a better bet to vote against someone eventually breaking a record, even when they are "on a pace" to so do early in the season. This may be one origin for the phrase "sucker bet"—a gambling wager where your expected return is significantly lower than your bet.

Another area where one expects to see a lot of anomalous results is when the dataset is split into ever-finer categorizations that end up having few observations in them, and thus subject to much greater variability. For example, should we be overly surprised if Albert Pujols doesn't seem to bat well in domed stadiums at night when batting second against left-handed pitching? The pundits look for "causes" for these kinds of extremes when they should just be marveling at the beauty of natural variation and the effects of sample size. A similar and probably more important misleading effect occurs when our data are on the effectiveness of some medical treatment, and we try to attribute positive or negative results to ever-finer-grained classifications of the clinical subjects.

Random processes are a fundamental part of nature and ubiquitous in our day-to-day lives. Most people do not understand them, or worse, fall under an "illusion of control" and believe they have influence over how events progress. Thus, there is an almost mystical belief in the ability of a new coach, CEO, or president to "turn things around." Part of these strong beliefs may result from the operation of regression toward the mean or the natural unfolding of any random process. We continue to get our erroneous beliefs reconfirmed when cause is attributed when none may actually be present. As humans we all wish to believe we can affect our future, but when events have dominating stochastic components, we are obviously not in complete control. There appears to be a fundamental clash between our ability to recognize the operation of randomness and the need for control in our lives.

An appreciation for how random processes might operate can be helpful in navigating the uncertain world we live in. When investments with Bernie Madoff give perfect 12% returns, year after year, with no exceptions and no variability, alarms should go off. If we see a supposed scatterplot of two fallible variables with a least-squares line imposed but where the actual data points have been withdrawn, remember that the relationship is not perfect. Or when we monitor error in quality assurance and control for various manufacturing or diagnostic processes (for example, application of radiation in medicine), and the tolerances become consistently beyond the region where we should generally expect the process to vary, a need to stop and recalibrate may be necessary. It is generally important to recognize that data interpretation may be a long-term process, with a need to appreciate variation appearing around a trend line. Thus, the immediacy of some major storms does not vitiate a longer-term perspective on global climate change. Remember the old meteorological adage:

climate is what you expect; weather is what you get. Relatedly, it is important to monitor processes we have some personal responsibility for (such as our own lipid panels when we go for physicals), and to assess when unacceptable variation appears outside of our normative values.

Besides having an appreciation for randomness in our day-to-day lives, there is also a flip side: if you don't see randomness when you probably should, something is amiss. The Bernie Madoff example noted above is a salient example, but there are many such deterministic traps awaiting the gullible. When something seems just too good to be true, most likely it isn't. A recent ongoing case in point involves the Dutch social psychologist, Diederik Stapel, and the massive fraud he committed in the very best psychology journals in the field. A news item by G. Vogel in *Science* (2011, *334*, 579) has the title, "Psychologist Accused of Fraud on 'Astonishing Scale'." Basically, in dozens of published articles and doctoral dissertations he supervised, Stapel never failed to obtain data showing the clean results he expected to see at the outset. As any practicing researcher in the behavioral sciences knows, this is just too good to be true. We give a short quotation from the *Science* news item (October 31, 2011) commenting on the Tilberg University report on the Stapel affair (authored by a committee headed by the well-known Dutch psycholinguist, Willem Levelt):

> Stapel was "absolute lord of the data" in his collaborations ... many of Stapel's datasets have improbable effect sizes and other statistical irregularities, the report says. Among Stapel's colleagues, the description of data as too good to be true "was a heartfelt compliment to his skill and creativity."

The report discusses the presence of consistently large effects being found; few missing data and outliers; hypotheses rarely refuted. Journals publishing Stapel's articles did not question the omission of details about the source of the data. As understated by Levelt, "We see that the scientific checks and balances process has failed at several levels." In a related article in the *New York Times* by Benedict Carey (November 2, 2011), "Fraud Case Seen as a Red Flag for Psychology Research," the whole field of psychology is now taken to task, appropriately we might add, in how research has generally been done and evaluated in the field. Part of the Levelt Committee report that deals explicitly with data and statistical analysis is redacted below:

> *The data were too good to be true; the hypotheses were almost always confirmed; the effects were improbably large; missing data, or impossible, out-of-range data, are rare or absent.*

This is possibly the most precarious point of the entire data fraud. Scientific criticism and approach failed on all fronts in this respect. The falsification of hypotheses is a fundamental principle of science, but was hardly a part of the research culture surrounding Mr. Stapel. The only thing that counted was verification. However, anyone with any research experience, certainly in this sector, will be aware that most hypotheses that people entertain do not survive. And if they do, the effect often vanishes with replication. The fact that Mr. Stapel's hypotheses were always confirmed should have caused concern, certainly when in most cases the very large "effect sizes" found were clearly out of line with the literature. Rather than concluding that this was all improbable, instead Mr. Stapel's experimental skills were taken to be phenomenal. "Too good to be true" was meant as a genuine compliment to his skill and creativity. Whereas all these excessively neat findings should have provoked thought, they were embraced. If other researchers had failed, that was assumed to be because of a lack of preparation, insight, or experimental skill. Mr. Stapel became the model: the standard. Evidently only Mr. Stapel was in a position to achieve the precise manipulations needed to make the subtle effects visible. People accepted, if they even attempted to replicate the results for themselves, that they had failed because they lacked Mr. Stapel's skill. However, there was usually no attempt to replicate, and certainly not independently. The few occasions when this did happen, and failed, were never revealed, because the findings were not publishable.

In other words, scientific criticism has not performed satisfactorily on this point. Replication and the falsification of hypotheses are cornerstones of science. Mr. Stapel's verification factory should have aroused great mistrust among colleagues, peers and journals.

As a supervisor and dissertation advisor, Mr. Stapel should have been expected to promote this critical attitude among his students. Instead, the opposite happened. A student who performed his own replications with no result was abandoned to his fate rather than praised and helped.

Strange, improbable, or impossible data patterns; strange correlations; identical averages and standard deviations; strange univariate distributions of variables.

The actual data displayed several strange patterns that should have been picked up. The patterns are related to the poor statistical foundation of Mr. Stapel's data fabrication approach (he also tended to make denigrating remarks about statistical methods). It has emerged that some of the fabrication involved simply "blindly" entering numbers based on the desired bivariate relationships, and by cutting and pasting data columns. This approach sometimes gave rise to strange data patterns. Reordering the data matrix by size of a given variable sometimes produces a matrix in which one column is identical to another, which is therefore the simple result of cutting and pasting certain scores. It was also possible for a variable that would normally score only a couple of per cent "antisocial," for no reason and unexpectedly suddenly to show "antisocial" most of the time. Independent replication yielded exactly the same averages and standard deviations. Two independent variables that always correlated positively, conceptually and in other research, now each had the right expected effects on the dependent

> variable, but correlated negatively with each other. There was no consistent checking of data by means of simple correlation matrices and univariate distributions. It is to the credit of the whistle blowers that they did discover the improbabilities mentioned above.
>
> Finally, a lamentable element of the culture in social psychology and psychology research is for everyone to keep their own data and not make them available to a public archive. This is a problem on a much larger scale, as has recently become apparent. Even where a journal demands data accessibility, authors usually do not comply ... Archiving and public access to research data not only makes this kind of data fabrication more visible, it is also a condition for worthwhile replication and meta-analysis. (pp. 13-15)

2 Regression Toward the Mean

Regression toward the mean is a phenomenon that will occur whenever dealing with fallible measures with a less-than-perfect correlation. The word "regression" was first used by Galton in his 1886 article, "Regression Towards Mediocrity in Hereditary Stature." Galton showed that heights of children from very tall or short parents regress toward mediocrity (that is, toward the mean) and exceptional scores on one variable (parental height) are not matched with such exceptionality on the second (child height). This observation is purely due to the fallibility for the various measures and the concomitant lack of a perfect correlation between the heights of parents and their children.

Regression toward the mean is a ubiquitous phenomenon, and given the name "regressive fallacy" whenever cause is ascribed where none exists. Generally, interventions are undertaken if processes are at an extreme (for example, a crackdown on speeding or drunk driving as fatalities spike, treatment groups formed from individuals who are seriously depressed, or individuals selected because of extreme good or bad behaviors). In all such instances, whatever remediation is carried out will be followed by some lessened value on a response variable. Whether the remediation was itself causative is problematic to assess given the universality of regression toward the mean.

There are many common instances where regression may lead to invalid reasoning: I went to my doctor and my pain has now lessened; I instituted corporal punishment and behavior has improved; he was jinxed by a *Sports Illustrated* cover because subsequent performance was poorer (also known as the "sophomore jinx"); although he hadn't had a hit in some time, he was "due," and the coach played him; and so on. More generally, any time one optimizes with respect to a given sample of data by constructing prediction functions of some kind, there is an implicit use and reliance on data extremities. In other words, the various measures of goodness of fit or prediction calculated need to be cross-validated either on new data or by a clever sample reuse strategy such as the well-known jackknife or bootstrap procedures. The degree of "shrinkage" seen in our measures based on this cross-validation is an indication of the fallibility of our measures and the (in)adequacy of the given sample sizes.

The misleading interpretive effects engendered by regression toward the mean are legion, particularly when we wish to interpret observational studies for some indication of causality. There is a continual violation of the traditional adage that "the rich get richer and the poor get poorer," in favor of "when you are at the top, the only way is down." Extreme scores are never quite as extreme as they first appear. Many of these regression artifacts are discussed in the cautionary source, *A Primer on Regression Artifacts* (Campbell & Kenny, 1999), including the various difficulties encountered in trying to equate intact groups by matching or analysis of covariance. Statistical equating creates the illusion but not the reality of equivalence. As summarized by Campbell and Kenny, "the failure to understand the likely direction of bias when statistical equating is used is one of the most serious difficulties in contemporary data analysis" (p. 85).

The historical prevalence of the regression fallacy is considered by Stephen Stigler in his 1997 article entitled "Regression Towards the Mean, Historically Considered" (*Statistical Methods in Medical Research*, *6*, 103–114). Stigler labels it "a trap waiting for the unwary, who were legion" (p. 112). He relates a story that we excerpt below about a Northwestern University statistician falling into the trap in 1933:

> The most spectacular instance of a statistician falling into the trap was in 1933, when a Northwestern University professor named Horace Secrist unwittingly wrote a whole book on the subject, *The Triumph of Mediocrity in Business*. In over 200 charts and tables, Secrist "demonstrated" what he took to be an important economic phenomenon, one that likely lay at the root of the great depression: a tendency for firms to grow more mediocre over time. Secrist was aware of Galton's work; he cited it and used Galton's terminology. The preface even acknowledged "helpful criticism" from such statistical luminaries as HC Carver (the editor of the *Annals of Mathematical Statistics*), Raymond Pearl, EB Wilson, AL Bowley, John Wishart and Udny Yule. How thoroughly these statisticians were informed of Secrist's work is unclear, but there is no evidence that they were successful in alerting him to the magnitude of his folly (or even if they noticed it). Most of the reviews of the book applauded it. But there was one dramatic exception: in late 1933 Harold Hotelling wrote a devastating review, noting among other things that

> "The seeming convergence is a statistical fallacy, resulting from the method of grouping. These diagrams really prove nothing more than that the ratios in question have a tendency to wander about." (p. 112)

Stigler goes on to comment about the impact of the Secrist-Hotelling episode for the recognition of the importance of regression toward the mean:

> One would think that so public a flogging as Secrist received for his blunder would wake up a generation of social scientists to the dangers implicit in this phenomenon, but that did not happen. Textbooks did not change their treatment of the topic, and if there was any increased awareness of it, the signs are hard to find. In the more than two decades between the Secrist-Hotelling exchange in 1933 and the publication in 1956 of a perceptively clear exposition in a textbook by W Allen Wallis and Harry Roberts, I have only encountered the briefest acknowledgements. (p. 113)

A variety of phrases seem to get attached whenever regression toward the mean is probably operative. We have the "winner's curse," where someone is chosen from a large pool (such as of job candidates), who then doesn't live up to expectations; or when we attribute some observed change to the operation of "spontaneous remission." As Campbell and Kenny noted, "many a quack has made a good living from regression toward the mean." Or, when a change of diagnostic classification results upon repeat testing for an individual given subsequent one-on-one tutoring (after being placed, for example, in a remedial context). More personally, there is "editorial burn-out" when someone is chosen to manage a prestigious journal at the apex of a career, and things go quickly downhill from that point.

3 Incorporating Reliability Corrections in Prediction

As discussed in the previous section, a recognition of when regression toward the mean might be operative can assist in avoiding the "regressive fallacy." In addition to this cautionary usage, the same regression-toward-the-mean phenomenon can make a positive contribution to the task of prediction with fallible information, and particularly in how such prediction can be made more accurate by correcting for the unreliability of the available variables. To make the argument a bit more formal, we assume an implicit underlying model for how any observed score, X, might be constructed additively from a true score, T_X, and an error score, E_X, where E_X is typically considered uncorrelated with T_X: $X = T_X + E_X$. The distribution of the observed variable over, say, a population of individuals, involves two sources of variability in the true and the error scores. If interests center on structural models among true scores, some correction should be made to the observed variables because the common regression models implicitly assume that all variables are measured without error. But before "errors-in-variables" models are briefly discussed, our immediate concern will be with how best to predict a true score from the observed score.[1]

The estimation, $\hat{T}_X$, of a true score from an observed score, X, was derived using the regression model by Kelley in the 1920s (Kelley,

[1] When an observed score is directly used as a prediction for the true score, the prediction is referred to as "non-regressive" and reflects an over-confidence in the fallible observed score as a direct reflection of the true score. One commonly used baseball example is to consider an "early-in-the-season" batting average (an "observed" score) as a direct prediction of an "end-of-the-season batting average (a presumed "true" score). As given by Kelley's equation in the text, better estimates of the true scores would regress the observed scores toward the average of the observed scores.

1947), with a reliance on the algebraic equivalence that the squared correlation between observed and true score is the reliability. If we let $\hat{\rho}$ be the estimated reliability, Kelley's equation can be written as

$$\hat{T}_X = \hat{\rho}X + (1 - \hat{\rho})\bar{X} \ ,$$

where $\bar{X}$ is the mean of the group to which the individual belongs. In other words, depending on the size of $\hat{\rho}$, a person's estimate is partly due to where the person is in relation to the group—upward if below the mean, downward if above. The application of this statistical tautology in the examination of group differences provides such a surprising result to the statistically naive that this equation has been labeled "Kelley's Paradox" (Wainer, 2005, pp. 67–70).

In addition to obtaining a true score estimate from an obtained score, Kelly's regression model also provides a standard error of estimation (which in this case is now referred to as the standard error of measurement). An approximate 95% confidence interval on an examinee's true score is given by

$$\hat{T}_X \pm 2\hat{\sigma}_X((\sqrt{1 - \hat{\rho}})\sqrt{\hat{\rho}}) \ ,$$

where $\hat{\sigma}_X$ is the (estimated) standard deviation of the observed scores. By itself, the term $\hat{\sigma}_X((\sqrt{1 - \hat{\rho}})\sqrt{\hat{\rho}})$ is the standard error of measurement, and is generated from the usual regression formula for the standard error of estimation but applied to Kelly's model that predicts true scores. The standard error of measurement most commonly used in the literature is not Kelly's but rather $\hat{\sigma}_X\sqrt{1 - \hat{\rho}}$, and a 95% confidence interval taken as the observed score plus or minus twice this standard error. An argument can be made that this latter procedure leads to "reasonable limits" (after Gulliksen, 1950) whenever

$\hat{\rho}$ is reasonably high, and the obtained score is not extremely deviant from the reference group mean. Why we should assume these latter preconditions and not use the more appropriate procedure to begin with, reminds us of a Bertrand Russell quotation (1919, p. 71): "The method of postulating what we want has many advantages; they are the same as the advantages of theft over honest toil."[2]

[2] The standard error of measurement (SEM) can play a significant role in the legal system as to who is eligible for execution. The recent Supreme Court case of *Hall v. Florida* (2014) found unconstitutional a "bright-line" Florida rule about requiring an I.Q. score of 70 or below to forestall execution due to intellectual disability. We redact part of this ruling as it pertains to the SEM of an I.Q. test:

FREDDIE LEE HALL, PETITIONER v. FLORIDA

ON WRIT OF CERTIORARI TO THE SUPREME COURT OF FLORIDA

[May 27, 2014]

JUSTICE KENNEDY delivered the opinion of the Court.

This Court has held that the Eighth and Fourteenth Amendments to the Constitution forbid the execution of persons with intellectual disability (*Atkins v. Virginia*). Florida law defines intellectual disability to require an IQ test score of 70 or less. If, from test scores, a prisoner is deemed to have an IQ above 70, all further exploration of intellectual disability is foreclosed. This rigid rule, the Court now holds, creates an unacceptable risk that persons with intellectual disability will be executed, and thus is unconstitutional.

...

On its face, the Florida statute could be consistent with the views of the medical community noted and discussed in *Atkins*. Florida's statute defines intellectual disability for purposes of an *Atkins* proceeding as "significantly subaverage general intellectual functioning existing concurrently with deficits in adaptive behavior and manifested during the period from conception to age 18." ... The statute further defines "significantly subaverage general intellectual functioning" as "performance that is two or more standard deviations from the mean score on a standardized intelligence test." ... The mean IQ test score is 100. The concept of standard deviation describes how scores are dispersed in a population. Standard deviation is distinct from standard error of measurement, a concept which describes the reliability of a test and is discussed further below. The standard deviation on an IQ test is approximately 15 points, and so two standard deviations is approximately 30 points. Thus a test taker who performs "two or more standard deviations from the mean" will score approximately 30 points below the mean on an IQ test, i.e., a score of approximately 70 points.

On its face this statute could be interpreted consistently with *Atkins* and with the conclusions this Court reaches in the instant case. Nothing in the statute precludes Florida from

There are several remarkable connections between Kelley's work

taking into account the IQ test's standard error of measurement, and as discussed below there is evidence that Florida's Legislature intended to include the measurement error in the calculation. But the Florida Supreme Court has interpreted the provisions more narrowly. It has held that a person whose test score is above 70, including a score within the margin for measurement error, does not have an intellectual disability and is barred from presenting other evidence that would show his faculties are limited. ... That strict IQ test score cutoff of 70 is the issue in this case.

Pursuant to this mandatory cutoff, sentencing courts cannot consider even substantial and weighty evidence of intellectual disability as measured and made manifest by the defendant's failure or inability to adapt to his social and cultural environment, including medical histories, behavioral records, school tests and reports, and testimony regarding past behavior and family circumstances. This is so even though the medical community accepts that all of this evidence can be probative of intellectual disability, including for individuals who have an IQ test score above 70. ... ("[T]he relevant clinical authorities all agree that an individual with an IQ score above 70 may properly be diagnosed with intellectual disability if significant limitations in adaptive functioning also exist"); ... ("[A] person with an IQ score above 70 may have such severe adaptive behavior problems ... that the person's actual functioning is comparable to that of individuals with a lower IQ score").

Florida's rule disregards established medical practice in two interrelated ways. It takes an IQ score as final and conclusive evidence of a defendant's intellectual capacity, when experts in the field would consider other evidence. It also relies on a purportedly scientific measurement of the defendant's abilities, his IQ score, while refusing to recognize that the score is, on its own terms, imprecise.

The professionals who design, administer, and interpret IQ tests have agreed, for years now, that IQ test scores should be read not as a single fixed number but as a range. ... Each IQ test has a "standard error of measurement," ... often referred to by the abbreviation "SEM." A test's SEM is a statistical fact, a reflection of the inherent imprecision of the test itself. ... An individual's IQ test score on any given exam may fluctuate for a variety of reasons. These include the test-taker's health; practice from earlier tests; the environment or location of the test; the examiner's demeanor; the subjective judgment involved in scoring certain questions on the exam; and simple lucky guessing.

The SEM reflects the reality that an individual's intellectual functioning cannot be reduced to a single numerical score. For purposes of most IQ tests, the SEM means that an individual's score is best understood as a range of scores on either side of the recorded score. The SEM allows clinicians to calculate a range within which one may say an individual's true IQ score lies. ... In addition, because the test itself may be flawed or administered in a consistently flawed manner, multiple examinations may result in repeated similar scores, so that even a consistent score is not conclusive evidence of intellectual functioning.

Despite these professional explanations, Florida law used the test score as a fixed number, thus barring further consideration of other evidence bearing on the question of intellectual in the first third of the twentieth century and the modern theory of statistical estimation developed in the last half of the century. In considering the model for an observed score, X, to be a sum of a true score, T, and an error score, E, plot the observed test scores on the x-axis and their true scores on the y-axis. As noted by Galton in the 1880s (Galton, 1886), any such scatterplot suggests two regression lines. One is of true score regressed on observed score (generating Kelley's true score estimation equation given in the text); the second is the regression of observed score being regressed on true score (generating the use of an observed score to directly estimate the observed score). Kelley clearly knew the importance for measurement theory of this distinction between two possible regression lines in a true-score versus observed-score scatterplot. The quotation given below is from his 1927 text, *Interpretation of Educational Measurements*. The reference to the "last section" is where the true score was estimated directly by the observed score; the "present section" refers to his true score regression estimator:

> This tendency of the estimated true score to lie closer to the mean than the obtained score is the principle of regression. It was first discovered by Francis Galton and is a universal phenomenon in correlated data. We may now characterize the procedure of the last and present sections by saying that in the last section regression was not allowed for and in the present it is. If the reliability is very high, then there is little difference between [the two methods], so that this second technique, which is slightly the more laborious, is not demanded, but if the reliability is low, there is much difference in individual outcome, and the refined procedure is always to be used in making

disability. For professionals to diagnose – and for the law then to determine – whether an intellectual disability exists once the SEM applies and the individual's IQ score is 75 or below the inquiry would consider factors indicating whether the person had deficits in adaptive functioning. These include evidence of past performance, environment, and upbringing.

individual diagnoses. (p. 177)

Kelley's preference for the refined procedure when reliability is low (that is, for the regression estimate of true score) is due to the standard error of measurement being smaller (unless reliability is perfect); this is observable directly from the formulas given earlier. There is a trade-off in moving to the regression estimator of the true score in that a smaller error in estimation is paid for by using an estimator that is now biased. Such trade-offs are common in modern statistics in the use of "shrinkage" estimators (for example, ridge regression, empirical Bayes methods, James–Stein estimators). Other psychometricians, however, apparently just don't buy the trade-off; for example, see Gulliksen (*Theory of Mental Tests*; 1950); Gulliksen wrote that "no practical advantage is gained from using the regression equation to estimate true scores" (p. 45). We disagree—who really cares about bias when a generally more accurate prediction strategy can be defined?

What may be most remarkable about Kelley's regression estimate of true score is that it predates the work in the 1950s on "Stein's Paradox" that shook the foundations of mathematical statistics. A readable general introduction to this whole statistical kerfuffle is the 1977 *Scientific American* article by Bradley Efron and Carl Morris, "Stein's Paradox in Statistics" (*236*(5), 119-127). When reading this popular source, keep in mind that the class referred to as James–Stein estimators (where bias is traded off for lower estimation error) includes Kelley's regression estimate of the true score. We give an excerpt below from Stephen Stigler's 1988 Neyman Memorial Lecture, "A Galtonian Perspective on Shrinkage Estimators" (*Statisti-*

cal Science, 1990, *5*, 147-155), that makes this historical connection explicit:

> The use of least squares estimators for the adjustment of data of course goes back well into the previous century, as does Galton's more subtle idea that there are two regression lines. ... Earlier in this century, regression was employed in educational psychology in a setting quite like that considered here. Truman Kelley developed models for ability which hypothesized that individuals had true scores ... measured by fallible testing instruments to give observed scores ... ; the observed scores could be improved as estimates of the true scores by allowing for the regression effect and shrinking toward the average, by a procedure quite similar to the Efron–Morris estimator. (p. 152)

Before we leave the topic of true score estimation by regression, we might also note what it does not imply. When considering an action for an individual where the goal is to help make, for example, the right level of placement in a course or the best medical treatment and diagnosis, then using group membership information to obtain more accurate estimates is the appropriate course to follow. But if we are facing a contest, such as awarding scholarships, or offering admission or a job, then it is inappropriate (and ethically questionable) to search for identifiable subgroups that a particular person might belong to and then adjust that person's score accordingly. Shrinkage estimators are "group blind." Their use is justified for whatever population is being observed; it is generally best for accuracy of estimation to discount extremes and "pull them in" toward the (estimated) mean of the population.

In the topic of errors-in-variables regression, we try to compensate for the tacit assumption in regression that all variables are measured without error. Measurement error in a response variable does not bias the regression coefficients per se, but it does increase standard errors and thereby reduces power. This is generally a common effect: unreliability attenuates correlations and reduces power even in standard ANOVA paradigms. Measurement error in the predictor variables biases the regression coefficients. For example, for a single predictor, the observed regression coefficient is the "true" value multiplied by the reliability coefficient. Thus, without taking account of measurement error in the predictors, regression coefficients will generally be underestimated, producing a biasing of the structural relationship among the true variables. Such biasing may be particularly troubling when discussing econometric models where unit changes in observed variables are supposedly related to predicted changes in the dependent measure; possibly the unit changes are more desired at the level of the true scores.

Milton Friedman's 1992 article entitled "Do Old Fallacies Ever Die?" (*Journal of Economic Literature*, *30*, 2129-2132), gives a downbeat conclusion regarding errors-in-variables modeling:

> Similarly, in academic studies, the common practice is to regress a variable Y on a vector of variables X and then accept the regression coefficients as supposedly unbiased estimates of structural parameters, without recognizing that all variables are only proxies for the variables of real interest, if only because of measurement error, though generally also because of transitory factors that are peripheral to the subject under consideration. I suspect that the regression fallacy is the most common fallacy in the statistical analysis of economic data, alleviated only occasionally by consideration of the bias introduced when "all variables are subject to error." (p. 2131)

4 Complete Enumeration versus Sampling in the Census

The basic sampling model implies that when the size of the population is effectively infinite, this does not affect the accuracy of our estimate, which is driven solely by sample size. Thus, if we want a more precise estimate, we need only draw a larger sample.[3] For some reason, this confusion resurfaces and is reiterated every ten years when the United States Census is planned, where the issue of complete enumeration, as demanded by the Constitution, and the problems of undercount are revisited. We begin with a short excerpt from a *New York Times* article by David Stout (April 2, 2009), "Obama's Census Choice Unsettles Republicans." The quotation it contains from John Boehner in relation to the 2010 census is a good instance of the "resurfacing confusion"; also, the level of Boehner's statistical reasoning skills should be fairly clear.

> Mr. Boehner, recalling that controversy [from the early 1990s when Mr. Groves pushed for statistically adjusting the 1990 census to make up for an undercount], said Thursday that "we will have to watch closely to ensure the 2010 census is conducted without attempting similar statistical sleight of hand."

There has been a continuing and decades-long debate about the efficacy of using surveys to correct the census for an undercount. The

[3] Courts have been distrustful of sampling versus complete enumeration, and have been so for a long time. A case in 1955, for example, involved Sears, Roebuck, and Company and the City of Inglewood (California). The Court ruled that a sample of receipts was inadequate to estimate the amount of taxes that Sears had overpaid. Instead, a costly complete audit or enumeration was required. For a further discussion of this case, see R. Clay Sprowls, "The Admissibility of Sample Data into a Court of Law: A Case History," *UCLA Law Review*, *4*, 222–232, 1956–1957.

arguments against surveys are based on a combination of partisan goals and ignorance. Why? First, the census is a big, costly, and complicated procedure. And like all such procedures, it will have errors. For example, there will be errors where some people are counted more than once, such as an affluent couple with two homes being visited by census workers in May in one and by different workers in July at the other, or they are missed entirely. Some people are easier to count than others. Someone who has lived at the same address with the same job for decades, and who faithfully and promptly returns census forms, is easy to count. Someone else who moves often, is a migrant laborer or homeless and unemployed, is much harder to count. There is likely to be an undercount of people in the latter category. Republicans believe those who are undercounted are more likely to vote Democratic, and so if counted, the districts they live in will get increased representation that is more likely to be Democratic. The fact of an undercount can be arrived at through just logical considerations, but its size must be estimated through surveys. Why is it we can get a better estimate from a smallish survey than from an exhaustive census? The answer is that surveys are, in fact, small. Thus, their budgets allow them to be done carefully and everyone in the sampling frame can be tracked down and included (or almost everyone).[4] A complete enumeration is a big deal, and even though census workers try hard, they have a limited (although large) budget that does not allow the same level of precision. Because of the enormous size of the census task, increasing the budget to any plausible level will still not be enough to get everyone. A number of well-designed surveys will do a better job at a fraction of the cost.

[4] A sampling frame is the list of all those in the population that can be sampled.

The Supreme Court ruling in *Department of Commerce v. United States House of Representatives* (1999) seems to have resolved the issue of sampling versus complete enumeration in a Solomon-like manner. For purposes of House of Representatives apportionment, complete enumeration is required with all its problems of "undercount." For other uses of the Census, however, "undercount" corrections that make the demographic information more accurate are permissable And these corrected estimates could be used in differential resource allocation to the states. Two items are given in an appendix below: a short excerpt from the American Statistical Association *amicus* brief for this case, and the syllabus from the Supreme Court ruling.

5 Appendix: Brief for American Statistical Association as Amicus Curiae, Department of Commerce v. United States House of Representatives

Friend of the Court brief from the American Statistical Association —

ASA takes no position on the appropriate disposition of this case or on the legality or constitutionality of any aspect of the 2000 census. ASA also takes no position in this brief on the details of any proposed use of statistical sampling in the 2000 census.

ASA is, however, concerned to defend statistically designed sampling as a valid, important, and generally accepted scientific method for gaining accurate knowledge about widely dispersed human populations. Indeed, for reasons explained in this brief, properly designed sampling is often a better and more accurate method of gaining such knowledge than an inevitably incomplete attempt to survey all members of such a population. Therefore, in principle, statistical sampling applied to the census "has the potential to increase the quality and accuracy of the count and to reduce costs." . . . There are no sound scientific grounds for rejecting all use of statistical sampling in the 2000 census.

As its argument in this brief, ASA submits the statement of its Blue Ribbon Panel that addresses the relevant statistical issues. ASA respectfully submits this brief in hopes that its explanation of these points will be helpful to the Court.

6 Appendix: Department of Commerce v. United States House of Representatives

Syllabus from the Supreme Court ruling: The Constitution's Census Clause authorizes Congress to direct an "actual Enumeration" of the American public every 10 years to provide a basis for apportioning congressional representation among the States. Pursuant to this authority, Congress has enacted the Census Act (Act), ... delegating the authority to conduct the decennial census to the Secretary of Commerce (Secretary). The Census Bureau (Bureau), which is part of the Department of Commerce, announced a plan to use two forms of statistical sampling in the 2000 Decennial Census to address a chronic and apparently growing problem of "undercounting" of some identifiable groups, including certain minorities, children, and renters. In early 1998, two sets of plaintiffs filed separate suits challenging the legality and constitutionality of the plan. The suit in No. 98-564 was filed in the District Court for the Eastern District of Virginia by four counties and residents of 13 States. The suit in No. 98-404 was filed by the United States House of Representatives in the District Court for the District of Columbia. Each of the courts held that the plaintiffs satisfied the requirements for Article III standing, ruled that the Bureau's plan for the 2000 census violated the Census Act, granted the plaintiffs' motion for summary judgment, and permanently enjoined the planned use of statistical sampling to determine the population for congressional apportionment purposes. On direct appeal, this Court consolidated the cases for oral argument.

Held:

1. Appellees in No. 98-564 satisfy the requirements of Article III standing. In order to establish such standing, a plaintiff must allege personal injury fairly traceable to the defendant's allegedly unlawful conduct and likely to

be redressed by the requested relief. . . . A plaintiff must establish that there exists no genuine issue of material fact as to justiciability or the merits in order to prevail on a summary judgment motion. . . . The present controversy is justiciable because several of the appellees have met their burden of proof regarding their standing to bring this suit. In support of their summary judgment motion, appellees submitted an affidavit that demonstrates that it is a virtual certainty that Indiana, where appellee Hofmeister resides, will lose a House seat under the proposed census 2000 plan. That loss undoubtedly satisfies the injury-in-fact requirement for standing, since Indiana residents' votes will be diluted by the loss of a Representative. . . . Hofmeister also meets the second and third standing requirements: There is undoubtedly a "traceable" connection between the use of sampling in the decennial census and Indiana's expected loss of a Representative, and there is a substantial likelihood that the requested relief—a permanent injunction against the proposed uses of sampling in the census—will redress the alleged injury. Appellees have also established standing on the basis of the expected effects of the use of sampling in the 2000 census on intrastate redistricting. Appellees have demonstrated that voters in nine counties, including several of the appellees, are substantially likely to suffer intrastate vote dilution as a result of the Bureau's plan. Several of the States in which the counties are located require use of federal decennial census population numbers for their state legislative redistricting, and States use the population numbers generated by the federal decennial census for federal congressional redistricting. Appellees living in the nine counties therefore have a strong claim that they will be injured because their votes will be diluted vis-à-vis residents of counties with larger undercount rates. The expected intrastate vote dilution satisfies the injury-in-fact, causation, and redressibility requirements.

2. The Census Act prohibits the proposed uses of statistical sampling to determine the population for congressional apportionment purposes. In 1976, the provisions here at issue took their present form. Congress revised 13 U. S. C. §141(a), which authorizes the Secretary to "take a decennial census . . . in such form and content as he may determine, including the use of sampling procedures." This broad grant of authority is informed, however, by the narrower and more specific §195. As amended in 1976, §195 provides: "Except

for the determination of population for purposes of [congressional] apportionment … the Secretary shall, if he considers it feasible, authorize the use of … statistical … 'sampling' in carrying out the provisions of this title." Section 195 requires the Secretary to use sampling in assembling the myriad demographic data that are collected in connection with the decennial census, but it maintains the longstanding prohibition on the use of such sampling in calculating the population for congressional apportionment. Absent any historical context, the "except/shall" sentence structure in the amended §195 might reasonably be read as either permissive or prohibitive. However, the section's interpretation depends primarily on the broader context in which that structure appears. Here, that context is provided by over 200 years during which federal census statutes have uniformly prohibited using statistical sampling for congressional apportionment. The Executive Branch accepted, and even advocated, this interpretation of the Act until 1994.

3. Because the Court concludes that the Census Act prohibits the proposed uses of statistical sampling in calculating the population for purposes of apportionment, the Court need not reach the constitutional question presented.

References

[1] Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

[2] Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

[3] Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

[4] Levelt Committee. (2011, October 31). *Interim report regarding the breach of scientific integrity committed by Prof. D. A. Stapel*. Tilburg, The Netherlands: Tilburg University.

[5] Russell, B. (1919). *Introduction to mathematical philosophy.* New York: Macmillan.

[6] Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures.* Princeton, NJ: Princeton University Press.

[7] Wainer, H. (2009). *Picturing the uncertain world: How to understand, communicate, and control uncertainty through graphical display.* Princeton, NJ: Princeton University Press.

Module 7: Probabilistic (Mis)Reasoning and Related Confusions

bioinformatics: a synergistic fusion of huge data bases and bad statistics
data mining: panning for gold in a sewer
– Stephen Senn (*Dicing with Death*, 2003)

Abstract: The introductory module started with the well-known case of Sally Clark and how a misunderstanding about probabilistic independence helped lead to her wrongful imprisonment for killing her two children. The present module will provide more examples of mistaken probabilistic reasoning, with many involving misinterpretations of conditional probability. We will revisit the O.J. Simpson criminal case where his defense team took advantage of what is termed the "defendant's fallacy," as well as some specious reasoning about conditional probability (perpetrated by Alan Dershowitz). Several additional high-profile legal cases will be mentioned that were mishandled because of the prosecutor's fallacy, much like that of Sally Clark. One is recent – the Dutch nurse, Lucia de Berk, was accused of multiple deaths at the hospitals she worked at in the Netherlands; another is much older and involves the turn-of-the-century (the late 1800s, that is) case of Alfred Dreyfus, the much maligned French Jew who was falsely imprisoned for espionage.

Contents

1 The (Mis)assignment of Probabilities

Clear probabilistic reasoning requires a good understanding of how conditional probabilities are defined and operate. There are many day-to-day contexts we face where decisions might best be made from conditional probabilities, if we knew them, instead of from marginal information. When deciding on a particular medical course of action, for example, it is important to condition on personal circumstances of age, risk factors, family medical history, and our own psychological needs and makeup. A fairly recent and controversial instance of this, where the conditioning information is "age," is reported in the *New York Times* article by Gina Kolata, "Panel Urges Mammograms at 50, Not 40" (November 16, 2009). The failure to consider conditional instead of marginal probabilities is particularly grating for many of us who follow various sporting activities and enjoy second-guessing managers, quarterbacks, sports commentators, and their ilk. As an example, consider the "strike-'em-out-throw-'em-out" double play in baseball, where immediately after the batter has swung and missed at a third strike or taken a called third strike, the catcher throws out a base runner attempting to steal second or third base. Before such a play occurs, announcers routinely state that the runner "will or will not be sent" because the "batter strikes out only some percentage of the time." The issue of running or not shouldn't be based on the marginal probability of the batter striking out but on some conditional probability (for example, how often does the batter strike out when faced with a particular count or type of pitcher). For many other instances, however, we might be content not to base our decisions on conditional information; for example, always wear a seat belt irrespective of the type or length of trip being taken.

Although the assignment of probabilities to events consistent with the mutually exclusive event rule may lead to an internally valid system mathematically, there is still no assurance that this assignment is "meaningful," or bears any empirical validity for observable long-run expected frequencies. There seems to be a never-ending string of misunderstandings in the way probabilities can be generated that are either blatantly wrong, or more subtly incorrect, irrespective of the internally consistent system they might lead to. Some of these problems are briefly sketched below, but we can only hope to be representative of a few possibilities, not exhaustive.

One inappropriate way of generating probabilities is to compute the likelihood of some joint occurrence after some of the outcomes are already known. For example, there is the story about the statistician who takes a bomb aboard a plane, reasoning that if the probability of one bomb on board is small, the probability of two is infinitesimal. Or, during World War I, soldiers were actively encouraged to use fresh shell holes as shelter because it was very unlikely for two shells to hit the same spot during the same day. And the Minnesota Twins baseball manager who bats for an individual who earlier in the game hit a home run because it would be very unlikely for him to hit two home runs in the same game. Although these slightly amusing stories may provide obvious misassignments of probabilities, other related situations are more subtle. For example, whenever coincidences are culled or "hot spots" identified from a search of available information, the probabilities that are then regenerated for these situations may not be valid. There are several ways of saying this: when some set of observations is the source of an initial suspicion, those same observations should not be used in a calculation that then tests the validity of the suspicion. In Bayesian terms, you should not obtain the posterior probabilities from the same information that gave you the prior probabilities. Alternatively said, it makes no sense to do formal hypothesis assessment by finding estimated probabilities when the data themselves have suggested the hypothesis in the first place. Some cross-validation strategy is necessary; for example, collecting independent data. Generally, when some process of search or optimization has been used to identify an unusual situation (for instance, when a "good" regression equation is found through a step-wise procedure [see Freedman, 1983, for a devastating critique]; when data are "mined" and unusual patterns identified; when DNA databases are searched for "cold-hits" against evidence left at a crime scene; when geographic "hot spots" are identified for, say, some particularly unusual cancer; or when the whole human genome is searched for clues to common diseases), the same methods for assigning probabilities before the particular situation was identified are generally no longer appropriate after the fact.[1]

A second general area of inappropriate probability assessment con-

[1] A particularly problematic case of culling or locating "hot spots" is that of residential cancer-cluster identification. A readable account is by Atul Gawande, "The Cancer-Cluster Myth," *New Yorker*, February 8, 1999. For the probability issues that arise in searching the whole human genome for clues to some condition, see "Nabbing Suspicious SNPS: Scientists Search the Whole Genome for Clues to Common Diseases" (Regina Nuzzo, *ScienceNews*, June 21, 2008).

cerns the model postulated to aggregate probabilities over several events. Campbell (1974, p. 126) cites an article in the *New York Herald Tribune* (May, 1954) stating that if the probability of knocking down an attacking airplane were .15 at each of five defensive positions before reaching the target, then the probability of knocking down the plane before it passed all five barriers would be .75 ($5 \times .15$), this last value being the simple sum of the individual probabilities—and an inappropriate model. If we could correctly assume independence between the Bernoulli trials at each of the five positions, a more justifiable value would be one minus the probability of passing all barriers successfully: $1.0 - (.85)^5 \approx .56$. The use of similar binomial modeling possibilities, however, may be specious—for example, when dichotomous events occur simultaneously in groups (such as in the World Trade Center disaster on 9/11/01); when the success proportions are not valid; when the success proportions change in value over the course of the trials; or when time dependencies are present in the trials (such as in tracking observations above and below a median over time). In general, when wrong models are used to generate probabilities, the resulting values may have little to do with empirical reality. For instance, in throwing dice and counting the sum of spots that result, it is not true that each of the integers from two through twelve is equally likely. The model of what is equally likely may be reasonable at a different level (for example, pairs of integers appearing on the two dice), but not at all aggregated levels. There are some stories, probably apocryphal, of methodologists meeting their demises by making these mistakes for their gambling patrons.

Flawed calculations of probability can have dire consequences within our legal systems, as the case of Sally Clark and related others make clear. One broad and current area of possible misunderstanding of probabilities is in the context of DNA evidence (which is exacerbated in the older and more fallible system of identification through fingerprints).[2] In the use of DNA evidence (and with fingerprints), one must be concerned with the Random Match Probability (RMP): the likelihood that a randomly selected unrelated person from the population would match a given DNA profile. Again, the use of independence in RMP estimation is questionable; also, how does the RMP relate to, and is it relevant for, "cold-hit" searches in DNA databases. In a confirmatory identification case, a suspect is first identified by non-DNA evidence; DNA evidence is then used to corroborate traditional police investigation. In a "cold-hit" framework, the suspect is first identified by a search of DNA databases; the DNA evidence is thus used to identify the suspect as perpetrator, to the exclusion of others, directly from the outset (this is akin to shooting an arrow into a tree and then drawing a target around it). Here, traditional police work is no longer the focus. For a thorough discussion of the probabilistic context surrounding DNA evidence, which extends with even greater force to fingerprints, the article by Jonathan Koehler is recommended ("Error and Exaggeration in the Presentation of DNA Evidence at Trial," *Jurimetrics Journal*, *34*, 1993–1994, 21–39). We excerpt part of the introduction to this article below:

DNA identification evidence has been and will continue to be powerful evidence against criminal defendants. This is as it should be. In general, when blood, semen or hair that reportedly matches that of a defendant is found on

[2] Two informative articles on identification error using fingerprints ("Do Fingerprints Lie?", Michael Specter, *New Yorker*, May 27, 2002), and DNA ("You Think DNA Evidence is Foolproof? Try Again," Adam Liptak, *New York Times*, March 16, 2003).

or about a victim of violent crime, one's belief that the defendant committed the crime should increase, based on the following chain of reasoning:

Match Report ⇒ True Match ⇒ Source ⇒ Perpetrator

First a reported match is highly suggestive of a true match, although the two are not the same. Errors in the DNA typing process may occur, leading to a false match report. Second, a true DNA match usually provides strong evidence that the suspect who matches is indeed the source of the trace, although the match may be coincidental. Finally, a suspect who actually is the source of the trace may not be the perpetrator of the crime. The suspect may have left the trace innocently either before or after the crime was committed.

In general, the concerns that arise at each phase of the chain of inferences are cumulative. Thus, the degree of confidence one has that a suspect is the source of a recovered trace following a match report should be somewhat less than one's confidence that the reported match is a true match. Likewise, one's confidence that a suspect is the perpetrator of a crime should be less than one's confidence that the suspect is the source of the trace.

Unfortunately, many experts and attorneys not only fail to see the cumulative nature of the problems that can occur when moving along the inferential chain, but they frequently confuse the probabilistic estimates that are reached at one stage with estimates of the others. In many cases, the resulting misrepresentations and misinterpretation of these estimates lead to exaggerated expressions about the strength and implications of the DNA evidence. These exaggerations may have a significant impact on verdicts, possibly leading to convictions where acquittals might have been obtained.

This Article identifies some of the subtle, but common, exaggerations that have occurred at trial, and classifies each in relation to the three questions that are suggested by the chain of reasoning sketched above: (1) Is a reported match a true match? (2) Is the suspect the source of the trace? (3) Is the suspect the perpetrator of the crime? Part I addresses the first question and discusses ways of defining and estimating the false positive error rates at DNA laboratories. Parts II and III address the second and third questions, respectively. These sections introduce the "source probability error" and "ultimate issue error" and show how experts often commit these errors at

> trial with assistance from attorneys on *both* sides. (pp. 21–22)

In 1989, and based on urging from the FBI, the National Research Council (NRC) formed the Committee on DNA Technology in Forensic Science, which issued its report in 1992 (*DNA Technology in Forensic Science*; or more briefly, NRC I). The NRC I recommendation about the cold-hit process was as follows:

> The distinction between finding a match between an evidence sample and a suspect sample and finding a match between an evidence sample and one of many entries in a DNA profile databank is important. The chance of finding a match in the second case is considerably higher. . . . The initial match should be used as probable cause to obtain a blood sample from the suspect, but only the statistical frequency associated with the additional loci should be presented at trial (to prevent the selection bias that is inherent in searching a databank). (p. 124)

A follow-up report by a second NRC panel was published in 1996 (*The Evaluation of Forensic DNA Evidence*; or more briefly, NRC II), having the following main recommendation about cold-hit probabilities and using the "database match probability" or DMP:

> When the suspect is found by a search of DNA databases, the random-match probability should be multiplied by N, the number of persons in the database. (p. 161)

The term "database match probability" (DMP) is somewhat unfortunate. This is not a real probability but more of an expected number of matches given the RMP. A more legitimate value for the probability that another person matches the defendant's DNA profile would be $1 - (1 - \frac{1}{\text{RMP}})^N$, for a database of size N; that is, one minus the probability of no matches over N trials. For example, for an RMP of 1/1,000,000 and an N of 1,000,000, the above probability of another match is .632; the DMP (not a probability) number is 1.00, being the product of N and RMP. In any case, NRC II made the recommendation of using the DMP to give a measure of the accuracy of a cold-hit match, and did not support the more legitimate "probability of another match" using the formula given above (possibly because it was considered too difficult?):[3]

> A special circumstance arises when the suspect is identified not by an eyewitness or by circumstantial evidence but rather by a search through a large DNA database. If the only reason that the person becomes a suspect is that his DNA profile turned up in a database, the calculations must be modified. There are several approaches, of which we discuss two. The first, advocated by the 1992 NRC report, is to base probability calculations solely on loci not used in the search. That is a sound procedure, but it wastes information, and if too many loci are used for identification of the suspect, not enough might be left for an adequate subsequent analysis. . . . A second procedure is to apply a simple correction: Multiply the match probability by the size of the database searched. This is the procedure we recommend. (p. 32)

2 More on Bayes' Rule and the Confusion of Conditional Probabilities

The case of Sally Clark discussed in the introductory module and the commission of the prosecutor's fallacy that lead to her conviction is not an isolated occurrence. There was the recent miscarriage of justice in the Netherlands involving a nurse, Lucia de Berk, accused of

[3] As noted repeatedly by Gigerenzer and colleagues (e.g., Gigerenzer, 2002; Gigerenzer et al., 2007), it also may be best for purposes of clarity and understanding, to report probabilities using "natural frequencies." For example, instead of saying that a random match probability is .01, this could be restated alternatively that for this population, 1 out of every 10,000 men would be expected to show a match. The use of natural frequencies supposedly provides a concrete reference class for a given probability that then helps interpretation.

multiple deaths at the hospitals where she worked. This case aroused the international community of statisticians to redress the apparent injustices visited upon Lucia de Berk. One source for background, although now somewhat dated, is Mark Buchanan at the *New York Times* online opinion pages ("The Prosecutor's Fallacy," May 16, 2007). The Wikipedia article on Lucia de Berk provides the details of the case and the attendant probabilistic arguments, up to her complete exoneration in April 2010.

A much earlier and historically important *fin de siecle* case, is that of Alfred Dreyfus, the much maligned French Jew, and captain in the military, who was falsely imprisoned for espionage. In this case, the nefarious statistician was Alphonse Bertillon, who through a very convoluted argument reported a small probability that Dreyfus was "innocent." This meretricious probability had no justifiable mathematical basis and was generated from culling coincidences involving a document, the handwritten *bordereau* (without signature) announcing the transmission of French military information. Dreyfus was accused and convicted of penning this document and passing it to the (German) enemy. The "prosecutor's fallacy" was more or less invoked to ensure a conviction based on the fallacious small probability given by Bertillon. In addition to Émile Zola's well-known article, *J'accuse* … !, in the newspaper *L'Aurore* on January 13, 1898, it is interesting to note that turn-of-the-century well-known statisticians and probabilists from the French Academy of Sciences (among them Henri Poincaré) demolished Bertillon's probabilistic arguments, and insisted that any use of such evidence needs to proceed in a fully Bayesian manner, much like our present understanding of evidence in current forensic science and the proper place of probabilistic argumentation.[4]

We observe the same general pattern in all of the miscarriages

[4] By all accounts, Bertillon was a dislikable person. He is best known for the development of the first workable system of identification through body measurements; he named this "anthropometry" (later called "bertillonage" by others). We give a brief quotation about Bertillon from *The Science of Sherlock Holmes* by E. J. Wagner (2006):

> And then, in 1882, it all changed, thanks to a twenty-six-year old neurasthenic clerk in the Paris Police named Alphonse Bertillon. It is possible that Bertillon possessed some social graces, but if so, he was amazingly discreet about them. He rarely spoke, and when he did, his voice held no expression. He was bad-tempered and avoided people. He suffered from an intricate variety of digestive complaints, constant headaches, and frequent nosebleeds. He was narrow-minded and obsessive.
>
> Although he was the son of the famous physician and anthropologist Louis Adolphe Bertillon and had been raised in a highly intellectual atmosphere appreciative of science, he had managed to be thrown out of a number of excellent schools for poor grades. He had been unable to keep a job. His employment at the police department was due entirely to his father's influence. But this misanthropic soul managed to accomplish what no one else had: he invented a workable system of identification.
>
> Sherlock Holmes remarks in *The Hound of the Baskervilles*, "The world is full of obvious things which nobody by any chance ever observes." It was Bertillon who first observed the obvious need for a scientific method of identifying criminals. He recalled discussions in his father's house about the theory of the Belgian statistician Lambert Adolphe Jacques Quetelet, who in 1840 had suggested that there were no two people in the world who were exactly the same size in all their measurements. (pp. 97–98)

Bertillonage was widely used for criminal identification in the decades surrounding the turn-of-the-century. It was eventually supplanted by the use of fingerprints, as advocated by Sir Francis Galton in his book, *Finger Prints*, published in 1892. A short extraction from Galton's introduction mentions Bertillon by name:

> My attention was first drawn to the ridges in 1888 when preparing a lecture on Personal Identification for the Royal Institution, which had for its principal object an account of the anthropometric method of Bertillon, then newly introduced into the prison administration of France. Wishing to treat the subject generally, and having a vague knowledge of the value sometimes assigned to finger marks, I made inquiries, and was surprised to find, both how much had been done, and how much there remained to do, before establishing their theoretical value and practical utility.

One of the better known photographs of Galton (at age 73) is a Bertillon record from a visit Galton made to Bertillon's laboratory in 1893 (a Google search using the two words "Galton" and "Bertillon" will give the image).

of justice involving the prosecutor's fallacy. A very small reported probability of "innocence" is reported, typically obtained incorrectly either by culling, misapplying the notion of statistical independence, or using an inappropriate statistical model. This probability is calculated by a supposed expert with some credibility in court: Roy Meadow for Clark, Henk Elffers for de Berk, Alphonse Bertillon for Dreyfus. The prosecutor's fallacy then takes place, leading to a conviction for the crime. Various outrages ensue from the statistically literate community, with the eventual emergence of some "statistical good guys" hoping to redress the wrongs done: Richard Gill for de Berk, Henri Poincaré (among others) for Dreyfus, the Royal Statistical Society for Clark. After long periods of time, convictions are eventually overturned, typically after extensive prison sentences have already been served. We can only hope to avoid similar miscarriages of justice in cases yet to come by recognizing the tell-tale pattern of occurrences for the prosecutor's fallacy.

Any number of conditional probability confusions can arise in important contexts and possibly when least expected. A famous instance of such a confusion was in the O.J. Simpson case, where one conditional probability, say, $P(A|B)$, was equated with another, $P(A|B \text{ and } D)$. We quote the clear explanation of this obfuscation

Besides anthropometry, Bertillon contributed several other advances to what would now be referred to as "forensic science." He standardized the criminal "mug shot," and the criminal evidence picture through "metric photography." Metric photography involves taking pictures before a crime scene is disturbed; the photographs had mats printed with metric frames placed on the sides. As in "mug shots," photographs are generally taken of both the front and side views of a scene. Bertillon also created other forensic techniques, for example, forensic document examination (but in the case of Dreyfus, this did not lead to anything good), the use of galvanoplastic compounds to preserve footprints, the study of ballistics, and the dynamometer for determining the degree of force used in breaking and entering.

by Krämer and Gigerenzer (2005):

> Here is a more recent example from the U.S., where likewise $P(A|B)$ is confused with $P(A|B$ and $D)$. This time the confusion is spread by Alan Dershowitz, a renowned Harvard Law professor who advised the O.J. Simpson defense team. The prosecution had argued that Simpson's history of spousal abuse reflected a motive to kill, advancing the premise that "a slap is a prelude to homicide." Dershowitz, however, called this argument "a show of weakness" and said: "We knew that we could prove, if we had to, that an infinitesimal percentage—certainly fewer than 1 of 2,500—of men who slap or beat their domestic partners go on to murder them." Thus, he argued that the probability of the event K that a husband killed his wife if he battered her was small, $P(K|\text{battered}) = 1/2{,}500$. The relevant probability, however, is not this one, as Dershowitz would have us believe. Instead, the relevant probability is that of a man murdering his partner given that he battered her and that she was murdered, $P(K|\text{battered and murdered})$. This probability is about 8/9. It must of course not be confused with the probability that O.J. Simpson is guilty; a jury must take into account much more evidence than battering. But it shows that battering is a fairly good predictor of guilt for murder, contrary to Dershowitz's assertions. (p. 228)

Avoiding the prosecutor's fallacy is one obvious characteristic of correct probabilistic reasoning in legal proceedings. A related specious argument on the part of the defense is the "defendant's fallacy" (Committee on DNA Technology in Forensic Science, 1992, p. 31). Suppose that for an accused individual who is innocent, there is a one-in-a-million chance of a match (such as for DNA, blood, or fiber). In an area of, say, 10 million people, the number of matches expected is 10 even if everyone tested is innocent. The defendant's fallacy would be to say that because 10 matches are expected in a city of 10 million, the probability that the accused is innocent is 9/10. Because this latter probability is so high, the evidence of a match for the accused cannot be used to indicate a finding of guilt, and therefore, the evidence of a match should be excluded. A version of this fallacy appeared (yet again) in the O.J. Simpson murder trial; we give a short excerpt about the defendant's fallacy that is embedded in the Wikipedia article on the prosecutor's fallacy :

> A version of this fallacy arose in the context of the O.J. Simpson murder trial where the prosecution gave evidence that blood from the crime scene matched Simpson with characteristics shared by 1 in 400 people. The defense retorted that a football stadium could be filled full of people from Los Angeles who also fit the grouping characteristics of the blood sample, and therefore the evidence was useless. The first part of the defenses' argument that there are several other people that fit the blood grouping's characteristics is true, but what is important is that few of those people were related to the case, and even fewer had any motivation for committing the crime. Therefore, the defenses' claim that the evidence is useless is untrue.

We end this chapter with two additional fallacies involving conditional probabilities that were also reviewed by Krämer and Gigerenzer (2005). One will be called the facilitation fallacy, and the second, the category (mis)representation fallacy.

The facilitation fallacy argues that because a conditional probability, $P(B|A)$, is "large," the event B must therefore be facilitative for A (i.e., it must be true that $P(A|B) > P(A)$). As an example, suppose that among all people involved in an automobile accident, the majority are male; or, $P(\text{male}|\text{accident})$ is "large." But this does not imply that being male is facilitative of having an accident (i.e., it is not necessarily true that $P(\text{accident}|\text{male}) > P(\text{accident})$. There could be, for example, many more male drivers on the road than female drivers, and even though accident rates per mile may be the same for males and females, males will be in the majority when only those individuals involved in an accident are considered.

The category (mis)representation fallacy begins with the true observation that if B is facilitative of A, so that $P(A|B) > P(A)$, then $\bar{B}$ must be inhibitive of A; that is, $P(A|\bar{B}) < P(A)$. The fallacy is to then say that all subsets of $\bar{B}$ must also be inhibitive of A as well.

To paraphrase a hypothetical example given by Krämer and Gigerenzer(2005), suppose an employer hires 158 out of 1000 applicants (among the 1000, 200 are black, 200 are Hispanic, and 600 are white). Of the 158 new hires, 38 are non-white (36 are Hispanic and 2 are black), and 120 are white. Being white is facilitative of being hired:

$P(\text{hired}|\text{white}) = \frac{120}{600} = .20 > P(\text{hired}) = \frac{158}{1000} = .158$

And being nonwhite is inhibitive of being hired:

$P(\text{hired}|\text{nonwhite}) = \frac{38}{400} = .095 < P(\text{hired}) = .158$

But note that although being black is inhibitive of being hired:

$P(\text{hired}|\text{black}) = \frac{2}{200} = .01 < P(\text{hired}) = .158$,

the same is not true for the Hispanic subset:

$P(\text{hired}|\text{Hispanic}) = \frac{36}{200} = .18$ is greater than $P(\text{hired}) = .158$.

References

[1] Campbell, S. K. (1974). *Flaws and fallacies in statistical thinking*. Englewood Cliffs, NJ: Prentice-Hall.

[2] Freedman, D. A. (1983). A note on screening regression equations. *American Statistician*, *37*, 152–155.

[3] Gigerenzer, G. (2002). *Calculated risks: How to know when numbers deceive you*. New York: Simon & Schuster.

[4] Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

[5] Krämer, W., & Gigerenzer, G. (2005). How to confuse with statistics or: The use and misuse of conditional probabilities. *Statistical Science*, *20*, 223–230.

Module 8: Probabilistic Reasoning, Forensic Evidence, and the Relevance of Base Rates

The Gileadites seized the fords of the Jordan before the Ephraimites arrived. And when any Ephraimite who escaped said, “Let me cross over,” the men of Gilead would say to him, “Are you an Ephraimite?” If he said, “No,” then they would say to him, “Then say, ‘Shibboleth’!” And he would say, “Sibboleth,” for he could not pronounce it right. Then they would take him and kill him at the fords of the Jordan. There fell at that time forty-two thousand Ephraimites.
— Judges 12:5-6

Abstract: The topics developed in this module have at least a tacit connection to Bayes’ theorem, and specifically to how base rates operate formally in the use of Bayes’ theorem as well as more informally for several legally-related contexts. A number of topic areas are pursued: the general unreliability of eyewitness identification and testimony; polygraph testing; the assessment of blood alcohol level; the legal status and use of base rates; racial and ethnic profiling; false confessions; police interrogations; and the overall dismal state of the forensic “sciences.”

An earlier Module 4 discussed the relevance of base rates in the evaluation of diagnostic tests and did so in several important contexts. One involved the Meehl and Rosen (1955) notion of “clinical efficiency” where prediction with a diagnostic test could be shown to outperform prediction using simple base rates. A second was a critique of the area under a Receiver Operating Characteristic curve (the AUC) as the sole mechanism for evaluating how well a particular diagnostic test performs; in general, the AUC is independent of base rates and fails to assess how well a diagnostic instrument does in specific populations that have relatively low base rates for the characteristic to be detected. When base rates are equal, test sensitivity and the positive predictive value (PPV) are equal (and so are the negative predictive value (NPV) and test specificity). Because of these equivalences, simple functions of the PPV and NPV make sense in communicating just how well or how badly a diagnostic instrument performs.

Contents

1 Bayes' Rule and the Importance of Base Rates

In the formulation of Bayes' rule given in Module 1, the two prior probabilities, $P(A)$ and $P(B)$, are also known as "base rates"; that is, in the absence of other information, how often do the events A and B occur. Base rates are obviously important in the conversion of $P(B|A)$ into $P(A|B)$, but as shown by Tversky and Kahneman, and others (for example, Tversky and Kahneman, 1974), base rates are routinely ignored when using various reasoning heuristics. The example given Module 1 on the importance of base rates in eyewitness identification involved the classic blue and black taxi cab problem; the example was made-up for clarity, but the principle it illustrates has far-reaching real-world implications.

Some interesting commonalities are present across several forensic and medical domains where a knowledge of Bayes' theorem and the use of prior probabilities (or, base rates) may be crucial to the presentation of science-based recommendations, but which are then subsequently ignored (or discounted) by those very groups to which they are addressed. One area causing a great deal of controversy in the latter part of 2009 was the United States Preventive Services Task Force recommendations on cancer screening in women, particularly regarding when mammograms should start and their frequency. It is clear from the reactions in the media and elsewhere (for example, Congress), that irrespective of what may be reasonable science-based guidelines for women in general, on an individual level they will probably have no force whatsoever, despite recent reassuring results that targeted therapy is just as effective at saving lives without early detection.

Another arena in which Bayes' theorem has a role is in assessing and quantifying in a realistic way the probative (that is, legal-proof) value of eyewitness testimony. The faith the legal system has historically placed in eyewitnesses has been shaken by the advent of forensic DNA testing. In the majority of the numerous DNA exonerations occurring over the last twenty years, mistaken eyewitness identifications have been involved. A 2009 article by Wells, Memon, and Penrod ("Eyewitness Evidence: Improving Its Probative Value," in the series *Psychological Science in the Public Interest*), highlights the place that psychology and probabilistic reasoning have in this endeavor. We quote part of the abstract to give the flavor of the review:[1]

> Decades before the advent of forensic DNA testing, psychologists were questioning the validity of eyewitness reports. Hugo Münsterberg's writings in the early part of the 20th century made a strong case for the involvement of psychological science in helping the legal system understand the vagaries of eyewitness testimony. But it was not until the mid-to-late 1970s that psychologists began to conduct programmatic experiments aimed at understanding the extent of error and the variables that govern error when eyewitnesses give accounts of crimes they have witnessed. Many of the experiments conducted in the late 1970s and throughout the 1980s resulted in articles by psychologists that contained strong warnings to the legal system that eyewitness evidence was being overvalued by the justice system in the sense that its impact on triers of fact (e.g., juries) exceeded its probative (legal-proof) value. Another message of the research was that the validity of eyewitness reports depends a great deal on the procedures that are used to obtain those reports and that the legal system was not using the best procedures. (p. 45)

A third area in which Bayesian notions are crucial to an understanding of what is possible, is in polygraph examinations and the quality of information that they can or cannot provide. Again, what appears to happen is that people want desperately to believe in some rational mechanism for detecting liars and cheats, and thereby increase one's sense of security and control. So, irrespective of the statistical evidence marshalled, and probably because nothing else is really offered to provide even an illusion of control in identifying prevarication, lie detector tests still get done, and a lot of them. An illuminating tale is Fienberg and Stern's, "In Search of the Magic Lasso: The Truth About the Polygraph," (2005) and the work of the

[1] A very informative *New Yorker* article on eyewitness evidence is by Atul Gawande ("Under Suspicion," January 8, 2001). A more recent news item from *Nature*, concentrates specifically on how lines-ups are (ill)conducted: "Eyewitness Identification: Line-Ups on Trial" (*Nature*, Laura Spinney, May 21, 2008).

National Research Council Committee to Review the Scientific Evidence on the Polygraph (2003). We give the abstract of the Fienberg and Stern article below, followed by three telling paragraphs from their concluding section:[2]

In the wake of controversy over allegations of espionage by Wen Ho Lee, a nuclear scientist at the Department of Energy's Los Alamos National Laboratory, the department ordered that polygraph tests be given to scientists working in similar positions. Soon thereafter, at the request of Congress, the department asked the National Research Council (NRC) to conduct a thorough study of polygraph testing's ability to distinguish accurately between lying and truth-telling across a variety of settings and examinees, even in the face of countermeasures that may be employed to defeat the test. This paper tells some of the story of the work of the Committee to Review the Scientific Evidence on the Polygraph, its report and the reception of that report by the U.S. government and Congress. (p. 249)

At the outset, we explained the seemingly compelling desire for a device that can assist law enforcement and intelligence agencies to identify criminals,

[2] An interesting historical subplot in the development of lie detection involved William Moulton Marston. Marston is usually given credit for promoting the development of an instrument for lie detection based on systolic blood pressure. His doctoral dissertation in experimental psychology at Harvard (1921) was entitled *Systolic Blood Pressure and Reaction-Time Symptoms of Deception and of Constituent Mental States*. It has been suggested (by none other than Marston's son) that it was actually Elizabeth Marston, William's wife, who was the motivation for his work on lie detection and its relation to blood pressure (quoting the son, "when she got mad or excited, her blood pressure seemed to climb"). In any case, Marston lived with two women in a polyamorous relationship—Elizabeth Holloway Marston, his wife, and Olive Byrne. Both these two women served as exemplars and inspirations for Marston's more well-known contribution to American life—the creation of the character and comic strip, *Wonder Women*, in the early 1940s under the pseudonym of Charles Moulton. Supposedly, it was Elizabeth's idea to create a female superhero who could triumph not with fists or firepower, but with love. This character would have a Magic Lasso (or a Golden Lasso, or a Lasso of Truth) that would force anyone captured by it to obey and tell the truth. So, besides introducing Wonder Woman and a lie detection instrument to a United States audience, Marston is credited with several additional cultural introductions. For more detail the reader is referred to the Wikipedia article on Marston.

> spies and saboteurs when direct evidence is lacking. The polygraph has long been touted as such a device. In this article and in the NRC report on which it draws, we explain the limited scientific basis for its use, the deep uncertainty about its level of accuracy and the fragility of the evidence supporting claims of accuracy in any realistic application.
>
> How should society, and the courts in particular, react to such a situation? At a minimum they should be wary about the claimed validity of the polygraph and its alternatives for use in the myriad settings in which they are used or proposed for use. This is especially relevant to current forensic uses of the polygraph. We believe that the courts have been justified in casting a skeptical eye on the relevance and suitability of polygraph test results as legal evidence. Generalizing from the available scientific evidence to the circumstances of a particular polygraph examination is fraught with difficulty. Further, the courts should extend their reluctance to rely upon the polygraph to the many quasiforensic uses that are emerging, such as in sex offender management programs. The courts and the legal system should not act as if there is a scientific basis for many, if any, of these uses. They need to hear the truth about lie detection.
>
> As this paper was going to press in January 2005, the Department of Energy finally announced its proposed revised polygraph rules in the Federal Register. They provide a detailed plan for implementing the plan outlined in Deputy Secretary McSlarrow's September 2003 testimony. [Note: This was to only do 4,500 lie detector tests rather than the usual 20,000.] But no other federal agency has stepped forward with a plan to curb the use of polygraphs. All of them have heard the truth about polygraphs as we know it, but they have failed to acknowledge it by action. (p. 259)

We mention one last topic where a knowledge of Bayes' rule might help in arguing within another arena of forensic evidence: the assessment of blood alcohol content (BAC). The United States Supreme Court heard arguments in January of 2010 (Briscoe v. Virginia, 2010) about crime analysts being required to make court appearances, and to (presumably) testify about the evidence and its reliability that they present now only in written form. The case was spurred in part by a California woman convicted of vehicular manslaughter with a supposed blood alcohol level two hours after the accident above the legal limit of .08. The woman denied being drunk but did admit to taking two shots of tequila (with Sprite chasers).[3]

There are several statistically related questions pertaining to the use of a dichotomous standard for BAC (usually, .08) as a definitive indication of impairment and, presumably, of criminal liability when someone is injured in an accident. Intuitively, it would seem that the same level of BAC might lead to different levels of impairment conditional on individual characteristics. Also, was this value set based on scientifically credible data? A variety of different BAC tests

[3] The woman's name is Virginia Hernandez Lopez; see, for example, Adam Liptak, *New York Times* (December 19, 2009), "Justices Revisit Rule Requiring Lab Testimony." In the actual case being orally argued of *Briscoe v. Virginia* (2010), the Court merely sent it back to a lower court in light of a recently decided case (*Melendez-Diaz v. Massachusetts* (2009)), which held that it is unconstitutional for a prosecutor to submit a chemical drug test report without the testimony of the scientist who conducted the test.

A more recent (5-4) Supreme Court ruling in *Bullcoming v. New Mexico* (2011) reaffirmed the *Melendez-Diaz* decision, saying that "surrogate testimony" would not suffice, and substitutes were not acceptable in crime lab testimony. The first paragraph of the syllabus in the *Bullcoming* opinion follows:

> The Sixth Amendment's Confrontation Clause gives the accused "[i]n all criminal prosecutions … the right … to be confronted with the witnesses against him." In Crawford v. Washington … this Court held that the Clause permits admission of "[t]estimonial statements of witnesses absent from trial … only where the declarant is unavailable, and only where the defendant has had a prior opportunity to cross-examine." Later, in Melendez-Diaz v. Massachusetts … the Court declined to create a "forensic evidence" exception to Crawford, holding that a forensic laboratory report, created specifically to serve as evidence in a criminal proceeding, ranked as "testimonial" for Confrontation Clause purposes. Absent stipulation, the Court ruled, the prosecution may not introduce such a report without offering a live witness competent to testify to the truth of the report's statements.

could be used (for example, urine, blood, saliva, breath, hair); thus, there are all the possible interchangeability and differential reliability issues that this multiplicity implies.

The two most common alternatives to the supposedly most accurate blood test are based on urine and breath. Urine tests indicate the presence of alcohol in a person's system, but it takes up to two hours for the alcohol to show up. A positive urine test does not necessarily mean the person was under the influence of alcohol at the time of the test. Rather, it detects and measures usage within the last several days. Breath alcohol does not directly measure BAC but the amount of supposed "alcohol" in one's breath (as well as all chemically similar compounds and extraneous material such as vomit), and can be influenced by many external factors—cell phones, gasoline, blood, exercise, holding one's breath, and so on. We point the reader to an entry, "Blood Alcohol Testing in Drunk Driving Cases," posted by a lawyer, Aaron Larson, on the "expertlaw.com" website (2000).

A knowledge of Bayes' theorem and the way in which sensitivity, specificity, the positive predictive value, and the prior probability all operate together may at times be helpful to you or to others in mitigating the effects that a single test may have on one's assessment of culpability. There are many instances where the error rates associated with an instrument are discounted, and it is implicity assumed that an "observed value" is the "true value." The example of blood alcohol level just discussed seems to be, on the face of it, a particularly egregious example. But there are other tests that could be usefully approached with an understanding of Bayes' rule, such as drug/steroid/human growth hormone use in athletes, blood doping in bicycle racers, polygraph tests for spying/white collar crime, fingerprint or eyewitness (mis)identification, or laser gun usage for speeding tickets. We are not saying that a savvy statistician armed with a knowledge of how Bayes' theorem works can "beat the rap," but it couldn't hurt. Anytime a judgment is based on a single fallible instrument, the value of the positive predictive value assumes a great importance in establishing guilt or innocence.[4]

[4] We point to two items regarding lie detection that are relevant to making judgments based on a fallible instrument. One is by Margaret Talbot on using brain scans to detect lying ("Duped: Can Brain Scans Uncover Lies?," *New Yorker*, July 2, 2007); the other debunks voice-based lie detection: "The Truth Hurts: Scientists Question Voice-Based Lie Detection" (Rachel Ehrenberg, *ScienceNews*, June 22, 2010). A more general review devoted to lie detection by Vrij, Granhag, and Porter, appeared in *Psychological Science in the Public Interest* ("Pitfalls and Opportunities in Nonverbal and Verbal Lie Detection," 2010, *11*, 89–121). This article discusses behaviors that are not the best diagnostic indicators of lying. The term "illusory correlation," refers to a false but widely held belief in a relationship between two behaviors, for example, the drawing of big eyes in a Draw-A-Person projective test and a person's paranoia. In lying, there are the two illusory correlations of gaze aversion and nervousness.

The notion that gaze aversion reflects lying appears in our common idiomatic language in phrases such as "he won't look me in the eye." An editorial accompanying the review article cited above (Elizabeth Loftus, "Catching Liars," 2010, *11*, 87–88), comments directly on the cross-racial problem of using gaze aversion to suggest someone is lying:

> Using gaze aversion to decide that someone is lying can be dangerous for that someone's health and happiness. And—what was news to me—some cultural or ethnic groups are more likely to show gaze aversion. For example, Blacks are particularly likely to show gaze aversion. So imagine now the problem that might arise when a White police officer interviews a Black suspect and interprets the gaze aversion as evidence of lying. This material needs to be put in the hands of interviewers to prevent this kind of cross-racial misinterpretation. (p. 87)

Coupled with a human tendency to engage in confirmation bias when an illusory correlation is believed, and to look for evidence of some type of "tell" such as "gaze aversion," we might once again remind ourselves to "lawyer up" early and often.

The illusory connection between nervousness and lying is so strong it has been given the name of "the Othello error." A passage from the Vrij et al. (2010) review provides a definition:

> A common error in lie detection is to too readily interpret certain behaviors, particularly signs of nervousness, as diagnostic of deception. A common mistake for lie detectors is the failure to consider that truth tellers (e.g., an innocent suspect or defendant) can be as nervous as liars. Truth tellers can be nervous as a result of being accused of wrongdoing or as a result of fear of not being believed, because they too could face negative consequences if they are not believed ... The misinterpretation of signs of nervousness in truth tellers as signs of deceit is referred to as the Othello error by deception researchers ... based on Shakespeare's character. Othello falsely accuses his wife Desdemona of infidelity, and he tells her to confess

1.1 The (Legal) Status of the Use of Base rates

The use of base rates in the context of various legal proceedings, criminal matters, and subject identification has been problematic. The quotation that just opened this chapter shows the historical range for which base rates have come into consideration in a variety of (quasi-)legal settings. This section reviews several of these areas in more detail.

Shibboleth: This word comes directly from the Old Testament Biblical quotation (Judges 12:5-6) regarding the Gileadites and Ephraimites. It refers to any distinguishing practice, usually one of language, associated with social or regional origin that identifies its speaker as being a member of a group. There are a number of famous shibboleths: German spies during World War II mispronounced the initial "sch" in the Dutch port city's name of Scheveningen (and thereby could be "caught"); during the Battle of the Bulge, American soldiers used knowledge of baseball to tell whether there were German infiltrators in American uniforms; United States soldiers in the Pacific used the word "lollapalooza" to identify the Japanese enemy because a repeat of the word would come back with a beginning pronunciation of "rorra."[5]

because he is going to kill her for her treachery. When Desdemona asks Othello to summon Cassio (her alleged lover) so that he can testify her innocence, Othello tells her that he has already murdered Cassio. Realizing that she cannot prove her innocence, Desdemona reacts with an emotional outburst, which Othello misinterprets as a sign of her infidelity. The Othello error is particularly problematic in attempting to identify high-stakes lies because of the observer's sense of urgency and a host of powerful cognitive biases that contribute to tunnel-vision decision making ... (p. 98)

[5] Or, asking a person to say "rabbit" to see if he is Elmer Fudd.

Criminal trials: As noted in Module 1, Rule 403 of the *Federal Rules of Evidence* explicitly disallows the introduction of base rate information that would be more prejudicial than having value as legal proof. For instance, base rate information about which demographic groups commit which crimes and which don't would not be admissible under Rule 403. Although Rule 403 was given in MOdule 1, it is repeated below for completeness of the present discussion.

> Rule 403. Exclusion of Relevant Evidence on Grounds of Prejudice, Confusion, or Waste of Time: Although relevant, evidence may be excluded if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury, or by considerations of undue delay, waste of time, or needless presentation of cumulative evidence.

Racial profiling: Although the Arizona governor, Jan Brewer, vehemently denied the label of racial profiling attached to its Senate Bill 1070, her argument comes down to officers knowing an illegal alien when they see one, and this will never depend on racial profiling because that, she says, "is illegal." How an assessment of "reasonable suspicion" would be made is left to the discretion of the officers—possibly a shibboleth will be used, such as speaking perfect English without an accent (or as the then governor of the state adjoining Arizona (Arnold Schwarzenegger) said: "I was also going to go and give a speech in Arizona but with my accent, I was afraid they were going to deport me back to Austria."). The reader is referred to the *New York Times* article by Randal C. Archibold ("Arizona Enacts Stringent Law on Immigration," April 23, 2010) that states succinctly the issues involved in Arizona's "Papers, Please" law.[6]

[6] As discussed in training videos for Arizona law-enforcement personnel, police can consider a variety of characteristics in deciding whether to ask about an individual's immigration

Constitutional protections: Two constitutional amendments protect the rights of individuals residing in the United States. The first amendment discussed is the Fourteenth, with its three operative clauses:

— The Citizenship Clause provides a broad definition of citizenship, overruling the decision in *Scott v. Sandford* (1857), which held that blacks could not be citizens of the United States. Those who follow current politics might note that this clause makes anyone born in the United States a citizen. Calls for its repeal are heard routinely from the political right, with the usual laments about "tourism babies," or those born to illegal immigrants. Irrespective of the citizenship of the parents, a baby born to someone temporarily in the United States is a United States citizen by default, and therefore, under all the protections of its laws.

— The Due Process Clause prohibits state and local governments from depriving persons of life, liberty, or property without steps being taken to insure fairness.

— The Equal Protection Clause requires the States to provide equal protection under the law to all people within its jurisdiction. This was the basis for the unanimous opinion in the famous *Brown v. Board of Education* (1954).

These three clauses are part of only one section of the Fourteenth

status: does the person speak poor English, look nervous, is he traveling in an overcrowded vehicle, wearing several layers of clothing in a hot climate, hanging out in areas where illegal immigrants look for work, does not have identification, does he try to run away, ... See Amanda Lee Myers, "Seventh Lawsuit Filed Over Ariz. Immigration Law" (Associated Press, July 10, 2010). It is difficult to see how any convincing statistical argument could be formulated that the use of behaviors correlated with ethnicity and race does not provide a *prima facie* case for racial profiling.

Amendment, which follows:

> Section 1. All persons born or naturalized in the United States, and subject to the jurisdiction thereof, are citizens of the United States and of the State wherein they reside. 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.

Although the "due process" and "equal protection" clauses seem rather definitive, the United States judicial system has found ways to circumvent their application when it was viewed necessary. One example discussed fully in Module 3 is the Supreme Court decision in *McCleskey v. Kemp* (1987) on racial disparities in the imposition of the death penalty (in Georgia). But probably the most blatant disregard of "equal protection" was the Japanese-American internment and relocation of about 110,000 individuals living along the United States Pacific coast in the 1940s. These "War Relocation Camps" were authorized by President Roosevelt on February 19, 1942, with the infamous *Executive Order 9066*. The Supreme Court opinion (6 to 3) in *Korematsu v. United States* (1944) upheld the constitutionality of *Executive Order 9066*. The majority opinion written by Hugo Black argued that the need to protect against espionage outweighed Fred Korematsu's individual rights and the rights of Americans of Japanese descent. In dissent, Justices Robert Jackson and Frank Murphy commented about both the bad precedent this opinion set and the racial issues it presented. We quote part of these two dissenting opinions:

> Murphy: I dissent, therefore, from this legalization of racism. Racial dis-

crimination in any form and in any degree has no justifiable part whatever in our democratic way of life. It is unattractive in any setting, but it is utterly revolting among a free people who have embraced the principles set forth in the Constitution of the United States. All residents of this nation are kin in some way by blood or culture to a foreign land. Yet they are primarily and necessarily a part of the new and distinct civilization of the United States. They must, accordingly, be treated at all times as the heirs of the American experiment, and as entitled to all the rights and freedoms guaranteed by the Constitution.

Jackson: A military order, however unconstitutional, is not apt to last longer than the military emergency. Even during that period, a succeeding commander may revoke it all. But once a judicial opinion rationalizes such an order to show that it conforms to the Constitution, or rather rationalizes the Constitution to show that the Constitution sanctions such an order, the Court for all time has validated the principle of racial discrimination in criminal procedure and of transplanting American citizens. The principle then lies about like a loaded weapon, ready for the hand of any authority that can bring forward a plausible claim of an urgent need. Every repetition imbeds that principle more deeply in our law and thinking and expands it to new purposes.

. . .

Korematsu was born on our soil, of parents born in Japan. The Constitution makes him a citizen of the United States by nativity and a citizen of California by residence. No claim is made that he is not loyal to this country. There is no suggestion that apart from the matter involved here he is not law abiding and well disposed. Korematsu, however, has been convicted of an act not commonly a crime. It consists merely of being present in the state whereof he is a citizen, near the place where he was born, and where all his life he has lived. . . . [H]is crime would result, not from anything he did, said, or thought, different than they, but only in that he was born of different racial stock. Now, if any fundamental assumption underlies our system, it is that guilt is personal and not inheritable. Even if all of one's antecedents had been convicted of treason, the Constitution forbids its penalties to be visited upon him. But here is an attempt to make an otherwise innocent act

> a crime merely because this prisoner is the son of parents as to whom he had no choice, and belongs to a race from which there is no way to resign. If Congress in peace-time legislation should enact such a criminal law, I should suppose this Court would refuse to enforce it.

Congress passed and President Reagan signed legislation in 1988 apologizing for the internment on behalf of the United States government. The legislation noted that the actions were based on "race prejudice, war hysteria, and a failure of political leadership." Over $1.6 billion was eventually dispersed in reparations to the interned Japanese-Americans and their heirs.

The other main amendment that has an explicit rights protection as its focus is the Fourth (from the Bill of Rights); its purpose is to guard against unreasonable searches and seizures, and to require a warrant to be judicially sanctioned and supported by "probable cause." The text of the amendment follows:

> The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

Various interpretations of the Fourth Amendment have been made through many Supreme Court opinions. We mention two here that are directly relevant to the issue of law-enforcement application of base rates, and for (racial) profiling: *Terry v. Ohio* (1968) and *Whren v. United States* (1996). The Wikipedia summaries are given in both cases:

> *Terry v. Ohio* . . . (1968) was a decision by the United States Supreme Court which held that the Fourth Amendment prohibition on unreasonable searches

> and seizures is not violated when a police officer stops a suspect on the street and frisks him without probable cause to arrest, if the police officer has a reasonable suspicion that the person has committed, is committing, or is about to commit a crime and has a reasonable belief that the person “may be armed and presently dangerous.” . . .
>
> For their own protection, police may perform a quick surface search of the person’s outer clothing for weapons if they have reasonable suspicion that the person stopped is armed. This reasonable suspicion must be based on “specific and articulable facts” and not merely upon an officer’s hunch. This permitted police action has subsequently been referred to in short as a “stop and frisk,” or simply a “Terry stop.”

> *Whren v. United States* . . . (1996) was a United States Supreme Court decision which “declared that any traffic offense committed by a driver was a legitimate legal basis for a stop,” [and] . . . “the temporary detention of a motorist upon probable cause to believe that he has violated the traffic laws does not violate the Fourth Amendment’s prohibition against unreasonable seizures, even if a reasonable officer would not have stopped the motorist absent some additional law enforcement objective.”

In a dissenting opinion in *Terry v. Ohio* (1968), Justice William O. Douglas strongly disagreed with permitting a stop and search without probable cause:

> I agree that petitioner was “seized” within the meaning of the Fourth Amendment. I also agree that frisking petitioner and his companions for guns was a “search.” But it is a mystery how that “search” and that “seizure” can be constitutional by Fourth Amendment standards, unless there was “probable cause” to believe that (1) a crime had been committed or (2) a crime was in the process of being committed or (3) a crime was about to be committed.
>
> The opinion of the Court disclaims the existence of “probable cause.” If loitering were in issue and that was the offense charged, there would be “probable cause” shown. But the crime here is carrying concealed weapons; and there is no basis for concluding that the officer had “probable cause” for believing that that crime was being committed. Had a warrant been sought,

a magistrate would, therefore, have been unauthorized to issue one, for he can act only if there is a showing of "probable cause." We hold today that the police have greater authority to make a "seizure" and conduct a "search" than a judge has to authorize such action. We have said precisely the opposite over and over again.

. . .

There have been powerful hydraulic pressures throughout our history that bear heavily on the Court to water down constitutional guarantees and give the police the upper hand. That hydraulic pressure has probably never been greater than it is today.

Yet if the individual is no longer to be sovereign, if the police can pick him up whenever they do not like the cut of his jib, if they can "seize" and "search" him in their discretion, we enter a new regime. The decision to enter it should be made only after a full debate by the people of this country.

The issues of racial profiling and policies of "stop-question-and-frisk" are ongoing and particularly divisive in big urban areas such as New York City. To get a sense of this continuing controversy, the reader is referred to the *New York Times* article by Al Baker and Ray Rivera (October 26, 2010), "Study Finds Street Stops by N.Y. Police Unjustified." Several excerpts from this article follow that should illustrate well the contentiousness of the "stop-question-and-frisk" policies of the New York City Police Department.

The study was conducted on behalf of the Center for Constitutional Rights, which is suing the New York Police Department for what the center says is a widespread pattern of unprovoked and unnecessary stops and racial profiling in the department's stop-question-and-frisk policy. The department denies the charges.

...

Police Commissioner Raymond W. Kelly has rejected the accusation of racial profiling, and said the racial breakdown of the stops correlated to the racial breakdown of crime suspects. Mr. Kelly has also credited the tactic with helping to cut crime to low levels in the city and with getting guns off the street.

...

The United States Supreme Court has held that in order for police officers to stop someone, they must be able to articulate a reasonable suspicion of a crime. To frisk them, they must have a reasonable belief that the person is armed and dangerous.

Darius Charney, a lawyer for the Center for Constitutional Rights, said the study crystallized the primary complaints in the lawsuit. "It confirms what we have been saying for the last 10 or 11 years, which is that stop-and-frisk patterns – it is really race, not crime, that is driving this," Mr. Charney said.

Mr. Kelly, responding to the professor's study, said on Tuesday, "I think you have to understand this was an advocacy paper." He also noted that Professor Fagan was paid well to produce the report.

Government institution protections: Although government institutions should protect rights guaranteed by the Constitution, there have been many historical failures. Many of these (unethical) intrusions are statistical at their core, where data are collected on individuals who may be under surveillance only for having unpopular views. To give a particularly salient and egregious example involving the FBI, J. Edgar Hoover, Japanese-American internment, and related topics, we redact the Wikipedia entry on the Custodial Detention Index (under the main heading of "FBI Index") used by the FBI from the 1930s to the 1970s (with various renamed successor indices, such as Rabble-Rouser, Agitator, Security, Communist, Administrative):

The Custodial Detention Index (CDI), or Custodial Detention List was formed in 1939-1941, in the frame of a program called variously the "Custodial Detention Program" or "Alien Enemy Control."

J. Edgar Hoover described it as having come from his resurrected General Intelligence Division in Washington:

"This division has now compiled extensive indices of individuals, groups, and organizations engaged in subversive activities, in espionage activities, or any activities that are possibly detrimental to the internal security of the United States. The Indexes have been arranged not only alphabetically but also geographically, so that at any rate, should we enter into the conflict abroad, we would be able to go into any of these communities and identify individuals or groups who might be a source of grave danger to the security of this country. These indexes will be extremely important and valuable in a grave emergency."

Congressmen Vito Marcantonio called it "terror by index cards." ...

The Custodial Detention Index was a list of suspects and potential subversives, classified as "A," "B," and "C"; the ones classified as "A" were destined to be immediately arrested and interned at the outbreak of war. Category A were leaders of Axis-related organizations, category B were members deemed "less dangerous" and category C were sympathizers. The actual assignment of the categories was, however, based on the perceived individual commitment to the person's native country, rather than the actual potential to cause harm; leaders of cultural organizations could be classified as "A," members of non-Nazi and pro-Fascist organizations.

The program involved creation of individual dossiers from secretly obtained information, including unsubstantiated data and in some cases, even hearsay and unsolicited phone tips, and information acquired without judicial warrants by mail covers and interception of mail, wiretaps and covert searches. While the program targeted primarily Japanese, Italian, and German "enemy aliens," it also included some American citizens. The program was run without Congress-approved legal authority, no judicial oversight and outside of the official legal boundaries of the FBI. A person against which an accusation was made was investigated and eventually placed on the index; it was not removed until the person died. Getting on the list was easy; getting off of it was virtually impossible.

According to the press releases at the beginning of the war, one of the purposes of the program was to demonstrate the diligence and vigilance of the government by following, arresting and isolating a previously identified group of people with allegedly documented sympathies for Axis powers and potential for espionage or fifth column activities. The list was later used for Japanese-American internment.

Attorney General Francis Biddle, when he found out about the Index, labeled it "dangerous, illegal" and ordered its end. However, J. Edgar Hoover simply renamed it the Security Index, and told his people not to mention it.

USA PATRIOT Act: The attitude present during World War II that resident Japanese-Americans had a proclivity for espionage has now changed after September 11, 2001, to that of Middle Eastern men having a proclivity for committing terrorist acts. The acronym of being arrested because of a DWB ("driving while black") has now been altered to FWM ("flying while Muslim"). Section 412 of the *USA PATRIOT Act* allows the United States Attorney General to detain aliens for up to seven days without bringing charges when the detainees are certified as threats to national security. The grounds for detention are the same "reasonable suspicion" standard of *Terry v. Ohio* (1968). The Attorney General certification must state that there are "reasonable grounds to believe" the detainee will commit espionage or sabotage, commit terrorist acts, try to overthrow the government, or otherwise behave in a way that would endanger national security. After seven days, the detention may continue if the alien is charged with a crime or violation of visa conditions. When circumstances prohibit the repatriation of a person for an immigration offense, the detention may continue indefinitely if recertified by the attorney general every six months. Under the *USA PATRIOT Act*, a person confined for a violation of conditions of United States entry but who cannot be deported to the country of origin, may be indefinitely confined without criminal charges ever being filed.

Profiling, ethnic or otherwise, has been an implicit feature of

United States society for some time. The particular targets change, but the idea that it is permissible to act against specific individuals because of group membership does not. In the 1950s there were popular radio and television programs, such as *The FBI in Peace and War* or *I Led 3 Lives* about the double-agent Herbert Philbrick. These all focused on the Red menace in our midst, bent on overthrowing our form of government. It is instructive to remember our history whenever a new group is targeted for surveillance, and to note that the *Smith Act of 1940* (also known as the *Alien Registration Act*) is still on the books; the enabling "membership clause" and other conditions in the *Smith Act* follow:

> Whoever knowingly or willfully advocates, abets, advises, or teaches the duty, necessity, desirability, or propriety of overthrowing or destroying the government of the United States or the government of any State, Territory, District or Possession thereof, or the government of any political subdivision therein, by force or violence, or by the assassination of any officer of any such government; or
>
> Whoever, with intent to cause the overthrow or destruction of any such government, prints, publishes, edits, issues, circulates, sells, distributes, or publicly displays any written or printed matter advocating, advising, or teaching the duty, necessity, desirability, or propriety of overthrowing or destroying any government in the United States by force or violence, or attempts to do so; or
>
> Whoever organizes or helps or attempts to organize any society, group, or assembly of persons who teach, advocate, or encourage the overthrow or destruction of any such government by force or violence; or becomes or is a member of, or affiliates with, any such society, group, or assembly of persons, knowing the purposes thereof
>
> Shall be fined under this title or imprisoned not more than twenty years, or both, and shall be ineligible for employment by the United States or any department or agency thereof, for the five years next following his conviction.

If two or more persons conspire to commit any offense named in this section, each shall be fined under this title or imprisoned not more than twenty years, or both, and shall be ineligible for employment by the United States or any department or agency thereof, for the five years next following his conviction.

As used in this section, the terms "organizes" and "organize," with respect to any society, group, or assembly of persons, include the recruiting of new members, the forming of new units, and the regrouping or expansion of existing clubs, classes, and other units of such society, group, or assembly of persons.

Eyewitness reliability and false confessions: Several troublesome forensic areas exist in which base rates can come into nefarious play. One is in eyewitness testimony and how base rates are crucial to assessing the reliability of a witness's identification. The criminal case of "In Re As.H (2004)" reported in Module 9 illustrates this point well, particularly as it deals with cross-racial identification, memory lapses, how lineups are done, and so forth. Also, we have the earlier taxicab anecdote of Module 1. One possibly unexpected use that we turn to next involves base rate considerations in "false confessions." False confessions appear more frequently than we might expect and also in some very high profile cases. The most sensationally reported example may be the Central Park jogger incident of 1989, in which five African and Hispanic Americans all falsely confessed. To give a better sense of the problem, a short abstract is given below from an informative review article by Saul Kassin in the *American Psychologist* (2005, *60*, 215–228), entitled "On the Psychology of Confessions: Does Innocence Put Innocents at Risk":

The Central Park jogger case and other recent exonerations highlight the problem of wrongful convictions, 15% to 25% of which have contained con-

> fessions in evidence. Recent research suggests that actual innocence does not protect people across a sequence of pivotal decisions: (a) In preinterrogation interviews, investigators commit false-positive errors, presuming innocent suspects guilty; (b) naively believing in the transparency of their innocence, innocent suspects waive their rights; (c) despite or because of their denials, innocent suspects elicit highly confrontational interrogations; (d) certain commonly used techniques lead suspects to confess to crimes they did not commit; and (e) police and others cannot distinguish between uncorroborated true and false confessions. It appears that innocence puts innocents at risk, that consideration should be given to reforming current practices, and that a policy of videotaping interrogations is a necessary means of protection. (p. 215)

To put this issue of false confession into a Bayesian framework, our main interest is in the term, $P(\text{guilty} \mid \text{confess})$. Based on Bayes' rule this probability can be written as

$$\frac{P(\text{confess} \mid \text{guilty})P(\text{guilty})}{P(\text{confess} \mid \text{guilty})P(\text{guilty}) + P(\text{confess} \mid \text{not guilty})P(\text{not guilty})} .$$

The most common interrogation strategy taught to police officers is the 9-step Reid Technique.[7] The proponents of the Reid Technique hold two beliefs: that $P(\text{confess} \mid \text{not guilty})$ is zero, and that they never interrogate innocent people, so the prior probability, $P(\text{guilty})$, is 1.0. Given these assumptions, it follows that if a confession is given, the party must be guilty. There is no room for error in the Reid system; also, training in the Reid system does not increase accuracy of an initial prior assessment of guilt but it does greatly increase confidence in that estimate. We thus have a new wording for an old adage: "never in error and never in doubt."

[7] A discussion of how police interrogation operates was written (and available online) by Julia Layton (May 18, 2006), "How Police Interrogation Works."

A number of psychological concerns are present with how interrogations are done in the United States. Innocent people are more likely to waive their *Miranda* rights (so unfortunately, they can then be subjected to interrogation); but somehow this does not seem to change an interrogator's prior probability of guilt.[8] People have a naive faith in the power of their own innocence to set them free. They maintain a belief in a just world where people get what they deserve and deserve what they get. People are generally under an illusion of transparency where they overestimate the extent that others can see their true thoughts. When in doubt, just remember the simple words—"I want a lawyer." (Or, in the idiom of the *Law & Order* series on TV, always remember to "lawyer-up.") If an interrogation proceeds (against our recommendation), it is a guilt-presumptive process that unfolds (it is assumed from the outset that $P(\text{guilty})$ is 1.0). False incriminating evidence can be presented to you (in contrast to the U.K, which is surprising because the United Kingdom doesn't have a "Bill of Rights"). Some people who are faced with false evidence may even begin to believe they are guilty. The interrogation process is one of social influence, with all the good cards stacked on one side of the table. It does not even have to be videotaped, so any post-confession argument of psychological coercion is hard to make.

As part of our advice to "lawyer up" if you happen to find yourself in a situation where you could be subjected to interrogation (and regardless of whether you believe yourself to be innocent or not),

[8] A minimal statement of a Miranda warning is given in the Supreme Court case of *Miranda v. Arizona* (1966): "You have the right to remain silent. Anything you say can and will be used against you in a court of law. You have the right to speak to an attorney, and to have an attorney present during any questioning. If you cannot afford a lawyer, one will be provided for you at government expense."

there is now a further need to be verbally obvious about invoking one's *Miranda* rights—counterintuitively, you have to be clear and audible in your wish not to talk. The Supreme Court issued the relevant ruling in June 2010. An article reviewing the decision from the *Los Angeles Times* (David G. Savage, "Supreme Court Backs Off Strict Enforcement of Miranda Rights," June 2, 2010) provides a cautionary piece of advice for those of us who might someday fall into the clutches of the criminal system through no fault of our own.

1.2 Forensic Evidence Generally

Most of us learn about forensic evidence and how it is used in criminal cases through shows such as *Law & Order*. Rarely, if ever, do we learn about evidence fallibility and whether it can be evaluated through the various concepts introduced to this point, such as base rates, sensitivity, specificity, prosecutor or defendant fallacy, or the positive predictive value. Contrary to what we may come to believe, evidence based on things such as bite marks, fibers, and voice prints are very dubious. As one example, we give the conclusion of a conference presentation by Jean-François Bonastre and colleagues (2003), entitled "Person Authentication by Voice: A Need for Caution":

> Currently, it is not possible to completely determine whether the similarity between two recordings is due to the speaker or to other factors, especially when: (a) the speaker does not cooperate, (b) there is no control over recording equipment, (c) recording conditions are not known, (d) one does not know whether the voice was disguised and, to a lesser extent, (e) the linguistic content of the message is not controlled. Caution and judgment must be exercised when applying speaker recognition techniques, whether human or automatic, to account for these uncontrolled factors. Under more constrained or calibrated situations, or as an aid for investigative purposes,

> judicious application of these techniques may be suitable, provided they are not considered as infallible.
>
> At the present time, there is no scientific process that enables one to uniquely characterize a person's voice or to identify with absolute certainty an individual from his or her voice. (p. 35)

Because of the rather dismal state of forensic science in general, Congress in 2005 authorized "the National Academy of Sciences to conduct a study on forensic science, as described in the Senate report" (H. R. Rep. No. 109-272). The Senate Report (No. 109-88, 2005) states in part: "While a great deal of analysis exists of the requirements in the discipline of DNA, there exists little to no analysis of the remaining needs of the community outside of the area of DNA. Therefore ... the Committee directs the Attorney General to provide [funds] to the National Academy of Sciences to create an independent Forensic Science Committee. This Committee shall include members of the forensics community representing operational crime laboratories, medical examiners, and coroners; legal experts; and other scientists as determined appropriate."[9]

[9] Implications of the NRC study have also appeared in the popular media. For example, an article from the *New York Times* by Clyde Haberman (May 18, 2014), entitled "DNA Analysis Exposes Flaws in an Inexact Forensic Science," emphasizes the fallibility of a heretofore staple of forensic science – microscopic hair analysis. We provide several paragraphs from the article:

> This week's offering from Retro Report, a series of video documentaries that re-examine major stories from the past, zeros in on microscopic hair analysis, a staple of forensics for generations. It was long accepted as a virtually unerring technique to prove that this suspect – without a doubt, Your Honor – was the criminal. Wasn't a hair found at the scene?
>
> But with the advent of DNA analysis in the late 1980s, apparent matches of hair samples ultimately proved to be not quite as flawless as people had been led to believe. Instances of wrongful imprisonment make that clear. Retro Report focuses on one such case, that of Kirk Odom, a Washington man who was found guilty of rape in 1981 and spent two decades behind bars. The Federal Bureau of Investigation's vaunted crime lab had asserted that hairs

The results of this National Research Council (NRC) study appeared in book form in 2009 from the National Academies Press (the

taken from his head were microscopically like – meaning virtually indistinguishable from – one found on the victim's nightgown. In time, however, DNA testing established that Mr. Odom was not the rapist, as he had asserted all along. Unfortunately for him, that official conclusion came late. By then, he had completed his prison sentence, a man done in by discredited forensic testimony.

Other lab techniques have had their reliability in the courtroom called into question. A 2009 report by a committee of the National Academy of Sciences found "serious problems" with an assortment of methods routinely relied on by prosecutors and the police. They included fingerprinting, blood typing, weapons identification, shoe print comparisons, handwriting, bite marks and – yes – hair testing. DNA was the game changer. The 2009 report said that, with the exception of nuclear DNA analysis, "no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source."

This is not to say that these techniques are no good at all. Indeed, the F.B.I. still affirms its faith in microscopic hair analysis, particularly as a first look. But it now tries to follow that procedure with a deeper and more certain investigation that uses DNA sampling, and it has done so for 18 years. Nonetheless, many forensic methods no longer come wrapped in the shield of invincibility they once widely enjoyed (especially among those prone to take TV shows literally). Fingerprints get blurred, bullets get smashed, blood specimens get tainted, hairs get mischaracterized.

...

The Innocence Project, a nonprofit group based in New York that uses DNA testing to help clear people wrongly convicted of crimes, has played a notable role in casting doubt on how forensic science is applied. Nationwide over the past 25 years, the project says, 316 people sent to prison have been exonerated through DNA analysis; 18 of them served time on death row. Hair comparisons performed by crime labs were factors in nearly one-fourth of those cases.

Even the F.B.I., while asserting the validity of hair analysis, has effectively acknowledged past problems.

In 2012, in an understanding reached with the Innocence Project and the National Association of Criminal Defense Lawyers, the F.B.I. agreed to a more cautious approach to stay squarely within the confines of known science. No absolutes. The bureau would now say, for example, only that a specific person could be included in, or could be excluded from, a "pool of people of unknown size" who might be the source of a specific hair sample. There would also be no statements of statistical probability. In addition, the F.B.I. says it is examining more than 2,500 old cases that lacked DNA evidence, to determine if hair analysis, of itself, played a role in guilty verdicts. It is unclear how far along this review is.

quotations just given are from this source): *Strengthening Forensic Science in the United States: A Path Forward.* The Summary of this NRC report provides most of what we need to know about the state of forensic science in the United States, and what can or should be done. The material that follows is an excerpt from the NRC Summary chapter:

Problems Relating to the Interpretation of Forensic Evidence:

Often in criminal prosecutions and civil litigation, forensic evidence is offered to support conclusions about "individualization" (sometimes referred to as "matching" a specimen to a particular individual or other source) or about classification of the source of the specimen into one of several categories. With the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source. *In terms of scientific basis, the analytically based disciplines generally hold a notable edge over disciplines based on expert interpretation.* [italics added for emphasis] But there are important variations among the disciplines relying on expert interpretation. For example, there are more established protocols and available research for fingerprint analysis than for the analysis of bite marks. There also are significant variations within each discipline. For example, not all fingerprint evidence is equally good, because the true value of the evidence is determined by the quality of the latent fingerprint image. These disparities between and within the forensic science disciplines highlight a major problem in the forensic science community: The simple reality is that the interpretation of forensic evidence is not always based on scientific studies to determine its validity. This is a serious problem. Although research has been done in some disciplines, there is a notable dearth of peer-reviewed, published studies establishing the scientific bases and validity of many forensic methods.

The Need for Research to Establish Limits and Measures of Performance:

In evaluating the accuracy of a forensic analysis, it is crucial to clarify the type of question the analysis is called on to address. Thus, although some techniques may be too imprecise to permit accurate identification of a specific individual, they may still provide useful and accurate information about questions of classification. For example, microscopic hair analysis may provide reliable evidence on some characteristics of the individual from which the specimen was taken, but it may not be able to reliably match the specimen with a specific individual. However, the definition of the appropriate question is only a first step in the evaluation of the performance of a forensic technique. A body of research is required to establish the limits and measures of performance and to address the impact of sources of variability and potential bias. Such research is sorely needed, but it seems to be lacking in most of the forensic disciplines that rely on subjective assessments of matching characteristics. These disciplines need to develop rigorous protocols to guide these subjective interpretations and pursue equally rigorous research and evaluation programs. The development of such research programs can benefit significantly from other areas, notably from the large body of research on the evaluation of observer performance in diagnostic medicine and from the findings of cognitive psychology on the potential for bias and error in human observers.

The Admission of Forensic Science Evidence in Litigation:

Forensic science experts and evidence are used routinely in the service of the criminal justice system. DNA testing may be used to determine whether sperm found on a rape victim came from an accused party; a latent fingerprint found on a gun may be used to determine whether a defendant handled the weapon; drug analysis may be used to determine whether pills found in a person's possession were illicit; and an autopsy may be used to determine the cause and manner of death of a murder victim. . . . for qualified forensic science experts to testify competently about forensic evidence, they must first find the evidence in a usable state and properly preserve it. A latent fingerprint that is badly smudged when found cannot be usefully saved, analyzed, or explained. An inadequate drug sample may be insufficient to allow for proper analysis. And, DNA tests performed on a contaminated or otherwise compromised sample cannot be used reliably to identify or eliminate an individual as the perpetrator of a crime. These are important matters involving the proper processing of forensic evidence. The law's greatest dilemma in its heavy reliance on forensic evidence, however, concerns the question of whether—and to what extent—there is *science* in any given forensic science discipline.

Two very important questions should underlie the law's admission of and reliance upon forensic evidence in criminal trials: (1) the extent to which a particular forensic discipline is founded on a reliable scientific methodology that gives it the capacity to accurately analyze evidence and report findings and (2) the extent to which practitioners in a particular forensic discipline rely on human interpretation that could be tainted by error, the threat of bias, or the absence of sound operational procedures and robust performance standards. These questions are significant. Thus, it matters a great deal whether an expert is qualified to testify about forensic evidence and whether the evidence is sufficiently reliable to merit a fact finder's reliance on the truth that it purports to support. Unfortunately, these important questions do not always produce satisfactory answers in judicial decisions pertaining to the admissibility of forensic science evidence proffered in criminal trials.

A central idea present throughout the collection of modules is that "context counts" and it "counts crucially." It is important both for experts and novices in how a question is asked, how a decision task is framed, and how forensic identification is made. People are primed by context whether as a victim making an eyewitness identification of a perpetrator, or as an expert making a fingerprint match. As an example of the latter, we have the 2006 article by Dror, Charlton, and Péron, "Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications" (*Forensic Science International*, *156*, 74–78). We give their abstract below:

We investigated whether experts can objectively focus on feature information in fingerprints without being misled by extraneous information, such as context. We took fingerprints that have previously been examined and assessed by latent print experts to make positive identification of suspects. Then we presented these same fingerprints again, to the same experts, but gave a context that suggested that they were a no-match, and hence the suspects could not be identified. Within this new context, most of the fingerprint experts made different judgments, thus contradicting their own previous identification decisions. Cognitive aspects involved in biometric identification can explain why experts are vulnerable to make erroneous identifications. (p. 74)

References

[1] Meehl, P., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194–215.

[2] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Module 9: Probability and Litigation

It is now generally recognized, even by the judiciary, that since all evidence is probabilistic—there are no metaphysical certainties—evidence should not be excluded merely because its accuracy can be expressed in explicitly probabilistic terms.

— Judge Richard A. Posner ("An Economic Approach to the Law of Evidence," *Stanford Law Review*, 1999)

Abstract: This module explores the connection between statements that involve probabilities and those phrases used for evidentiary purposes in the courts. We begin with Jack Weinstein, a federal judge in the Eastern District of New York, and his views on the place that probability has in litigation. Jack Weinstein may be the only federal judge ever to publish an article in a major statistics journal; his primary interests center around subjective probability and how these relate, among others, to the four levels of a "legal burden of proof": preponderance of the evidence; clear and convincing evidence; clear, unequivocal, and convincing evidence; and proof beyond a reasonable doubt. The broad topic area of probability scales and rulers is discussed in relation to several more specific subtopics: Jeremy Bentham and his suggestion of a "persuasion thermometer"; some of Jack Weinstein's legal rulings where probabilistic assessments were made: the cases of Vincent Gigante, Agent Orange, and Daniel Fatico. An appendix gives a redacted Weinstein opinion in this later Fatico case. Two other appendices are also given: the text of Maimonides' 290th Negative Commandment, and a District of Columbia Court of Appeals opinion "In re As.H" (2004) that dealt with the assignment of subjective probabilities and various attendant verbal phrases in eyewitness testimony.

Contents

1 Federal Judge Jack Weinstein, Eastern District of New York (Brooklyn)

The retirement of Supreme Court Justice John Paul Stevens in 2010 gave President Obama a second opportunity to nominate a successor who drew the ire of the Republican Party during the confirmation process (similar to the previous such hearing with "the wise Latina woman"). For one who might have enjoyed witnessing a collective apoplexy from the conservative right, we could have suggested that President Obama nominate Jack Weinstein, a sitting federal judge in the Eastern District of New York (Brooklyn), if it weren't for the fact that at 89 he was only one year younger than the retiring Stevens.

Weinstein is one of the most respected and influential judges in America. He has directly organized and presided over some of the most important mass tort cases of the last forty years (for example, Agent Orange, asbestos, tobacco, breast implants, DES, Zyprexa, handgun regulation, and repetitive-stress injuries).[1] For present purposes, our interest is in Weinstein's deep respect for science-based evidence in the judicial process, and in particular, for how he views probability and statistics as an intimate part of that process. He also may be the only federal judge ever to publish an article in a major statistics journal (*Statistical Science*, 1988, *3*, 286–297, "Litigation and Statistics"). This last work developed out of Weinstein's association in the middle 1980s with the National Academy of Science's Panel on Statistical Assessment as Evidence in the Courts. This panel produced the comprehensive Springer-Verlag volume, *The Evolving Role of Statistical Assessments as Evidence in the Courts* (1988; Stephen E. Fienberg, Editor).

The importance that Weinstein gives to the role of probability and statistics in the judicial process is best expressed by Weinstein himself (we quote from his *Statistical Science* article):

> The use of probability and statistics in the legal process is not unique to our times. Two thousand years ago, Jewish law, as stated in the Talmud, cautioned about the use of probabilistic inference. The medieval Jewish commentator Maimonides summarized this traditional view in favor of certainty when he noted:
>
> "The 290th Commandment is a prohibition to carry out punishment on a high probability, even close to certainty … No punishment [should] be

[1] A tort is a civil wrong; tort law concerns situations where a person's behavior has harmed someone else.

> carried out except where ... the matter is established in certainty beyond any doubt ... ”
>
> That view, requiring certainty, is not acceptable to the courts. We deal not with the truth, but with probabilities, in criminal as well as civil cases. Probabilities, express and implied, support every factual decision and inference we make in court. (p. 287)

Maimonides’ description of the 290th Negative Commandment is given in its entirety in an appendix to this module. According to this commandment, an absolute certainty of guilt is guaranteed by having two witnesses to exactly the same crime. Such a probability of guilt being identically one is what is meant by the contemporary phrase “without any shadow of a doubt.”

Two points need to be emphasized about this Mitzvah (Jewish commandment). One is the explicit unequalness of costs attached to the false positive and negative errors: “it is preferable that a thousand guilty people be set free than to execute one innocent person.” The second is in dealing with what would now be characterized as the (un)reliability of eyewitness testimony. Two eyewitnesses are required, neither is allowed to make just an inference about what happened but must have observed it directly, and exactly the same crime must be observed by both eyewitnesses. Such a high standard of eyewitness integrity might have made the current rash of DNA exonerations unnecessary.

Judge Weinstein’s interest in how probabilities could be part of a judicial process goes back some years before the National Research Council Panel. In one relevant opinion from 1978, *United States v. Fatico*, he wrestled with how subjective probabilities might be related to the four levels of a “legal burden of proof”; what level was required in this particular case; and, finally, was it then met. The four (ordered) levels are: preponderance of the evidence; clear and convincing evidence; clear, unequivocal, and convincing evidence; and proof beyond a reasonable doubt. The case in point involved proving that Daniel Fatico was a "made" member of the Gambino organized crime family, and thus could be given a "Special Offender" status. "The consequences of (such) a 'Special Offender' classification are significant. In most cases, the designation delays or precludes social furloughs, release to half-way houses and transfers to other correctional institutions; in some cases, the characterization may bar early parole" (text taken from the opinion). The summary of Weinstein's final opinion in the Fatico case follows:

> In view of prior proceedings, the key question of law now presented is what burden of proof must the government meet in establishing a critical fact not proved at a criminal trial that may substantially enhance the sentence to be imposed upon a defendant. There are no precedents directly on point.
>
> The critical factual issue is whether the defendant was a "made" member of an organized crime family. Clear, unequivocal and convincing evidence adduced by the government at the sentencing hearing establishes this proposition of fact.

The text of Weinstein's opinion in the Fatico case explains some of the connections between subjective probabilities, burdens of proof, and the need for different levels depending on the particular case (we note that the numerical values suggested in this opinion as corresponding to the various levels of proof, appear to be based only on Judge Weinstein's "best guesses"). We redact part of his opinion in an appendix to this module.

Other common standards used for police searches or arrests might also be related to an explicit probability scale. The lowest standard (perhaps a probability of 20%) would be "reasonable suspicion" to determine whether a brief investigative stop or search by any governmental agent is warranted (in the 2010 "Papers, Please" law in Arizona, a "reasonable suspicion" standard is set for requesting documentation). A higher standard would be "probable cause" to assess whether a search or arrest is warranted, or whether a grand jury should issue an indictment. A value of, say, 40% might indicate a "probable cause" level that would put it somewhat below a "preponderance of the evidence" criterion. In all cases, a mapping of such verbal statements to numerical values requires "wiggle" room for vagueness, possibly in the form of an interval estimate rather than a point estimate. As an example of this variability of assessment in the Fatico case, Judge Weinstein informally surveyed the judges in his district court regarding the four different standards of proof. The data are given in Table 1 (taken from Fienberg, 1988, p. 204). We leave it to you to decide whether you would want Judge 4 or 7 to hear your case.

2 Probability Scales and Rulers

The topic of relating a legal understanding of burdens of proof to numerical probability values has been around for a very long time. Fienberg (1988, p. 212) provides a short discussion of Jeremy Bentham's (1827) suggestion of a "persuasion thermometer," and some contemporary reaction to this idea from Thomas Starkie (1833).[2] We

[2] This is the same Bentham known for utilitarianism, and more amusingly, for the "auto-icon." A short section from the Wikipedia article on "Jeremy Bentham" describes the

Table 1: Probabilities associated with different standards of proof by judges in the Eastern District of New York.

Judge	Prepon-	Clear and	Clear,	Beyond a	
	derance	convincing	unequivocal,	reasonable	row
			and convincing	doubt	median
1	50+	60	70	85	65
2	51	65	67	90	66
3	50+	60-70	65-75	80	67
4	50+	67	70	76	69
5	50+	Standard is elusive		90	70
6	50+	70+	70+	85	70
7	50+	60	90	85	72
8	50+	70+	80+	95+	75
9	50.1	75	75	85	75
10	51		Cannot estimate		–
column median	50	66	70	85	

quote:

Jeremy Bentham appears to have been the first jurist to seriously propose

auto-icon:

As requested in his will, Bentham's body was dissected as part of a public anatomy lecture. Afterward, the skeleton and head were preserved and stored in a wooden cabinet called the "Auto-icon," with the skeleton stuffed out with hay and dressed in Bentham's clothes. Originally kept by his disciple Thomas Southwood Smith, it was acquired by University College London in 1850. It is normally kept on public display at the end of the South Cloisters in the main building of the college, but for the 100th and 150th anniversaries of the college, it was brought to the meeting of the College Council, where it was listed as "present but not voting."

The Auto-icon has a wax head, as Bentham's head was badly damaged in the preservation process. The real head was displayed in the same case for many years, but became the target of repeated student pranks, including being stolen on more than one occasion. It is now locked away securely.

that witnesses and judges numerically estimate their degrees of persuasion. Bentham (1827; Vol. 1, pp. 71–109) envisioned a kind of moral thermometer:

> The scale being understood to be composed of ten degrees—in the language applied by the French philosophers to thermometers, a decigrade scale—a man says, My persuasion is at 10 or 9, etc. affirmative, or at least 10, etc. negative . . .

Bentham's proposal was greeted with something just short of ridicule, in part on the pragmatic grounds of its inherent ambiguity and potential misuse, and in part on the more fundamental ground that legal probabilities are incapable of numerical expression. Thomas Starkie (1833) was merely the most forceful when he wrote:

> The notions of those who have supposed that mere moral probabilities or relations could ever be represented by numbers or space, and thus be subjected to arithmetical analysis, cannot but be regarded as visionary and chimerical. (p. 212)

Several particularly knotty problems and (mis)interpretations when it comes to assigning numbers to the possibility of guilt arise most markedly in eyewitness identification. Because cases involving eyewitness testimony are typically criminal cases, they demand burdens of proof "beyond a reasonable doubt"; thus, the (un)reliability of eyewitness identification becomes problematic when it is the primary (or only) evidence presented to meet this standard. As discussed extensively in the judgment and decision-making literature, there is a distinction between making a subjective estimate of some quantity, and one's confidence in that estimate once made. For example, suppose someone picks a suspect out of a lineup, and is then asked the (Bentham) question, "on a scale of from one to ten, characterize your level of 'certainty'." Does an answer of "seven or eight" translate into a probability of innocence of two or three out of ten? Exactly such confusing situations, however, arise. We give a fairly extensive redaction in an appendix to this module of an opinion from the District of Columbia Court of Appeals in a case named "In re As.H" (2004). It combines extremely well both the issues of eyewitness (un)reliability and the attempt to quantify that which may be better left in words; the dissenting Associate Judge Farrel noted pointedly: "I believe that the entire effort to quantify the standard of proof beyond a reasonable doubt is a search for fool's gold."

2.1 The Cases of Vincent Gigante and Agent Orange

Although Judge Weinstein's reputation may rest on his involvement with mass toxic torts, his most entertaining case occurred in the middle 1990s, with the murder-conspiracy and racketeering trial and conviction of Vincent Gigante, the boss of the most powerful Mafia family in the United States. The issue here was assessing the evidence of Gigante's mental fitness to be sentenced to prison, and separating such evidence from the putative malingering of Gigante. Again, Judge Weinstein needed to evaluate the evidence and make a probabilistic assessment ("beyond a reasonable doubt") that Gigante's trial was a "valid" one.

Apart from the great legal theater that the Gigante case provided, Judge Weinstein's most famous trials all involve the Agent Orange defoliant used extensively by the United States in Vietnam in the 1960s. Originally, he oversaw in the middle 1980s the $200 million settlement fund provided by those companies manufacturing the agent. Most recently, Judge Weinstein presided over the dismissal of the civil lawsuit filed on behalf of millions of Vietnamese individuals. The 233-page decision in this case is an incredible "read" about

United States polices during this unfortunate period in our country's history. The suit was dismissed not because of poor documentation of the effects of Agent Orange and various ensuing conditions, but because of other legal conditions. The judge concluded that even if the United States had been a Geneva Accord signatory (outlawing use of poisonous gases during war), Agent Orange would not have been banned: "The prohibition extended only to gases deployed for their asphyxiating or toxic effects on man, not to herbicides designed to affect plants that may have unintended harmful side effects on people" (In re "Agent Orange" Product Liability Litigation, 2005, p. 190). The National Academy of Science through its Institute of Medicine, regularly updates what we know about the effects of Agent Orange, and continues to document many associations between it and various disease conditions. The issues in the Vietnamese lawsuit, however, did not hinge on using a probability of causation assessment, but rather on whether, given the circumstances of the war, the United States could be held responsible for what it did in Vietnam in the 1960s.

3 Appendix: Maimonides' 290th Negative Commandment

"And an innocent and righteous person you shall not slay" — Exodus 23:7.

Negative Commandment 290

Issuing a Punitive Sentence Based on Circumstantial Evidence:

The 290th prohibition is that we are forbidden from punishing someone based on our estimation [without actual testimony], even if his guilt is virtually certain. An example of this is a person who was chasing after his enemy to kill him. The pursued escaped into a house and the pursuer entered the house after him. We enter the house after them and find the victim lying murdered, with the pursuer standing over him holding a knife, with both covered with blood. The Sanhedrin may not inflict the death penalty on this pursuer since there were no witnesses who actually saw the murder.

The Torah of Truth (Toras Emess) comes to prohibit his execution with G—d's statement (exalted be He), "Do not kill a person who has not been proven guilty."

Our Sages said in *Mechilta*: "If they saw him chasing after another to kill him and they warned him, saying, 'He is a Jew, a son of the Covenant! If you kill him you will be executed!' If the two went out of sight and they found one murdered, with the sword in the murderer's hand dripping blood, one might think that he can be executed. The Torah therefore says, 'Do not kill a person who has not been proven guilty.' "

Do not question this law and think that it is unjust, for there are some possibilities that are extremely probable, others that are extremely unlikely, and others in between. The category of "possible" is very broad, and if the Torah allowed the High Court to punish when the offense was very probable and almost definite (similar to the above example), then they would carry out punishment in cases which were less and less probable, until people would be constantly executed based on flimsy estimation and the judges' imagination. G—d (exalted be He), therefore "closed the door" to this possibility and forbid any punishment unless there are witnesses who are certain beyond a doubt that the event transpired and that there is no other possible explanation.

If we do not inflict punishment even when the offense is most probable, the worst that could happen is that someone who is really guilty will be found innocent. But if punishment was given based on estimation and circumstantial evidence, it is possible that someday an innocent person would be executed. And it is preferable and more proper that even a thousand guilty people be set free than to someday execute even one innocent person.

Similarly, if two witnesses testified that the person committed two capital offenses, but each one saw only one act and not the other, he cannot be executed. For example: One witness testified that he saw a person doing a *melachah* on *Shabbos* and warned him not to. Another witness testified that he saw the person worshipping idols and warned him not to. This person cannot be executed by stoning. Our Sages said, “If one witness testified that he worshipped the sun and the other testified that he worshipped the moon, one might think that they can be joined together. The Torah therefore said, ‘Do not kill a person who has not been proven guilty.’ ”

4 Appendix: The Redacted Text of Judge Weinstein’s Opinion in United States v. Fatico (1978)

We begin with the caution of Justice Brennan in Speiser v. Randall, about the crucial nature of fact finding procedures:

To experienced lawyers it is commonplace that the outcome of a lawsuit and hence the vindication of legal rights depends more often on how the factfinder appraises the facts than on a disputed construction of a statute or interpretation of a line of precedents. Thus the procedures by which the facts of the case are determined assume an importance fully as great as the validity of the substantive rule of law to be applied. And the more important the rights at stake, the more important must be the procedural safeguards surrounding those rights.

The “question of what degree of proof is required . . . is the kind of question which has traditionally been left to the judiciary to resolve . . . ”

Broadly stated, the standard of proof reflects the risk of winning or losing a given adversary proceeding or, stated differently, the certainty with which the party bearing the burden of proof must convince the factfinder.

As Justice Harlan explained in his concurrence in Winship, the choice of an appropriate burden of proof depends in large measure on society’s assessment of the stakes involved in a judicial proceeding.

In a judicial proceeding in which there is a dispute about the facts of some earlier event, the factfinder cannot acquire unassailably accurate knowledge of what happened. Instead, all the factfinder can acquire is a belief of what Probably happened. The intensity of this belief—the degree to which a factfinder is convinced that a given act actually occurred—can, of course, vary. In this regard, a standard of proof represents an attempt to instruct the factfinder concerning the degree of confidence our society thinks he should have in the correctness of factual conclusions for a particular type of adjudication. Although the phrases "preponderance of the evidence" and "proof beyond a reasonable doubt" are quantitatively imprecise, they do communicate to the finder of fact different notions concerning the degree of confidence he is expected to have in the correctness of his factual conclusions.

Thus, the burden of proof in any particular class of cases lies along a continuum from low probability to very high probability.

Preponderance of the Evidence:

As a general rule, a "preponderance of the evidence" [or] more probable than not standard, is relied upon in civil suits where the law is indifferent as between plaintiffs and defendants, but seeks to minimize the probability of error.

In a civil suit between two private parties for money damages, for example, we view it as no more serious in general for there to be an erroneous verdict in the defendant's favor than for there to be an erroneous verdict in the plaintiff's favor. A preponderance of the evidence standard therefore seems peculiarly appropriate; as explained most sensibly, it simply requires the trier of fact "to believe that the existence of a fact is more probable than its nonexistence before (he) may find in favor of the party who has the burden to persuade the (judge) of the fact's existence."

Quantified, the preponderance standard would be 50+% Probable.

Clear and Convincing Evidence:

In some civil proceedings where moral turpitude is implied, the courts utilize the standard of "clear and convincing evidence," a test somewhat stricter than preponderance of the evidence.

Where proof of another crime is being used as relevant evidence pursuant to Rules 401 to 404 of the Federal Rules of Evidence, the most common test articulated is some form of the "clear and convincing" standard.

Quantified, the probabilities might be in the order of above 70% under a clear and convincing evidence burden.

Clear, Unequivocal and Convincing Evidence:

"In situations where the various interests of society are pitted against restrictions on the liberty of the individual, a more demanding standard is frequently imposed, such as proof by clear, unequivocal and convincing evidence." The Supreme Court has applied this stricter standard to deportation proceedings, denaturalization cases, and expatriation cases. In Woodby, the Court explained:

To be sure, a deportation proceeding is not a criminal prosecution. But it does not syllogistically follow that a person may be banished from this country upon no higher degree of proof than applies in a negligence case. This Court has not closed its eyes to the drastic deprivations that may follow when a resident of this country is compelled by our Government to forsake all the bonds formed here and go to a foreign land where he often has no contemporary identification.

In terms of percentages, the probabilities for clear, unequivocal and convincing evidence might be in the order of above 80% under this standard.

Proof Beyond a Reasonable Doubt:

The standard of "proof beyond a reasonable doubt" is constitutionally mandated for elements of a criminal offense. Writing for the majority in Winship, Justice Brennan enumerated the "cogent reasons" why the "'reasonable-doubt' standard plays a vital role in the American scheme of criminal procedure" and "is a prime instrument for reducing the risk of convictions resting on factual error."

The accused during a criminal prosecution has at stake interest of immense importance, both because of the possibility that he may lose his liberty upon conviction and because of the certainty that he would be stigmatized by the conviction. Accordingly, a society that values the good name and freedom of every individual should not condemn a man for commission of a crime when there is reasonable doubt about his guilt. As we said in Speiser v. Randall, "There is always in litigation a margin of error, representing error in fact finding, which both parties must take into account. Where one party has at stake an interest of transcending value as a criminal defendant—his liberty—this margin of error is reduced as to him by the process of placing on the other party the burden of ... persuading the factfinder at the conclusion of the trial of his guilt beyond a reasonable doubt. Due process commands that no man shall lose his liberty unless the Government has borne the burden of ... convincing the factfinder of his guilt." ...

Moreover, use of the reasonable-doubt standard is indispensable to command the respect and confidence of the community in applications of the criminal law. It is critical that the moral force of the criminal law not be diluted by a standard of proof that leaves people in doubt whether innocent men are being condemned.

In capital cases, the beyond a reasonable doubt standard has been utilized for findings of fact necessary to impose the death penalty after a finding of guilt.

Many state courts, in interpreting state recidivism statutes, have held that proof of past crimes must be established beyond a reasonable doubt.

In civil commitment cases, where the stakes most resemble those at risk in a criminal trial, some courts have held that the beyond a reasonable doubt standard is required.

If quantified, the beyond a reasonable doubt standard might be in the range of 95+% Probable.

5 Appendix: District of Columbia Court of Appeals, "In re As.H" (2004)

DISTRICT OF COLUMBIA COURT OF APPEALS

IN RE AS.H.

SCHWELB, Associate Judge: This juvenile delinquency case is more than five years old. On January 20, 1999, following a factfinding hearing, As.H., then sixteen years of age, was adjudicated guilty of robbery. The sole evidence implicating As.H. in the offense was the testimony of the victim, Ms. Michal Freedhoff, who identified As.H. at a photo array almost a month after the robbery and again in court more than four months after that. Ms. Freedhoff described her level of certainty on both occasions, however, as "seven or eight" on a scale of one to ten. Because Ms. Freedhoff was obviously less than positive regarding her identification, and for other reasons described below, we conclude as a matter of law that the evidence was insufficient to prove beyond a reasonable doubt that As.H. was involved in the robbery. Accordingly, we reverse.

I. In the early morning hours of August 17, 1998, between 12:30 and 1:00 a.m., Ms. Freedhoff was robbed by three or more young men. The assailants knocked Ms. Freedhoff to the ground, threatened her with “a long piece of wood” which, Ms. Freedhoff believed, was “suppose[d] to look like a rifle,” ordered her to “shut up, bitch,” and robbed her of her purse and her personal electronic organizer. Ms. Freedhoff promptly reported the crime to the police. Officers detained a group of young men shortly after the robbery and arranged a show-up, but Ms. Freedhoff stated that the detained individuals were not the robbers. Indeed, she was “completely” certain that the individuals at the show-up were not the guilty parties.

Ms. Freedhoff testified that there were street lights in the area where the robbery occurred. She further stated that she had been outside in the street for some time, so that her eyes had become accustomed to the dark. Nevertheless, Ms. Freedhoff could not provide an informative description of her assailants. According to Detective Ross, she recalled nothing distinctive about their clothing; “young black males and baggy clothes” was his recollection of her report. At the factfinding hearing, which took place more than five months after the robbery, Ms. Freedhoff recalled that the robbers were teenagers, “two dark-skinned and one light,” each of a different height, and that “one had shorts and sneakers and another may have had a hat.” Ms. Freedhoff was also uncertain as to the role which the individual she tentatively identified as As.H. allegedly played in the robbery.

On September 11, 1998, Detective Ross showed Ms. Freedhoff an array of nine polaroid pictures and asked her if she recognized anyone who was involved in the offense. At a hearing on As.H.’s motion to suppress identification, Ms. Freedhoff testified as follows regarding this array:

Q: Now, Ms. Freedhoff, on that day did you identify any of the people in the photos as having been involved in the incident of August 16th?

A: Yes, I did.

Q: Which photos did you identify?

A: These two marked nine and [ten] I was very certain about and the two marked three and four I was less certain about.

Q: During the identification procedure, did you talk to the detective about your level of certainty?

A: Yes.

Q: In terms of nine and [ten], what was your level of certainty that those people were involved?

A: I [was] asked to rate them on a scale of—I believe it was one to [ten]—and I believe I said it was, that nine and [ten], I was seven or eight.

Q: And in terms of three and four, how did you rate those?

A: Six.

According to Ms. Freedhoff, the photograph of As.H. was No. 10. At the factfinding hearing, Ms. Freedhoff initially stated that she saw one of the robbers sitting in the courtroom, pointing out As.H. When asked which of the individuals in the array he was, Ms. Freedhoff "believed" that it "would be Number 10." However, when counsel for the District of Columbia again asked Ms. Freedhoff about her present level of certainty in making the identification—how certain are you?—the witness adhered to her previous estimate: "At the time, on a scale of one to [ten], I said that I was seven or eight."

According to Detective Ross, who also testified regarding the viewing of the photo array, Ms. Freedhoff was "comfortable in saying they could be the people that robbed her." Ross further disclosed that he "may have discussed with [Ms. Freedhoff] that I had a previous history with the persons that she had picked. They were my possible suspects in the case."

Without elaborating on his reasons, the trial judge denied As.H.'s motion to suppress identification and found As.H. guilty as charged. This appeal followed.

II. In evaluating claims of evidentiary insufficiency in juvenile delinquency appeals, we view the record "in the light most favorable to the [District], giving full play to the right of the judge, as the trier of fact, to determine credibility, weigh the evidence, and draw reasonable inferences ... We will reverse on insufficiency grounds only when the [District] has failed to produce evidence upon which a reasonable mind might fairly find guilt beyond a reasonable doubt." "Even identification testimony of a single eyewitness will be sufficient so long as a reasonable person could find the identification convincing beyond a reasonable doubt." Moreover, the District was not required to prove As.H.'s guilt beyond all doubt. "There is no rule of law which requires an identification to be positive beyond any shadow of doubt."

Nevertheless, the "[beyond a] reasonable doubt" standard of proof is a formidable one. It "requires the factfinder to reach a subjective state of near certitude of the guilt of the accused." Although appellate review is deferential, we have "the obligation to take seriously the requirement that the evidence in a criminal prosecution must be strong enough that a jury behaving rationally really could find it persuasive beyond a reasonable doubt." Moreover, "while [the trier of fact] is entitled to draw a vast range of reasonable inferences from evidence, [he or she] may not base [an adjudication of guilt] on mere speculation."[3]

In the present case, we have an eyewitness identification of questionable certitude, and the witness and the respondent are strangers. Ms. Freedhoff saw her assailants at night and under extremely stressful conditions. Moreover, this is a "pure" eyewitness identification case; there is no evidence linking As.H. to the robbery except for Ms. Freedhoff's statements upon viewing the array almost a month after the event and her testimony at the factfinding hearing more than five months after she was robbed.

The vagaries of eyewitness identification, and the potential for wrongful convictions or adjudications based upon such evidence, have long been recognized in the District of Columbia. More recently, in Webster v. United States, we summarized this concern as follows:

"[T]he identification of strangers is proverbially untrustworthy. The hazards of such testimony are established by a formidable number of instances in the records of English and American trials." FELIX FRANKFURTER, THE CASE OF SACCO AND VANZETTI (1927). Indeed, "[p]ositive identification of a person not previously known to the witness is perhaps the most fearful testimony known to the law of evidence." Even if the witness professes certainty, "it is well recognized that the most positive eyewitness is not necessarily the most reliable."

Here, the witness did not even profess certainty. Moreover, the present case concerns a hesitant cross-racial identification by a white woman of a

[3] The court emphasized in Crawley that as appellate judges, we have the responsibility in eyewitness identification cases "to draw upon our own experience, value judgments, and common sense in determining whether the [finding] reached was in keeping with the facts." Although this observation might be viewed today as an unduly activist formulation of an appellate court's function, it illustrates the concern of conscientious judges regarding the possibility that a mistaken identification may send an innocent person to prison.

black teenager, and "[i]t is well established that there exists a comparative difficulty in recognizing individual members of a race different from one's own." ELIZABETH LOFTUS, EYEWITNESS TESTIMONY; see State v. Cromedy, (discussing at length the difficulties in cross-racial identification and mandating a jury instruction on the subject in some cases); John P. Rutledge, They All Look Alike: The Inaccuracy of Cross-Racial Identifications, 28 AM. J. CRIM. L. 207 (2001).

It is in the context of these realities that we now turn to the dispositive issue in this appeal, namely, whether Ms. Freedhoff's testimony—the only evidence of As.H.'s participation in the robbery—was legally sufficient to support a finding of guilt beyond a reasonable doubt. The key fact is that, both when viewing the polaroid photographs and when testifying in open court, Ms. Freedhoff candidly characterized her level of "certainty"—i.e., of her being "very certain"—as seven or eight on a scale of one to ten. Her testimony leads inexorably to the conclusion that her level of uncertainty—i.e., the possibility that As.H. was not involved—was two or three out of ten—a 20% to 30% possibility of innocence. This differs dramatically from Ms. Freedhoff's complete certainty that the young men she viewed at the show-up on the night of the offense were not the robbers. The contrast between Ms. Freedhoff's statements in the two situations is revealing, and surely negates the "near certitude" that is required for a showing of guilt beyond a reasonable doubt. The "seven or eight out of ten" assessment is also consistent with Detective Ross' recollection of Ms. Freedhoff's account: As.H. and others "could be the people that robbed her," and As.H. "looked like" one of the kids. It is, of course, difficult (if not impossible) to place a meaningful numerical value on reasonable doubt. See generally Tribe, Trial by Mathematics: Precision and Ritual in the Legal Process, 84 HARV. L. REV 1329 (1971); Underwood, The Thumb on the Scales of Justice; Burden of Persuasion in Criminal Cases, 86 Yale L.J. 1299, 1309–11 (1977) (hereinafter Underwood). Professor Wigmore cites a study in which judges in Chicago were asked to:

translate into probability statements their sense of what it means to be convinced by a preponderance of the evidence, and to be convinced beyond a reasonable doubt. When responding to questionnaires, at least, the judges thought there was an important difference: almost a third of the responding judges put "beyond a reasonable doubt" at 100%, another third put it at 90% or 95%, and most of the rest put it at 80% or 85%. For the preponderance standard, by contrast, over half put it at 55%, and most of the rest put it between 60% and 75%.

Although the Chicago study alone is not dispositive of this appeal, it reveals that very few judges, if any, would have regarded an 80% probability as sufficient to prove guilt beyond a reasonable doubt, and that all of them would have considered a 70% probability as altogether inadequate. For the Chicago judges, Ms. Freedhoff's "certainty" appears to be well outside the ballpark for proof in a criminal case. In Fatico, nine judges of the United States District Court for the Eastern District of New York, responding to a poll by Judge Weinstein, the co-author of a leading text on evidence, suggested percentages of 76%, 80%, 85%, 85%, 85%, 85%, 90%, 90% and 95%, as reflecting the standard for proof beyond a reasonable doubt. Thus, at most, two of the nine judges polled by Judge Weinstein would have found the level of assurance voiced by Ms. Freedhoff sufficient to support a finding of guilt.[4]

But, argues the District, Ms. Freedhoff "was not asked for a level of accuracy or how sure she was, but, given the certainty of her identification, how high a level of certainty she had felt." Therefore, the argument goes, "the trier of fact can be confident that the witness felt that her identification was very certain." We do not find this contention at all persuasive. Taken to its logical conclusion, it would mean that if Ms. Freedhoff had expressed a level of certainty of one in ten—10%—this would be sufficient to support a finding of guilt. The notion that Ms. Freedhoff was assessing varying gradations of certainty, all of them very certain, is also at odds with what she told Detec-

[4] Commenting on the same Chicago study in one of its submissions, the District reveals only that "about one-third of the judges put it at 80%-85%." Unfortunately, by failing to mention that one third of the judges put "beyond a reasonable doubt" at 100% and that another third put it at 90%-95%, the District presents us with a misleading picture of the results of the study. Remarkably, the District then argues that we should affirm because judges who try to quantify reasonable doubt place it "not that far distant from Ms. Freedhoff's estimate." In fact, the contrast between the judges' estimates and Ms. Freedhoff's articulation is quite remarkable, and a contention that fails to take this contrast into account is necessarily fallacious.

tive Ross, namely, that As.H. "looks like" or "could have been" one of the robbers.[5]

Professor Lawrence Tribe has written:

[I]t may well be ... that there is something intrinsically immoral about condemning a man as a criminal while telling oneself, "I believe that there is a chance of one in twenty that this defendant is innocent, but a 1/20 risk of sacrificing him erroneously is one I am willing to run in the interest of the public's—and my own—safety."

It may be that Professor Tribe's proposition is more suited to the world of academe than to the less rarefied realities of the Superior Court's criminal docket—realities in which "beyond all doubt" presents an idealist's impossible dream, while "beyond a reasonable doubt" provides a workable standard. This case, however, is not like the hypothetical one that disturbed Professor Tribe. Here, the doubt of the sole identifying witness in a night-time robbery by strangers to her stood at two or three out of ten, or 20%-30%. We conclude, at least on this record, that this level of uncertainty constituted reasonable doubt as a matter of law. Accordingly, we reverse the adjudication of guilt and remand the case to the Superior Court with directions to enter a judgment of not guilty and to dismiss the petition.[6]

[5] The District also argues that, in open court, Ms. Freedhoff "unhesitatingly and positively" identified the respondent. As we have explained in Part I of this opinion, however, the full context of Ms. Freedhoff's courtroom testimony reveals that, five months after the robbery, she was no more certain of her identification than she had been when she viewed the photo array. Moreover, after Ms. Freedhoff had selected photographs at the array, Detective Ross revealed that he "had a previous history with the persons she had picked," and that they were his "possible suspects in the case." "[W]here ... the police consider an individual to be a possible perpetrator and a witness makes an initially ambiguous identification, there may develop a process of mutual bolstering which converts initial tentativeness into ultimate certainty." "The victim relies on the expertise of the officer and the officer upon the victim's identification."

[6] Our dissenting colleague argues that reasonable doubt is not susceptible of ready quantification, and we agree. But where, as in this case, the sole identifying witness described her own level of "certainty" as only seven or eight on a scale of ten, then, notwithstanding the difficulty of quantification in the abstract, this level of unsureness necessarily raises a reasonable doubt and negates the requisite finding of "near certitude" that As.H. was one of the robbers. Nothing in this opinion holds or even remotely suggests that a cross-racial identification is insufficient as a matter of law or that the trier of fact is required to discount

So ordered.

FARRELL, Associate Judge, dissenting: Less than a month after she was assaulted by three young men, the complainant, Ms. Freedhoff, identified two men from photographs as among the assailants. One was appellant. According to the detective who showed her the photographs, she did not hesitate in picking appellant, and at the hearing on appellant's motion to suppress the identification she twice stated that she had been "very certain" in selecting his photograph. At trial, although she could not remember appellant's exact role in the assault, she stated that she had been able to see all three assailants well, that the two people she was "certain of" in her identification "were probably the two" who had been "in front of [her]" during the assault, and that she had identified them because "they looked very familiar to [her] as being the people that were involved." Ms. Freedhoff was not given to quick accusations: at a show-up confrontation shortly after the assault, she had been "[completely] certain" that the individuals shown to her were not the assailants. The trial judge, sitting as trier of fact, found her testimony convincing and found appellant guilty beyond a reasonable doubt.

The majority sets that verdict aside. Although concededly unable to replicate Judge Mitchell's vantage point in assessing the complainant's demeanor and the strength of her belief as she recalled the robbery and identification, it concludes that the identification was too weak as a matter of law to support conviction. And it does so at bottom for one reason: when asked by the detective her level of certainty "on a scale of one to ten" in identifying appellant, Ms. Freedhoff had answered "seven or eight." This, in the majority's view, explains what she meant when she said she was "very certain," and a level of uncertainty of an uncorroborated eyewitness "st[anding] at two or three out of ten, or 20%-30%[,] . . . constituted reasonable doubt as a matter of law."

The majority thus decides that the trier of fact could not convict based on testimony of a victim who was as much as four-fifths certain of her iden-

such an identification. The reasonable doubt in this case arises from the witness' very limited certainty (seven or eight on a scale of ten) regarding her uncorroborated identification. The difficulties of eyewitness identification of strangers in general, as well as of cross-racial identification, provide the context in which the witness' uncertainty arose.

tification. I do not agree, basically because I believe that the entire effort to quantify the standard of proof beyond a reasonable doubt is a search for fool's gold. Ms. Freedhoff stated that she was very certain of her identification; she was questioned extensively about the circumstances of the photo display and the assault; and she offered reasons for her certainty. The fact that when asked to rate her certainty "on a scale of one to ten" she answered "seven or eight" cannot be decisive unless, like the majority, one is ready to substitute an unreliable, quantitative test of certainty for the intensely qualitative standard of proof beyond a reasonable doubt. Even in popular usage, the "scale of one to ten" as an indicator of belief is notoriously imprecise. People who in any ultimate—and unascertainable—sense probably share the same level of conviction may translate that very differently into numbers, and even the same person will change his mind from one moment to the next in assigning a percentage to his belief. Treating "one to ten" as a decisive indicator of the sufficiency of identification evidence thus elevates to a legal standard a popular measure that makes no claim at all to precision. As Wigmore stated long ago in this context, "The truth is that no one has yet invented or discovered a mode of measurement for the intensity of human belief. Hence there can be yet no successful method of communicating intelligibly ... a sound method of self-analysis for one's belief." Here, for example, Ms. Freedhoff equated "seven or eight" with being "very certain"; for all we know, she thought that any higher number would approach mathematical or absolute certainty, something the reasonable doubt standard does not require. The trial judge wisely did not view her attempt to furnish a numerical equivalent for her belief as conclusive, and neither should we.

The judicial straw polls cited by the majority merely confirm the futility of defining a percentual range (or "ball-park," to quote the majority) within which proof beyond a reasonable doubt must lie. Had Ms. Freedhoff added five percent to her belief-assessment (as much as "85%" rather than as much as "80%"), she would have come well within the range of, for example, Judge Weinstein's survey in Fatico. A factfinder's evaluation of credibility and intensity of belief should not be overridden by such inexact and even trivial differences of quantification.

Another aspect of the majority's opinion requires comment. It points to "[t]he vagaries of eyewitness identification," explains that this was a case of "cross-racial identification by a white woman of a black teenager," and cites to the "well established ... comparative difficulty in recognizing individual members of a race different from one's own." [quoting ELIZABETH LOFTUS, EYEWITNESS TESTIMONY]. It is not clear what the majority means by this discussion. The present appeal is not about whether a trier of fact may hear expert testimony or be instructed regarding the uncertainties of eyewitness identification, cross-racial or any other. Here the majority holds the identification insufficient as a matter of law, which implies that the trier of fact was required to discount the identification to an (undefined) extent because of the intrinsic weakness of eyewitness identifications generally or because this one was cross-racial. Either basis would be unprecedented. If, as I prefer to believe, that is not what the majority intends, then I respectfully suggest that the entire discussion of the point is dictum.

I would affirm the judgment of the trial court

References

[1] Bentham, J. (1827). *Rationale of judicial evidence, specially applied to English practice* (J. S. Mill, Ed.). London: Hunt and Clarke.

[2] Fienberg, S. E. (Ed.). (1988). *The evolving role of statistical assessments as evidence in the courts.* New York: Springer-Verlag.

[3] Starkie, T. (1833). *A practical treatise of the law of evidence and digest of proofs, in civil and criminal proceedings* (2nd ed.). London: J. and W. T. Clarke.

Module 10: Sleuthing with Probability and Statistics

My mother made me a scientist without ever intending to. Every Jewish mother in Brooklyn would ask her child after school: 'So? Did you learn anything today?' But not my mother. She always asked me a different question. 'Izzy,' she would say, 'did you ask a good question today?'

— Isidor Tabi (Nobel Prize in Physics, 1944; quotation given by John Barell in *Developing More Curious Minds*, 2003)

Abstract: Statistical sleuthing is concerned with the use of various probabilistic and statistical tools and methods to help explain or "tell the story" about some given situation. In this type of statistical detective work, a variety of probability distributions can prove useful as models for a given underlying process. These distributions include the Bernoulli, binomial, normal, Poisson (especially for spatial randomness and the assessment of "Poisson clumping"). Other elucidating probabilistic topics introduced include Benford's Law, the "birthday probability model," survival analysis and Kaplan-Meier curves, the Monty Hall problem, and what is called the "secretary problem" (or more pretentiously, the "theory of optimal stopping"). An amusing instance of the latter secretary problem is given as a *Car Talk* Puzzler called the "Three Slips of Paper"; a full listing of the script from the NPR show is included that aired on February 12, 2011.

Contents

1 Sleuthing Interests and Basic Tools

Modern statistics is often divided into two parts: exploratory and confirmatory. Confirmatory methods were developed over the first half of the 20th century, principally by Karl Pearson and Ronald Fisher. This was, and remains, a remarkable intellectual accomplishment. The goal of confirmatory methods is largely judicial: they are used to weigh evidence and make decisions. The aim of exploratory methods is different. They are useful in what could be seen as detective work; data are gathered and clues are sought to enable us to learn what might have happened. Exploratory analysis generates the hypotheses that are tested by the confirmatory methods. Surprisingly, the codification, and indeed the naming of exploratory data analysis, came after the principal work on the development of confirmatory methods was complete. John Tukey's (1977) influential book changed everything. He taught us that we should understand what might be true before we learn how well we have measured it.

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 (Stigler, 1986, pp. 215–216);

(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).

A variety of sleuthing approaches are available to help explain what might be occurring over a variety of different contexts. Some of those discussed in this monograph include Simpson's Paradox, Bayes' rule and base rates, regression toward the mean, the effects of culling on the identification of false positives and the subsequent inability to cross-validate, the operation of randomness and the difficulty in "faking" such a process, and confusions caused by misinterpreting conditional probabilities. 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 (Benford, 1938); 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.[1]

[1] 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. This is good news brought to you by Probability and Statistics!

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. For example, suppose a Bernoulli process produces a sequence of size n with r heads and $n - r$ tails. All arrangements of the r Hs and $n - r$ Ts should be equally likely (cutting, say, various sequences of size n all having r Hs and $n - r$ Ts from a much longer process); if not, possibly the process is not stationary or the assumption of independence is inappropriate. A similar use of the binomial would first estimate p from the long sequence, and then use this value to find the expected number of heads in sequences of a smaller size n; a long sequence could be partitioned into segments of this size and the observed number of heads compared to what would be expected. Again, a lack of fit between the observed and expected might suggest lack of stationarity or trial dependence (a more formal assessment of fit could be based on the usual chi-square goodness-of-fit test).

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 1 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

Table 1: Flying-bomb hits on London.

Number of hits	0	1	2	3	4	5 or more
Number of areas	229	211	93	35	7	1
Expected number	226.74	211.39	98.54	30.62	7.14	1.57

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 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").

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:

$$E(R) = \frac{2n_1n_2}{n_1 + n_2} + 1 \; ;$$

$$V(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} \; ;$$

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.

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.[2]

[2] The International Society for Clinical Biostatistics through its Subcommittee on Fraud published a position paper entitled "The Role of Biostatistics in the Prevention, Detection, and Treatment of Fraud in Clinical Trials" (Buyse et al., *Statistics in Medicine*, 1999, *18*, 3435–3451). Its purpose was to point out some of the ethical responsibilities the statistical community has in helping monitor clinical studies with public or personal health implications. The abstract is given below, but we still refer the reader directly to the article for more detail on a range of available statistical sleuthing tools (including Benford's law) that can assist in uncovering data fabrication and falsification:

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:

Recent cases of fraud in clinical trials have attracted considerable media attention, but relatively little reaction from the biostatistical community. In this paper we argue that biostatisticians should be involved in preventing fraud (as well as unintentional errors), detecting it, and quantifying its impact on the outcome of clinical trials. We use the term "fraud" specifically to refer to data fabrication (making up data values) and falsification (changing data values). Reported cases of such fraud involve cheating on inclusion criteria so that ineligible patients can enter the trial, and fabricating data so that no requested data are missing. Such types of fraud are partially preventable through a simplification of the eligibility criteria and through a reduction in the amount of data requested. These two measures are feasible and desirable in a surprisingly large number of clinical trials, and neither of them in any way jeopardizes the validity of the trial results. With regards to detection of fraud, a brute force approach has traditionally been used, whereby the participating centres undergo extensive monitoring involving up to 100 per cent verification of their case records. The cost-effectiveness of this approach seems highly debatable, since one could implement quality control through random sampling schemes, as is done in fields other than clinical medicine. Moreover, there are statistical techniques available (but insufficiently used) to detect "strange" patterns in the data including, but no limited to, techniques for studying outliers, inliers, overdispersion, underdispersion and correlations or lack thereof. These techniques all rest upon the premise that it is quite difficult to invent plausible data, particularly highly dimensional multivariate data. The multicentric nature of clinical trials also offers an opportunity to check the plausibility of the data submitted by one centre by comparing them with the data from all other centres. Finally, with fraud detected, it is essential to quantify its likely impact upon the outcome of the clinical trial. Many instances of fraud in clinical trials, although morally reprehensible, have a negligible impact on the trial's scientific conclusions. (pp. 3435–3436)

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."

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. (p. 65)

2 Survival Analysis

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 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. Explicitly, let the observed times of death for the N members under study be $t_1 \leq t_2 \leq \cdots \leq t_N$. Corresponding to each t_i is the number of members, n_i, "at risk" just prior to t_i; d_i is the number of deaths at time t_i. The Kaplan–Meier nonparametric maximum likelihood estimator, $\widehat{S(t)}$, is a product:

$$\widehat{S(t)} = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i}) \ .$$

When there is no right-censoring, n_i is just the number of survivors prior to time t_i; otherwise, n_i is the number of survivors minus the number of censored cases (by that time t_i). Only those surviving cases are still being observed (that is, not yet censored), and thus at risk of death. The function $\widehat{S(t)}$ is a nonincreasing step function, with steps at $t_i, 1 \leq i \leq N$; it is also usual to indicate the censored observations with tick marks on the graph of $\widehat{S(t)}$.

The original Kaplan and Meir article that appeared in 1958 (Kaplan, E. L., & Meier, P., "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, *53*, 457–481), is one of the most heavily cited papers in all of the sciences. It was featured as a "Citation Classic" in the June 13, 1983 issue of *Current Contents: Life Sciences*. As part of this recognition, Edward Kaplan wrote a short retrospective that we excerpt below:

This paper began in 1952 when Paul Meier at Johns Hopkins University (now at the University of Chicago) encountered Greenwood's paper on the duration

> of cancer. A year later at Bell Telephone Laboratories I became interested in the lifetimes of vacuum tubes in the repeaters in telephone cables buried in the ocean. When I showed my manuscript to John W. Tukey, he informed me of Meier's work, which already was circulating among some of our colleagues. Both manuscripts were submitted to the *Journal of the American Statistical Association*, which recommended a joint paper. Much correspondence over four years was required to reconcile our differing approaches, and we were concerned that meanwhile someone else might publish the idea.
>
> The nonparametric estimate specifies a discrete distribution, with all the probability concentrated at a finite number of points, or else (for a large sample) an actuarial approximation thereto, giving the probability in each of a number of successive intervals. This paper considers how such estimates are affected when some of the lifetimes are unavailable (censored) because the corresponding items have been lost to observation, or their lifetimes are still in progress when the data are analyzed. Such items cannot simply be ignored because they may tend to be longer-lived than the average. (p. 14)

To indicate the importance of the Kaplan–Meier estimator in sleuthing within the medical/pharmaceutical areas and elsewhere, we give the two opening paragraphs of Malcolm Gladwell's *New Yorker* article (May 17, 2010), entitled "The Treatment: Why Is It So Difficult to Develop Drugs for Cancer?":

> In the world of cancer research, there is something called a Kaplan–Meier curve, which tracks the health of patients in the trial of an experimental drug. In its simplest version, it consists of two lines. The first follows the patients in the "control arm," the second the patients in the "treatment arm." In most cases, those two lines are virtually identical. That is the sad fact of cancer research: nine times out of ten, there is no difference in survival between those who were given the new drug and those who were not. But every now and again—after millions of dollars have been spent, and tens of thousands of pages of data collected, and patients followed, and toxicological issues examined, and safety issues resolved, and manufacturing processes fine-tuned—the patients in the treatment arm will live longer than the patients in

> the control arm, and the two lines on the Kaplan–Meier will start to diverge.
>
> Seven years ago, for example, a team from Genentech presented the results of a colorectal-cancer drug trial at the annual meeting of the American Society of Clinical Oncology—a conference attended by virtually every major cancer researcher in the world. The lead Genentech researcher took the audience through one slide after another—click, click, click—laying out the design and scope of the study, until he came to the crucial moment: the Kaplan–Meier. At that point, what he said became irrelevant. The members of the audience saw daylight between the two lines, for a patient population in which that almost never happened, and they leaped to their feet and gave him an ovation. Every drug researcher in the world dreams of standing in front of thousands of people at ASCO and clicking on a Kaplan–Meier like that. "It is why we are in this business," Safi Bahcall says. Once he thought that this dream would come true for him. It was in the late summer of 2006, and is among the greatest moments of his life. (p. 69)

A great deal of additional statistical material involving survival functions can be helpful in our sleuthing endeavors. Survival functions may be compared over samples (for example, the log-rank test), and generalized to accommodate different forms of censoring; the Kaplan–Meier estimator has a closed-form variance estimator (for example, the Greenwood formula); various survival models can incorporate a mechanism for including covariates (for example, the proportional hazard models introduced by Sir David Cox; see Cox and Oakes (1984): *Analysis of Survival Data*). All of the usual commercial software (SAS, SPSS, SYSTAT) include modules for survival analysis. And, as might be expected, a plethora of cutting edge routines are in R, as well as in the Statistics Toolbox in MATLAB.

3 Sleuthing in the Media

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. We give two examples of this that may help: one is called the Monty Hall problem; the second is termed the Secretary problem.

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:[3]

> 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? (p. 16)

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

[3] As an interesting historical note, the “Monty Hall” problem has been a fixture of probability theory from at least the 1890s; it was named the problem of the “three caskets” by Henri Poincaré, and is more generally known as (Joseph) Bertrand’s Box Paradox

doubles the probability of winning from 1/3 to 2/3.[4]

An enjoyable diversion on Saturday mornings is the NPR radio show, *Car Talk*, with Click and Clack, The Tappet Brothers (aka Ray and Tom Magliozzi). A regular feature of the show, besides giving advice on cars, is The Puzzler; a recent example on Febuary 12, 2011 gives us another chance to "do the math." It is called the Three Slips of Paper, and it is stated as follows on the Car Talk website:

Three different numbers are chosen at random, and one is written on each of three slips of paper. The slips are then placed face down on the table. You have to choose the slip with the largest number. How can you improve your odds?

The answer given on the show:

Ray: This is from Norm Leyden from Franktown, Colorado. The date on it is 1974—I'm a little behind.

[4] To show the reach of the Monty Hall problem, we give the abstract for an article by Herbranson and Schroeder (2010): "Are Birds Smarter Than Mathematicians? Pigeons (*Columba livia*) Perform Optimally on a Version of the Monty Hall Dilemma" (*Journal of Comparative Psychology*, *124*, 1–13):

The "Monty Hall Dilemma" (MHD) is a well known probability puzzle in which a player tries to guess which of three doors conceals a desirable prize. After an initial choice is made, one of the remaining doors is opened, revealing no prize. The player is then given the option of staying with their initial guess or switching to the other unopened door. Most people opt to stay with their initial guess, despite the fact that switching doubles the probability of winning. A series of experiments investigated whether pigeons (*Columba livia*), like most humans, would fail to maximize their expected winnings in a version of the MHD. Birds completed multiple trials of a standard MHD, with the three response keys in an operant chamber serving as the three doors and access to mixed grain as the prize. Across experiments, the probability of gaining reinforcement for switching and staying was manipulated, and birds adjusted their probability of switching and staying to approximate the optimal strategy. Replication of the procedure with human participants showed that humans failed to adopt optimal strategies, even with extensive training. (p. 1)

Three different numbers are chosen at random, and one is written on each of three slips of paper. The slips are then placed face down on the table. The objective is to choose the slip upon which is written the largest number.

Here are the rules: You can turn over any slip of paper and look at the amount written on it. If for any reason you think this is the largest, you're done; you keep it. Otherwise you discard it and turn over a second slip. Again, if you think this is the one with the biggest number, you keep that one and the game is over. If you don't, you discard that one too.

Tommy: And you're stuck with the third. I get it.

Ray: The chance of getting the highest number is one in three. Or is it? Is there a strategy by which you can improve the odds?

Ray: Well, it turns out there is a way to improve the odds—and leave it to our pal Vinnie to figure out how to do it. Vinnie's strategy changes the odds to one in two. Here's how he does it: First, he picks one of the three slips of paper at random and looks at the number. No matter what the number is, he throws the slip of paper away. But he remembers that number. If the second slip he chooses has a higher number than the first, he sticks with that one. If the number on the second slip is lower than the first number, he goes on to the third slip.

Here's an example. Let's say for the sake of simplicity that the three slips are numbered 1000, 500, and 10.

Let's say Vinnie picks the slip with the 1000. We know he can't possibly win because, according to his rules, he's going to throw that slip out. No matter what he does he loses, whether he picks 500 next or 10. So, Vinnie loses—twice.

Now, let's look at what happens if Vinnie starts with the slip with the 500 on it. If he picks the 10 next, according to his rules, he throws that slip away and goes to the 1000.

Tommy: Whopee! He wins.

Ray: Right. And if Vinnie picks the 1000 next, he wins again!

Finally, if he picks up the slip with the 10 on it first, he'll do, what?

Tommy: Throw it out. Those are his rules.

Ray: Right. And if he should be unfortunate enough to pick up the one that says 500 next, he's going to keep it and he's going to lose. However, if his second choice is not the 500 one but the 1000 one, he's gonna keep that slip—and he'll win.

If you look at all six scenarios, Tommy will win one in three times, while Vinnie will win three times out of six.

Tommy: That's almost half!

Ray: In some countries.

One particularly rich area in probability theory that extends the type of *Car Talk* example just given is in the applied probability topic known as optimal stopping, or more colloquially, "the secretary problem." We paraphrase the simplest form of this problem from Thomas Ferguson's review paper in *Statistical Science* (1989), "Who Solved the Secretary Problem?": There is one secretarial position to be filled from among n applicants who are interviewed sequentially and in a random order. All applicants can be ranked from best to worse, with the choice of accepting an applicant based only on the relative ranks of those interviewed thus far. Once an applicant has been rejected, that decision is irreversible. Assuming the goal is to maximize the probability of selecting the absolute best applicant, it can be shown that the selection rules can be restricted to a class of strategies defined as follows: for some integer $r \geq 1$, reject the first $r - 1$ applicants and select the next who is best in the relative ranking of the applicants interviewed thus far. The probability of selecting the best applicant is $1/n$ for $r = 1$; for $r > 1$, it is

$$(\frac{r-1}{n}) \sum_{j=r}^{n} \frac{1}{j-1} \ .$$

For example, when there are 5 ($= n$) applicants, the probabilities of choosing the best for values of r from 1 to 5 are given in the following table:

r	Probability
1	$1/5 = .20$
2	$5/12 \approx .42$
3	$13/30 \approx .43$
4	$7/20 = .35$
5	$1/5 = .20$

Thus, because an r value of 3 leads to the largest probability of about .43, it is best to interview and reject the first two applicants and then pick the next relatively best one. For large n, it is (approximately) optimal to wait until about 37% ($\approx 1/e$) of the applicants have been interviewed and then select the next relatively best one. This also gives the probability of selecting the best applicant as .37 (again, $\approx 1/e$).

In the *Car Talk* Puzzler discussed above, $n = 3$ and Vinnie uses the rule of rejecting the first "interviewee" but then selects the next that is relatively better. The probability of choosing the best therefore increases from 1/3 to 1/2.

Any beginning statistics class should always include a number of formal tools to help work through puzzling situations. Several of these are mentioned elsewhere in this monograph: 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.

References

[1] Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, *78*, 551–572.

[2] Bernhardt, D., & Heston, S. (2010). Point shaving in college basketball: A cautionary tale for forensic economics. *Economic Inquiry*, *48*, 14–25.

[3] Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed., Vol. 1). New York: Wiley.

[4] Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. Reading, MA: Addison-Wesley.

[5] Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press / Harvard University Press.

[6] Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

[7] Wolfers, J. (2006). Point shaving: Corruption in NCAA basketball. *American Economic Review*, *96*, 279–283.

Module 11: Cross-validation and the Control of Error Rates

As is your sort of mind, so is your sort of search; you'll find what you desire.
— Robert Browning (1812–1889)

Abstract: This module emphasizes what might be termed "the practice of safe statistics." The discussion is split into three parts: (1) the importance of cross-validation for any statistical method that relies on an optimization process based on a given data set (or sample); (2) the need to exert control on overall error rates when carrying out multiple testing, even when that testing is done only implicitly; (3) in the context of "big data" and associated methods for "data mining," the necessity of some mechanism for ensuring the replicability of "found results."

Contents

1 Cross-validation

Many texts in statistics that include a discussion of (multiple) regression and related techniques give little weight to the topic of cross-validation, which we believe is crucial to the appropriate (and "safe") use of these methods.[1] Cross-validation might be discussed under the rubric of how does a result found for a particular sample of data "hold up" in a new sample. As a general illustration, consider (multiple) regression where the interest is in predicting a single dependent measure, Y, from a linear combination of K independent variables, $X_1, \ldots, X_K$. As a measure of how well a regression equation does in the sample, we typically use the squared correlation (R^2) between the values on Y and those predicted from the regression equation, say, $\hat{Y}$. This is a measure of how well an equation does on the same data set from which it was derived, typically through an optimization process of least-squares. Our real interest, however, may be in how well or badly the sample equation works generally. The sample equation has been optimized with respect to the particular data at hand, and therefore, it might be expected that the squared correlation is "inflated." In other words, the concern is with sample equation performance in a new group; this is the quintessential question of "cross-validation."

[1] We have one salient example in Module 2 where a lack of cross-validation lead to overly-optimistic estimates of how well actuarial predictions of violence could be made. This was the development of the COVR instrument in the MacArthur Study of Mental Disorder and Violence. In the training sample, 1 out of 3 predictions of "violence" were wrong; in the one small cross-validation study done somewhat later using completely "new" data, 2 out of 3 predictions of "violence" were incorrect. In fact, the COVR even failed to be clinically efficient in the Meehl and Rosen sense – the diagnostic test was outperformed by prediction using simple base rates.

There are several general strategies that can be used to approach the task of cross-validation:

a) Get *new* data and use the sample equation to predict Y and calculate the squared correlation between Y and $\hat{Y}$; denoting this squared correlation by R^2_{new}, the difference $R^2 - R^2_{new}$ is called "shrinkage" and measures the drop in how well one can predict with new data. The chief problem with this first approach is that new data may be "hard to come by" and/or very expensive.

b) We can first split the original sample into two parts; obtain the equation on one part (the "training set") and test how well it does on the second (the "test set"). This is a common method of cross-validation; the only possible down-side is when the original sample is not very big, and the smaller training sample might produce a more unstable equation than desirable.

c) *Sample reuse methods*: here, the original sample is split into K parts, with the equation obtained with $K-1$ of the parts aggregated together and then tested on the one part left out. This process is repeated K times, leaving one of the K parts out each time; it is called K-fold cross-validation. Given the increased computational power that is now readily available, this K-fold cross-validation is close to being a universal default option (and with K usually set at around 10).

At the extreme, if n subjects are in the original sample, n-fold cross-validation would leave one person out at a time. For this person left out, say person i, we obtain $\hat{Y}_i$ and then calculate the squared correlation between the Y_i's and $\hat{Y}_i$'s to see how well we cross-validate with a “new” sample. Each equation is constructed with $n - 1$ subjects so there should be more stability present than in approach (b).

1.1 An Example of a Binary Classifier

The term *discrimination* (in a nonpejorative statistical sense) can refer to the task of separating groups through linear combinations of variables maximizing a criterion, such as an F-ratio. The linear combinations themselves are commonly called Fisher’s linear discriminant functions. The related term *classification* refers to the task of allocating observations to existing groups, typically to minimize the cost and/or probability of misclassification. These two topics are intertwined, but here we briefly comment on the topic of classification when there are two groups (or in the current jargon, we will construct a “binary classifier”).

In the simple two-group situation, there are two populations, π_1 and π_2; π_1 is assumed to be characterized by a normal distribution with mean μ_1 and variance σ_X^2 (the density is denoted by $f_1(x)$); π_2 is characterized by a normal distribution with mean μ_2 and (common) variance σ_X^2 (the density is denoted by $f_2(x)$). Given an observation, say x_0, we wish to decide whether it should be assigned to π_1 or to π_2. Assuming that $\mu_1 \leq \mu_2$, a criterion point c is chosen; the rule then becomes: allocate to π_1 if $x_0 \leq c$, and to π_2 if $> c$. The probabilities of misclassification are given in the following chart:

		True	State
		π_1	π_2
Decision	π_1	$1-\alpha$	β
	π_2	α	$1-\beta$

In the terminology of our previous usage of Bayes' rule to obtain the positive predictive value of a test, and assuming that π_1 refers to a person having "it," and π_2 to not having "it," the sensitivity of the test is $1-\alpha$ (true positive); specificity is $1-\beta$, and thus, β refers to a false negative and α to a false positive.

To choose c so that $\alpha+\beta$ is smallest, select the point at which the densities are equal. A more complicated way of stating this decision rule is to allocate to π_1 if $f_1(x_0)/f_2(x_0) \geq 1$; if < 1, then allocate to π_2. Suppose now that the prior probabilities of being drawn from π_1 and π_2 are p_1 and p_2, respectively, where $p_1+p_2=1$. If c is chosen so the Total Probability of Misclassification (TPM) is minimized (that is, $p_1\alpha + p_2\beta$), the rule would be to allocate to π_1 if $f_1(x_0)/f_2(x_0) \geq p_2/p_1$; if $< p_2/p_1$, then allocate to π_2. Finally, to include costs of misclassification, $c(1|2)$ (for assigning to π_1 when actually coming from π_2), and $c(2|1)$ (for assigning to π_2 when actually coming from π_1), choose c to minimize the Expected Cost of Misclassification (ECM), $c(2|1)p_1\alpha + c(1|2)p_1\beta$, by the rule of allocating to π_1 if $f_1(x_0)/f_2(x_0) \geq (c(1|2)/c(2|1))(p_2/p_1)$; if $<$ $(c(1|2)/c(2|1))(p_2/p_1)$, then allocate to π_2.

Using logs, the last rule can be restated:

allocate to π_1 if $\log(f_1(x_0)/f_2(x_0)) \geq \log((c(1|2)/c(2|1))(p_2/p_1))$.

The left-hand side is equal to

$$(\mu_1 - \mu_2)(\sigma_X^2)^{-1} x_0 - (1/2)(\mu_1 - \mu_2)(\sigma_X^2)^{-1}(\mu_1 + \mu_2),$$

so the rule can be rephrased further:

allocate to π_1 if

$$x_0 \leq \{(1/2)(\mu_1 - \mu_2)(\sigma_X^2)^{-1}(\mu_1 + \mu_2) -$$

$$\log((c(1|2)/c(2|1))(p_2/p_1))\{\frac{\sigma_X^2}{-(\mu_1 - \mu_2)}\}$$

or

$$x_0 \leq \{(1/2)(\mu_1+\mu_2) - \log((c(1|2)/c(2|1))(p_2/p_1))\}\{\frac{\sigma_X^2}{(\mu_2 - \mu_1)}\} = c\,.$$

If the costs of misclassification are equal (that is, $c(1|2) = c(2|1)$), then the allocation rule is based on classification functions: allocate to π_1 if

$$[\frac{\mu_1}{\sigma_X^2} x_0 - (1/2)\frac{\mu_1^2}{\sigma_X^2} + \log(p_1)] - [\frac{\mu_2}{\sigma_X^2} x_0 - (1/2)\frac{\mu_2^2}{\sigma_X^2} + \log(p_2)] \geq 0\ .$$

The classifier just constructed has been phrased using population parameters, but to obtain a sample-based classifier, estimates are made for the population means and variances. Alternatively, a "dummy" binary dependent variable Y ($= 0$ for an observation in group 1; $= 1$ for an observation in group 2) can be predicted from X; the sample-based classifier is obtained in this way. Also, this process of using a binary Y but with K independent variables, $X_1, \ldots, X_K$, leads to a binary classifier based on more than one independent variable (and to what is called Fisher's linear discriminant function).[2]

[2] In the terminology of signal detection theory and the general problem of yes/no diag-

In moving to the sample where estimated quantities (sample means, variances, and covariances) are used for the population parameters, we can do more than just hope that the (sample) classification rule does well by carrying out a cross-validation. First, a misclassification table can be constructed based on simple resubstitution of the original data into the sample classification rule (where n_1 observations are in group π_1 and n_2 are in group π_2):

		True	State
		π_1	π_2
Decision	π_1	a	b
	π_2	c	d
sums		n_1	n_2

The apparent error rate (APR) is $(b + c)/n$, which is overly optimistic because it is optimized with respect to this sample. A K-fold cross-validation would give a less optimistic estimate; for example, letting $K = n$ and using the "hold out one-at-a-time" strategy, the following misclassification table might be obtained:

nostic decisions as discussed in Module 4, a plot of sensitivity (true positive probability) on the y-axis against $1-$ specificity on the x-axis as c varies, is an ROC curve (for Receiver Operating Characteristic). This ROC terminology originated in World War II in detecting enemy planes by radar (group π_1) from the noise generated by random interference (group π_2). The ROC curve is bowed from the origin of (0, 0) at the lower-left corner to (1.0, 1.0) at the upper right; it indicates the trade-off between increasing the probability of true positives and the increase of false positives. Generally, the adequacy of a particular diagnostic decision strategy is measured by the area under the ROC curve, with areas closer to 1.0 being better; that is, steeper bowed curves hugging the left wall and the top border of the square box. For a comprehensive introduction to diagnostic processes, see Swets, Dawes, and Monahan (2000).

		True	State
		π_1	π_2
Decision	π_1	a*	b*
	π_2	c*	d*
sums		n_1	n_2

To estimate the actual error rate (AER), we would use $(b^* + c^*)/n$, and would expect it to be greater than the APR.

2 Problems With Multiple Testing

A difficulty encountered with the use of automated software analyses is that of multiple testing, where the many significance values provided are all given as if each were obtained individually without regard for how many tests were performed. This situation gets exacerbated when the "significant" results are then culled, and only these are used in further analysis. A good case in point is reported in the next section on odd correlations where highly inflated correlations get reported in fMRI studies because an average is taken only over those correlations selected to have reached significance according to a stringent threshold. Such a context is a clear violation of a dictum given in many beginning statistics classes: you cannot legitimately test a hypothesis on the same data that first suggested it.

Exactly the same issue manifests itself, although in a more subtle, implicit form, in the modern procedure known as data mining. Data mining consists of using powerful graphical and algorithmic methods to view and search through high-dimensional data sets of moderate-to-large size, looking for interesting features. When such a feature is uncovered, it is isolated and saved. Implicit in the search, however, are many comparisons that the viewer makes and decides are not interesting. Because the searching and comparing is done in real time, it is difficult to keep track of how many "insignificant" comparisons were discarded before alighting on a significant one. Without knowing how many, we cannot judge the significance of the interesting features found without an independent confirmatory sample. Such independent confirmation is all too rarely done.

Uncontrolled data mining and multiple testing on some large (longitudinal) data sets can also lead to results that might best be labeled with the phrase "the oat bran syndrome." Here, a promising association is identified; the relevant scientists appear in the media and on various cable news shows; and an entrepreneurial industry is launched to take advantage of the supposed findings. Unfortunately, some time later, contradictory studies appear, possibly indicating a downside of the earlier recommendations, or at least no replicable effects of the type reported previously. The name "the oat bran syndrome" results from the debunked studies from the 1980s that had food manufacturers adding oat bran to absolutely everything, including beer, to sell products to people who wanted to benefit from the fiber that would supposedly prevent cancer.

To be more formal about the problem of multiple testing, suppose there are K hypotheses to test, $H_1, \ldots, H_K$, and for each, we set the criterion for rejection at the fixed Type I error value of α_k, $k = 1, \ldots, K$. If the event A_k is defined as the incorrect rejection of H_k

(that is, rejection when it is true), the Bonferroni inequality gives

$$P(A_1 \text{ or } \cdots \text{ or } A_K) \leq \sum_{k=1}^{K} P(A_k) = \sum_{k=1}^{K} \alpha_k \ .$$

Noting that the event (A_1 or $\cdots$ or A_K) can be verbally restated as one of "rejecting incorrectly *one or more* of the hypotheses," the experiment-wise (or overall) error rate is bounded by the sum of the K α values set for each hypothesis. Typically, we let $\alpha_1 = \cdots = \alpha_K = \alpha$, and the bound is then $K\alpha$. Thus, the usual rule for controlling the overall error rate through the Bonferroni correction sets the individual αs at some small value such as $.05/K$; the overall error rate is then guaranteed to be no larger than .05.

The problem of multiple testing and the failure to practice "safe statistics" appears in both blatant and more subtle forms. For example, companies may suppress unfavorable studies until those to their liking occur. A possibly apocryphal story exists about toothpaste companies promoting fluoride in their products in the 1950s and who repeated studies until large effects could be reported for their "look Ma, no cavities" television campaigns. This may be somewhat innocent advertising hype for toothpaste, but when drug or tobacco companies engage in the practice, it is not so innocent and can have a serious impact on our collective health. It is important to know how many things were tested to assess the importance of those reported. For example, when given only those items from some inventory or survey that produced significant differences between groups, be very wary!

People sometimes engage in a number of odd behaviors when doing multiple testing. We list a few of these below in summary form:

(a) It is not legitimate to do a Bonferroni correction post hoc; that is, find a set of tests that lead to significance, and then evaluate just this subset with the correction;

(b) Scheffé's method (and relatives) are the only true post-hoc procedures to control the overall error rate. An unlimited number of comparisons can be made (no matter whether identified from the given data or not), and the overall error rate remains constant;

(c) You cannot look at your data and then decide which planned comparisons to do;

(d) Tukey's method is not post hoc because you actually plan to do all possible pairwise comparisons;

(e) Even though the comparisons you might wish to test are independent (such as those defined by orthogonal comparisons), the problem of inflating the overall error rate remains; similarly, in performing a multifactor analysis of variance (ANOVA) or testing multiple regression coefficients, all of the tests carried out should have some type of control imposed on the overall error rate;

(f) It makes little sense to perform a multivariate analysis of variance before you go on to evaluate each of the component variables. Typically, a multivariate analysis of variance (MANOVA) is completely noninformative as to what is really occurring, but people proceed in any case to evaluate the individual univariate ANOVAs irrespective of what occurs at the MANOVA level; we may accept the null hypothesis at the overall MANOVA level but then illogically ask where the differences are at the level of the individual variables. Plan to do the individual comparisons beforehand, and avoid the uninterpretable overall MANOVA test completely.

We cannot leave the important topic of multiple comparisons without at least a mention of what is now considered one of the more powerful methods currently available: the False Discovery Rate (Benjamini & Hochberg, 1995). But even this method is not up to the most vexing of problems of multiplicity. We have already mentioned data mining as one of these; a second problem arises in the search for genetic markers. A typical paradigm in this crucial area is to isolate a homogeneous group of individuals, some of whom have a genetic disorder and others do not, and then to see if one can determine which genes are likely to be responsible. One such study is currently being carried out with a group of 200 Mennonites in Pennsylvania. Macular degeneration is common among the Mennonites, and this sample was chosen so that 100 of them had macular degeneration and a matched sample of 100 did not. The genetic structure of the two groups was very similar, and so the search was on to see which genes were found much more often in the group that had macular degeneration than in the control group. This could be determined with a t-test. Unfortunately, the usefulness of the t-test was diminished considerably when it had to be repeated for more than 100,000 separate genes. The Bonferroni inequality was no help, and the False Discovery Rate, while better, was still not up to the task. The search still goes on to find a better solution to the vexing problem of multiplicity.[3]

[3] The probability issues involved with searching through the whole genome are discussed in: "Nabbing Suspicious SNPS: Scientists Search the Whole Genome for Clues to Common

3 Odd Correlations

A recent article (Vul et al. 2009) in *Perspectives on Psychological Science*, has the intriguing title, “Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition” (renamed from the earlier and more controversial “Voodoo Correlations in Social Neuroscience”; note that the acronym fMRI stands for functional Magnetic Resonance Imaging, and is always written with a lower-case letter “f”). These authors comment on the extremely high (for example, greater than .80) correlations reported in the literature between brain activation and personality measures, and point out the fallaciousness of how they were obtained. Typically, huge numbers of separate correlations were calculated, and only the mean of those correlations exceeding some threshold (based on a very small significance level) are reported. It is tautological that these correlations selected for size must then be large in their average value. With no cross-validation attempted to see the shrinkage expected in these measures on new samples, we have sophistry at best. Any of the usual understanding of yardsticks provided by the correlation or its square, the proportion of shared variance, are inappropriate. In fact, as noted by Vul et al. (2009), these inflated mean correlations typically exceed the upper bounds provided by the correction for attenuation based on what the reliabilities should be for the measures being correlated.

An amusing critique of fMRI studies that fail to correct for multiple comparisons and control false positives involves the scan of a dead salmon’s brain and its response to human emotions (“Trawling

Diseases” (Regina Nuzzo, *ScienceNews*, June 21, 2008).

the Brain," Laura Sanders, December 19, 2009, *ScienceNews*). The original article was published in the *Journal of Serendipitous and Unexpected Results* (Craig Bennett, et al., 2010, *1*, 1–6), with the long title "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction." This tongue-in-cheek piece provides a cautionary lesson for anyone involved with the interpretation of fMRI research. A dead salmon's brain can display much of the same beautiful red-hot areas of activity in response to emotional scenes flashed to the (dead) salmon that would be expected for (alive) human subjects. We give the abstract below.

> With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of at least one false positive is almost certain. Proper correction for multiple comparisons should be completed during the analysis of these datasets, but is often ignored by investigators. To highlight the danger of this practice we completed an fMRI scanning session with a post-mortem Atlantic Salmon as the subject. The salmon was shown the same social perspective taking task that was later administered to a group of human subjects. Statistics that were uncorrected for multiple comparisons showed active voxel clusters in the salmon's brain cavity and spinal column. Statistics controlling for the familywise error rate (FWER) and false discovery rate (FDR) both indicated that no active voxels were present, even at relaxed statistical thresholds. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing proper multiple comparisons correction as standard practice when thresholding their data.

For conducting the "dead-salmon" study, the main authors, Craig Bennett and Michael Miller, received a 2012 Ig Nobel prize. They were interviewed shortly thereafter by Scott Simon for NPR's Weekend Edition. The transcript of this interview follows:

Host Scott Simon speaks with Craig Bennett and Michael Miller about being awarded a 2012 Ig Nobel prize for their paper on the brain waves of dead Atlantic Salmon, published in the *Journal of Serendipitous and Unexpected Results*.

SCOTT SIMON, HOST:

In a couple weeks, the prestigious Nobel Prizes will be announced. But this week, the Ig Nobels honored the silliest discoveries of 2012. A study on the physics of the ponytail; a paper on why coffee spills when you walk; and a prize for a group of psychologists who scanned the brain of an unpromising patient: a deceased Atlantic salmon. Even more unlikely were their findings: the dead fish had thoughts. Who knows – maybe dreams. Craig Bennett did the experiment and accepted the award with good humor, and a couple of fish jokes.

CRAIG BENNETT: Some have called functional neuroimaging, which is an important method for studying the human brain, a fishing expedition. Some have even called the results a red herring. But ...

SIMON: Craig Bennett and his colleague, Dr. Michael Miller, joins us now from studios at Harvard University. Gentlemen, thanks for being with us.

MICHAEL MILLER: Thank you, Scott.

: Yeah, it's good to be here.

SIMON: Is there any defensible reason to study the brain of a dead fish?

MILLER: Well, not for genuine, functional brain activities there's not.

: We wanted to illustrate kind of the absurdity of improper statistical approaches, that you can find false positives, or what is essentially garbage results. And using the incorrect statistical approach you can actually see that there are voxels of activity in the dead, frozen salmon's brain.

MILLER: You know, while the salmon was in the scanner, we were doing the testing exactly like a human would have been in there.

SIMON: I'm sorry, did you say to the postmortem salmon, just press this button in case you get antsy?

: We actually did, because we were also training our research assistants on the proper methods on how to interact with humans. And so not only did we give the experimental instructions to the salmon but we also were on the intercom asking if the salmon was OK throughout the experiment.

SIMON: Did you just go into Legal Seafood and say give me a mackerel - forgive me, an Atlantic salmon?

MILLER: It was a Saturday morning and we were conducting the testing very early so that we didn't interrupt the running of humans later in the day. So, I walked into the local supermarket at 6:30 in the morning, and I said, excuse me, gentlemen, I need a full-length Atlantic salmon. And I'm not a morning person, I just kind of added - for science. And they kind of looked at me funny, but then they were like, you know, we'll be happy to oblige. That'll be $27.50, and before I knew it, I had a full-length Atlantic salmon that was ready to scan.

SIMON: Gentlemen, I'm sorry if this question sounds indelicate, but when your experimentation was done, grilled or poached?

: Baked. That was dinner that night.

(LAUGHTER)

SIMON: Well, science was served, I expect, right?

: And science was tasty.

SIMON: Craig Bennett and Michael Miller, University of California Santa Barbara, won the Ig Nobel Prize this week. They joined us from Harvard University. Gentlemen, thanks for being with us.

MILLER: Thank you, Scott.

: Thanks.

SIMON: You can hear more highlights from the Ig Nobel Awards later this fall on a special Thanksgiving edition of NPR's SCIENCE FRIDAY. This is NPR News.

There are several ways to do corrections for multiple comparisons in fMRI. One is through the false discovery method already mentioned (e.g., Benjamini and Hochberg, 1995); another is the class of methods that control the familywise error rate which includes the

Bonferroni correction strategy, random field theory, and a general method based on permutation procedures. This later approach is discussed in detail in ”Nonparametric Permutation Tests For Functional Neuroimaging: A Primer with Examples” (Thomas E. Nichols and Andrew P. Holmes; *Human Brain Mapping*, *15*, 2001, 1–25); the abstract for this paper follows:

> Requiring only minimal assumptions for validity, nonparametric permutation testing provides a flexible and intuitive methodology for the statistical analysis of data from functional neuroimaging experiments, at some computational expense. ... [T]he permutation approach readily accounts for the multiple comparisons problem implicit in the standard voxel-by-voxel hypothesis testing framework. When the appropriate assumptions hold, the nonparametric permutation approach gives results similar to those obtained from a comparable Statistical Parametric Mapping approach using a general linear model with multiple comparisons corrections derived from random field theory. For analyses with low degrees of freedom, such as single subject PET/SPECT experiments or multi-subject PET/SPECT or fMRI designs assessed for population effects, the nonparametric approach employing a locally pooled (smoothed) variance estimate can outperform the comparable Statistical Parametric Mapping approach. Thus, these nonparametric techniques can be used to verify the validity of less computationally expensive parametric approaches. Although the theory and relative advantages of permutation approaches have been discussed by various authors, there has been no accessible explication of the method, and no freely distributed software implementing it. Consequently, there have been few practical applications of the technique. This article, and the accompanying MATLAB software, attempts to address these issues. The standard nonparametric randomization and permutation testing ideas are developed at an accessible level, using practical examples from functional neuroimaging, and the extensions for multiple comparisons described. Three worked examples from PET and fMRI are presented, with discussion, and comparisons with standard parametric approaches made where appropriate. Practical considerations are given throughout, and rele-

vant statistical concepts are expounded in appendices.

4 Cautionary Summary Comments

As a reminder of the ubiquitous effects of searching/selecting/optimization, and the identification of "false positives," we have mentioned some blatant examples here and in earlier modules—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). We quote from the abstract of a fairly recent review, "The Scientific Status of Projective Techniques" (Lilienfeld, Wood, & Garb, 2000):

> Although some projective instruments were better than chance at detecting child sexual abuse, there were virtually no replicated findings across independent investigative teams. This meta-analysis also provides the first clear evidence of substantial file-drawer effects in the projectives literature, as the effect sizes from published studies markedly exceeded those from unpublished studies. (p. 27)

The general failure to replicate is being continually (re)documented both in the scientific literature and in more public venues. In medicine, there is the work of John Ioannidis:

"Contradicted and Initially Stronger Effects in Highly Cited Clinical Research" (*Journal of the American Medical Association*, 2005, *294*, 218–228);

"Why Most Published Research Findings Are False" (*PLoS Medicine*, 2005, *2*, 696–701).

"Why Most Discovered True Associations Are Inflated" (*Epidemiology*, 2008, *19*, 640–648).[4]

[4] This particular Ioannidis article covers much more than just the field of medicine; its message is relevant to the practice of probabilistic reasoning in science more generally. The abstract follows:

> Newly discovered true (non-null) associations often have inflated effects compared with the true effect sizes. I discuss here the main reasons for this inflation. First, theoretical considerations prove that when true discovery is claimed based on crossing a threshold of statistical significance and the discovery study is underpowered, the observed effects are expected to be inflated. This has been demonstrated in various fields ranging from early stopped clinical trials to genome-wide associations. Second, flexible analyses coupled with selective reporting may inflate the published discovered effects. The vibration ratio (the ratio of the largest vs. smallest effect on the same association approached with different analytic choices) can be very large. Third, effects may be inflated at the stage of interpretation due to diverse conflicts of interest. Discovered effects are not always inflated, and under some circumstances may be deflated – for example, in the setting of late discovery of associations in sequentially accumulated overpowered evidence, in some types of misclassification from measurement error, and in conflicts causing reverse biases. Finally, I discuss potential approaches to this problem. These include being cautious about newly discovered effect

In the popular media, we have the discussion of the "decline effect" by Jonah Lehrer in the *New Yorker* (December 13, 2010), "The Truth Wears Off (Is These Something Wrong With the Scientific Method?)"; or from one of the nation's national newspapers, "Low-Salt Diet Ineffective, Study Finds. Disagreement Abounds" (*New York Times*, Gina Kolata, May 3, 2011). We give part of the first sentence of Kolata's article: "A new study found that low-salt diets increase the risk of death from heart attacks and strokes and do not prevent high blood pressure."

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?). Because of the importance of this cautionary tale for behavioral genetics research generally, we reproduce below a *News of the Week*

sizes, considering some rational down-adjustment, using analytical methods that correct for the anticipated inflation, ignoring the magnitude of the effect (if not necessary), conducting large studies in the discovery phase, using strict protocols for analyses, pursuing complete and transparent reporting of all results, placing emphasis on replication, and being fair with interpretation of results.

item from *Science*, written by Constance Holden (2009), "Back to the Drawing Board for Psychiatric Genetics":[5]

> Geneticists have long been immersed in an arduous and largely fruitless search to identify genes involved in psychiatric disorders. In 2003, a team led by Avshalom Caspi, now at Duke University in Durham, North Carolina, finally landed a huge catch: a gene variant that seemed to play a major role in whether people get depressed in response to life's stresses or sail through. The find, a polymorphism related to the neurotransmitter serotonin, was heralded as a prime example of "gene-environment interaction": whereby an environmental trigger influences the activity of a gene in a way that confers susceptibility. "Everybody was excited about this," recalls Kathleen Merikangas, a genetic epidemiologist at the National Institute of Mental Health (NIMH) in Bethesda, Maryland. "It was very widely embraced." Because of the well-established link between serotonin and depression, the study offered a plausible biological explanation for why some people are so much more resilient than others in response to life stresses.
>
> But an exhaustive new analysis published last week in *The Journal of the American Medical Association* suggests that the big fish may be a minnow at best.
>
> In a meta-analysis, a multidisciplinary team headed by Merikangas and ge-

[5] The general problem of exaggerated initially-found effects for a marker-allele association is discussed by Peter Kraft in his article "Curses – Winner's and Otherwise – in Genetic Epidemiology" (*Epidemiology*, 2008, *19*, 649–651). The abstract follows:

> The estimated effect of a marker allele from the initial study reporting the marker-allele association is often exaggerated relative to the estimated effect in follow-up studies (the "winner's curse" phenomenon). This is a particular concern for genome-wide association studies, where markers typically must pass very stringent significance thresholds to be selected for replication. A related problem is the overestimation of the predictive accuracy that occurs when the same data set is used to select a multilocus risk model from a wide range of possible models and then estimate the accuracy of the final model ("over-fitting"). Even in the absence of these quantitative biases, researchers can over-state the qualitative importance of their findings – for example, by focusing on relative risks in a context where sensitivity and specificity may be more appropriate measures. Epidemiologists need to be aware of these potential problems: as authors, to avoid or minimize them, and as readers, to detect them.

neticist Neil Risch of the University of California, San Francisco, reanalyzed data from 14 studies, including Caspi's original, and found that the cumulative data fail to support a connection between the gene, life stress, and depression. It's "disappointing—of all the [candidates for behavior genes] this seemed the most promising," says behavioral geneticist Matthew McGue of the University of Minnesota, Twin Cities.

The Caspi paper concluded from a longitudinal study of 847 New Zealanders that people who have a particular variant of the serotonin transporter gene are more likely to be depressed by stresses, such as divorce and job loss (*Science*, 18 July 2003, pp. 291–293; 386–389). The gene differences had no effect on depression in the absence of adversity. But those with a "short" version of the gene—specifically, an allele of the promoter region of the gene—were more likely to be laid low by unhappy experiences than were those with two copies of the "long" version, presumably because they were getting less serotonin in their brain cells.

Subsequent research on the gene has produced mixed results. To try to settle the issue, Merikangas says, "we really went through the wringer on this paper." The group started with 26 studies but eliminated 12 for various reasons, such as the use of noncomparable methods for measuring depression. In the end, they reanalyzed and combined data from 14 studies, including unpublished data on individual subjects for 10 of them.

Of the 14 studies covering some 12,500 individuals, only three of the smaller ones replicated the Caspi findings. A clear relationship emerged between stressful happenings and depression in all the studies. But no matter which way they sliced the accumulated data, the Risch team found no evidence that the people who got depressed from adverse events were more likely to have the suspect allele than were those who didn't.

Caspi and co-author Terrie Moffitt, also now at Duke, defend their work, saying that the new study "ignores the complete body of scientific evidence." For example, they say the meta-analysis omitted laboratory studies showing that humans with the short allele have exaggerated biological stress responses and are more vulnerable to depression-related disorders such as anxiety and posttraumatic stress disorder. Risch concedes that his team had to omit several supportive studies. That's because, he says, they wanted to focus as much as possible on attempts to replicate the original research, with comparable measures of depression and stress.

Many researchers find the meta-analysis persuasive. "I am not surprised by their conclusions," says psychiatric geneticist Kenneth Kendler of Virginia Commonwealth University in Richmond, an author of one of the supportive studies that was excluded. "Gene discovery in psychiatric illness has been very hard, the hardest kind of science," he says, because scientists are looking for multiple genes with very small effects.

Dorrett Boomsma, a behavior geneticist at Amsterdam's Free University, points out that many people have questioned the Caspi finding. Although the gene was reported to have an effect on depression only in the presence of life stress, she thinks it is "extremely unlikely that it would not have an independent effect" as well. Yet recent whole-genome association studies for depression, for which scientists scan the genomes of thousands of subjects for tens of thousands of markers, she adds, "do not say anything about [the gene]."

Some researchers nonetheless believe it's too soon to close the book on the serotonin transporter. ... geneticist Joel Gelernter of Yale University agrees with Caspi that the rigorous demands of a meta-analysis may have forced the Risch team to carve away too much relevant material. And NIMH psychiatrist Daniel Weinberger says he's not ready to discount brain-imaging studies showing that the variant in question affects emotion-related brain activity.

Merikangas believes the meta-analysis reveals the weakness of the "candidate gene" approach: genotyping a group of subjects for a particular gene variant and calculating the effect of the variant on a particular condition, as was done in the Caspi study. "There are probably 30 to 40 genes that have to do with the amount of serotonin in the brain," she says. So "if we just pull out genes of interest, ... we're prone to false positives." Instead, she says, most geneticists recognize that whole-genome scans are the way to go. McGue agrees that behavioral gene hunters have had to rethink their strategies. Just in the past couple of years, he says, it's become clear that the individual genes affecting behavior are likely to have "much, much smaller effects" than had been thought.

References

[1] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289–300.

[2] Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington H. L., ... Poulton, R. (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science*, *301*, 386–399.

[3] Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest*, *1*, 27–66.

[4] Swets, J. A, Dawes, R. M., & Monahan, J. (2000b). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.

[5] Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Module 12: An Olio of Topics in Applied Probabilistic Reasoning

To understand God's thoughts we must study statistics for these are the measure of His purpose.

– Florence Nightingale.

Abstract: The last module is a collection of topics in applied probabilistic reasoning that were all too small to command their own separate modules. Topics include: 1) the randomized response method as a way of asking sensitive questions and hopefully receiving truthful answers; 2) the use of surrogate end points (or proxies) in the study of some phenomenon where the connections to "real" outcomes of interest (for example, to mortality) are indirect and probabilistically linked (for example, to lowered cholesterol levels); 3) the comparison between a normative theory of choice and decision making derived from probability theory and actual human performance; 4) permutation tests and statistical inference derived directly from how a randomized controlled study was conducted. As an oddity that can occur for this type of statistical inference procedure, the famous 1954 Salk polio vaccine trials are discussed. Also, three brief subsections are given that summarize the jackknife, the bootstrap, and permutation tests involving correlational measures. This latter material is provided in an abbreviated form suitable for slide presentation in class, and where further explanatory detail would be given by an instructor.

Contents

1 The Randomized Response Method

As noted elsewhere, how questions are framed and the context in which they are asked are crucial for understanding the meaning of the given responses. This is true both in matters of opinion polling and for collecting data on, say, the health practices of subjects. In these situations, the questions asked are usually not sensitive, and when framed correctly, honest answers are expected. For more sensitive questions about illegal behavior, (reprehensible) personal habits, suspect health-related behaviors, questionable attitudes, and so on, asking a question outright may not garner a truthful answer.

The randomized response method is one mechanism for obtaining "accurate" data for a sensitive matter at a group level (but not at the individual level). It was first proposed in 1965 by Stanley Warner in the *Journal of the American Statistical Association*, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias" (*60*, 63–69). A modified strategy was proposed by Bernard Greenberg and colleagues in 1969, again in *JASA*: "The Unrelated Question Randomized Response Model: Theoretical Framework" (*64*, 520–539). We first illustrate Warner's method and then Greenberg's with an example.

Let $\mathcal{Q}$ be the question: "Have you ever smoked pot (and inhaled)?"; and $\bar{\mathcal{Q}}$ the complement: "Have you never smoked pot (and inhaled)?" With some known probability, θ, the subject is asked $\mathcal{Q}$; and with probability $(1-\theta)$, is given $\bar{\mathcal{Q}}$ to answer. The respondent determines which question is posed by means of a probability mechanism under his or her control. For example, if the respondent rolls a single die and a 1 or 2 appears, question $\mathcal{Q}$ is given; if 3, 4, 5, or 6 occurs, $\bar{\mathcal{Q}}$ is given. So, in this case, $\theta = 1/3$.

As notation, let p be the proportion in the population for which the true response to $\mathcal{Q}$ is "yes"; $1-p$ is then the proportion giving a "yes" to $\bar{\mathcal{Q}}$. Letting P_{yes} denote the observed proportion of "yes" responses generally, its expected value is $\theta p + (1-\theta)(1-p)$; thus, p can be estimated as

$$\hat{p}_w = \frac{P_{yes} - (1-\theta)}{2\theta - 1} \ ,$$

where the subscript w is used to denote Warner's method of estimation. Obviously, θ cannot be 1/2 because the denominator would then be zero; but all other values are legitimate. The extremes of θ

being 0 or 1, however, do not insure the “privacy” of a subject’s response because the question actually answered would then be known.

The Greenberg method is referred to as the unrelated (or innocuous) question technique. The complement question $\bar{\mathcal{Q}}$ is replaced with an unrelated question, say, $\mathcal{Q}_U$, with a known probability of giving a “yes” response, say γ. For example, $\mathcal{Q}_U$ could be “Flip a coin. Did you get a head?” Here, $\gamma = 1/2$ for a “yes” response; the expected value of P_{yes} is $\theta p + (1-\theta)\gamma$, leading to

$$\hat{p}_g = \frac{P_{yes} - (1-\theta)\gamma}{\theta} \ ,$$

where the subscript g now refers to Greenberg’s method of estimation.

To decide which strategy might be the better, the variances of the two estimates can be compared though closed-form formulas:

$$\mathrm{Var}(\hat{p}_w) = \frac{p(1-p)}{n} + \frac{\theta(1-\theta)}{n(2\theta-1)^2} \ ;$$

$$\mathrm{Var}(\hat{p}_g) =$$

$$\frac{p(1-p)}{n} + \frac{(1-\theta)^2\gamma(1-\gamma) + \theta(1-\theta)(p(1-\gamma)+\gamma(1-p))}{n\theta^2} \ ,$$

where the number of respondents is denoted by n. As an example, suppose θ is .6; the coin flip defines $\mathcal{Q}_U$ so γ is .5; and let the true proportion p be .3. Using the variance formulas above: $\mathrm{Var}(\hat{p}_w) = 6.21/n$ and $\mathrm{Var}(\hat{p}_g) = .654/n$. Here, the Greenberg “innocuous question” variance is only about a tenth of that for the Warner estimate, making the Greenberg method much more efficient in this instance (that is, the sampling variance for the Greenberg estimate is much less than that for the Warner estimate).

The use of innocuous questions is the most common implementation of a randomized response method. This is likely due to the generally smaller variance for the Greenberg estimator compared to that for Warner; also, the possible confusion caused by using "ever" and "never" and responding "yes" and "no" in Warner's method is avoided by the use of an innocuous question. As a practical example of the unrelated question implementation of randomized response, several excerpts are presented below from a *New York Times* article by Tom Rohan (August 22, 2013), entitled "Antidoping Agency Delays Publication of Research":

> Doping experts have long known that drug tests catch only a tiny fraction of the athletes who use banned substances because athletes are constantly finding new drugs and techniques to evade detection. So in 2011, the World Anti-Doping Agency convened a team of researchers to try to determine more accurately how many athletes use performance-enhancing drugs.
>
> More than 2,000 track and field athletes participated in the study, and according to the findings, which were reviewed by *The New York Times*, an estimated 29 percent of the athletes at the 2011 world championships and 45 percent of the athletes at the 2011 Pan-Arab Games said in anonymous surveys that they had doped in the past year.
>
> ...
>
> The project began in 2011 when the researchers created a randomized-response survey, a common research technique that is used to ask sensitive questions while ensuring a subject's confidentiality. The researchers conducted their interviews at two major track and field events: the world championships in Daegu, South Korea, and the Pan-Arab Games in Doha, Qatar.
>
> Athletes at the events answered questions on tablet computers and were asked initially to think of a birthday, either their own or that of someone close

to them. Then, depending on the date of the birthday, they were instructed to answer one of two questions that appeared on the same screen: one asked if the birthday fell sometime between January and June, and the other asked, "Have you knowingly violated anti-doping regulations by using a prohibited substance or method in the past 12 months?"

The study was designed this way, the researchers said, so only the athlete knew which of the two questions he or she was answering. Then, using statistical analysis, the researchers could estimate how many of the athletes admitted to doping.

The researchers noted that not every athlete participated, and those who did could have lied on the questionnaire, or chosen to answer the birthday question. They concluded that their results, which found that nearly a third of the athletes at the world championships and nearly half at the Pan-Arab Games had doped in the past year, probably underestimated the reality.

2 Surrogate End Points and Proxies

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 probabilistic interpretations. An earlier module on the (mis)reporting of data was devoted more extensively to the review of Gigerenzer et al. (2007), where the distinctions are made between survival and mortality rates, absolute versus relative risks, natural frequencies versus probabilities, among others. Generally, the presentation of information should be as honest, clear, and transparent as possible. One such example given by Gigerenzer et al. (2007) suggests the use of frequency statements instead of single-event probabilities, thereby removing the ambiguity of the reference class: instead of saying "there is a 30% to 50% probability of developing sexual problems with Prozac," use "out of every 10 patients who take Prozac, 3 to 5 experience a sexual problem." Thus, a male taking Prozac won't expect that 30% to 50% of his personal sexual encounters will result in a "failure."

In presenting data to persuade, and because of the "lead-time bias" medical screening produces, it is ethically questionable to promote any kind of screening based on improved five-year survival rates, or to compare such survival rates across countries where screening practices vary. As a somewhat jaded view of our current health situation, we have physicians practicing defensive medicine because there are no legal consequences for overdiagnosis and overtreatment, but only for underdiagnosis. Or, as the editor of the *Lancet* commented (as quoted by R. Horton, *New York Review of Books*, March 11, 2004), "journals have devolved into information laundering operations for the pharmaceutical industry." The issues involved in medical screening and its associated consequences are psychologically important; for example, months after false positives for HIV, mammograms, or prostate cancer, considerable and possibly dysfunctional anxiety may still exist.

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 probabilistically determined and 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.

Outside of the medical arena, proxies play prominently in the current climate-change debate. When actual surface temperatures are unavailable, surrogates for these are typically used (for example, tree-ring growth, coral accumulation, evidence in ice). Whether these are satisfactory stand-ins for the actual surface temperatures is questionable. Before automatically accepting a causal statement (for example, that greenhouse gases are wholly responsible for the apparent recent increase in earth temperature), pointed (statistical) questions should be raised, such as:

(a) why don't the tree-ring proxies show the effects of certain climate periods in our history—the Medieval Warm Period (circa 1200) and the Little Ice Age (circa 1600)?;

(b) over the last century or so, why has the tree-ring and surface temperature relationship been corrupted so that various graphical "tricks" need to be used to obtain the "hockey stick" graphic demonstrating the apparent catastrophic increase in earth temperature over the last century?;

(c) what effect do the various solar cycles that the sun goes through have on our climate; could these be an alternative mechanism for what we are seeing in climate change?;

(d) or, is it some random process and we are on the up-turn of something comparable to the Medieval Warm Period, with some later downturn expected into another Little Ice Age?

3 The Normative Theory of Probability and Human Decision Making

One important area of interest in developing statistical literacy skills and learning to reason probabilistically 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. We give a classic Tversky and Kahneman (1983) illustration to show how various reasoning heuristics might operate:

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination

> and social justice, and also participated in anti-nuclear demonstrations. Which . . . [is] more probable?
>
> 1. Linda is a bank teller.
> 2. Linda is a bank teller and is active in the feminist movement.

Eighty-five percent of one group of subjects chose option 2, even though the conjunction of two events must be less likely than either of the constituent events. Tversky and Kahneman argue that this "conjunction fallacy" occurs because the "representativeness heuristic" is being used to make the judgment; the second option seems more representative of Linda based on the description given for her.

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 (base rates) 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. (p. 1128)

Because retrievability can be influenced by differential familiarity and saliences, the probability of an event may not be best estimated by the ease to which occurrences come to mind. A third reasoning heuristic is one of *anchoring and adjustment*, which may also be prone to various biasing effects. Here, estimates are made based on some initial value that is then adjusted (Tversky & Kahneman, 1974).

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. The classic experiment on the fundamental attribution error is from Jones and Harris (1967); we quote a summary given in the Wikipedia article on the fundamental attribution error:

> Subjects read pro- and anti-Fidel Castro essays. Subjects were asked to rate the pro-Castro attitudes of the writers. When the subjects believed that the writers freely chose the positions they took (for or against Castro), they naturally rated the people who spoke in favor of Castro as having a more positive attitude toward Castro. However, contradicting Jones and Harris' initial hypothesis, when the subjects were told that the writer's positions were determined by a coin toss, they still rated writers who spoke in favor of Castro as having, on average, a more positive attitude towards Castro than those who spoke against him. In other words, the subjects were unable to see the influence of the situational constraints placed upon the writers; they could not refrain from attributing sincere belief to the writers.

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. We give an article that lays out the issues by Michael Stone of the *Portland Progressive Examiner* (March 5, 2010; "Toxic Politics: Liz Cheney's Keep America Safe 'Al Qaeda Seven' Ad"):

> Liz Cheney, daughter of former Vice President Dick Cheney and co-founder of the advocacy group "Keep America Safe," is taking heat for a controversial ad questioning the values of Justice Department lawyers who represented Guantanamo Bay detainees.
>
> Several top political appointees at the Justice Department previously worked as lawyers or advocates for 'enemy combatants' confined at Guantanamo Bay, Cuba. In their ad, Cheney's group derides the unidentified appointees as the 'Al Qaeda 7.' The ad implies the appointees share terrorist values.

Aside from questioning the values of these Justice Department lawyers, the ad is using fear and insinuations to smear both the Justice Department lawyers and the Obama administration.

Demonizing Department of Justice attorneys as terrorist sympathizers for their past legal work defending Gitmo detainees is wrong. The unfounded attacks are vicious, and reminiscent of McCarthyism.

Indeed, the ad itself puts into question Cheney's values, her patriotism, her loyalty. One thing is certain: her understanding of US history, the founding of our country, and the US Constitution, is left seriously wanting.

John Aloysius Farrell, writing in the Thomas Jefferson Street blog, for *US News and World Report*, explains:

There are reasons why the founding fathers ... in the Bill of Rights, strove to protect the rights of citizens arrested and put on trial by the government in amendments number 4, 5, 6, 7, and 8.

The founders had just fought a long and bloody revolution against King George, and knew well how tyrants like the British sovereign perpetuated power with arbitrary arrests, imprisonments, and executions. And so, along with guarantees like the right to due process, and protection from unreasonable searches and cruel and unusual punishment, the first patriots also included, in the Sixth Amendment, the right of an American to a speedy trial, by an impartial jury, with "the Assistance of Counsel for his defense."

John Adams regarded his defense of the British soldiers responsible for the Boston Massacre as one of the noblest acts of his life for good reason. Our adversarial system of justice depends upon suspects receiving a vigorous defense. That means all suspects must receive adequate legal counsel, including those accused of the most heinous crimes: murder, rape, child abuse and yes, even terrorism.

Defending a terrorist in court does not mean that one is a terrorist or shares terrorist values. Implying otherwise is despicable. Cheney's attacks are a dangerous politicization and polarization of the terrorism issue. Those who would honor our system of law and justice by defending suspected terrorists deserve our respect. Instead Cheney and her group smear these patriots in an attempt to score points against political enemies.

4 Permutation Tests and Statistical Inference

The aim of any well-designed experimental study is to make a causal claim, such as "the difference observed between two groups is *caused* by the different treatments administered." To make such a claim we need to know the counterfactual: what would have happened if this group had not received the treatment? This counterfactual is answered most credibly when subjects are assigned to the treatment and control groups at random. In this instance, there is no reason to believe that the group receiving the treatment condition would have reacted any differently (than the control condition) had it received the control condition. If there is no differential experimental mortality to obscure this initial randomness, one can even justify the analyses used by how the groups were formed (for example, by randomization tests, or their approximations defined by the usual analysis methods based on normal theory assumptions). As noted by R. A. Fisher (1971, p. 34), "the actual and physical conduct of an experiment must govern the statistical procedure of its interpretation." When the gold standard of inferring causality is not met, however, we are in the realm of quasi-experimentation, where causality must be approached differently.

An important benefit from designing an experiment with random assignment of subjects to conditions, possibly with blocking in various ways, is that the method of analysis through randomization tests is automatically provided. As might be expected, the original philosophy behind this approach is due to R. A. Fisher, but it also has been developed and generalized extensively by others (see Edgington & Onghena, 2007). In Fisher's time, and although randomization methods may have been the preferred strategy, approximations were developed based on the usual normal theory assumptions to serve as computationally feasible alternatives. But with this view, our standard methods are just approximations to what the preferred analyses should be. A short quotation from Fisher's *The Design of Experiments* (1971) makes this point well (and one that expands on the short phrase given in the previous paragraph):

> In these discussions it seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied. The arithmetical procedure of such an examination is tedious, and we shall only give the results of its application ... to show the possibility of an independent check on the more expeditious methods in common use. (p. 45)

A randomization (or permutation) test uses the given data to generate an exact null distribution for a chosen test statistic. The observed test statistic for the way the data actually arose is compared to this null distribution to obtain a p-value, defined as the probability (if the null distribution were true) of an observed test statistic being as or more extreme than what it actually was. Three situations lead to the most common randomization tests: K-dependent samples, K-independent samples, and correlation. When ranks are used instead of the original data, all of the common nonparametric tests arise. In practice, null randomization distributions are obtained either by complete enumeration, sampling (a Monte Carlo strategy), or through various kinds of large sample approximations (for example, normal or chi-squared distributions).

Permutation tests can be generalized beyond the usual correlational framework or that of K-dependent or K-independent samples. Much of this work falls under a rubric of combinatorial data analysis (CDA), where the concerns are generally with comparing various kinds of complete matrices (such as proximity or data matrices) using a variety of test statistics. The most comprehensive source for this material is Hubert (1987), but the basic matrix comparison strategies are available in a number of places, for example, see discussions of the "Mantel Test" in many packages in R (as one example, see the "Mantel–Hubert general spatial cross-product statistic" in the package, `spdep`). Even more generally, one can at times tailor a test statistic in nonstandard situations and then implement a permutation strategy for its evaluation through the principles developed in CDA.

The idea of repeatedly using the sample itself to evaluate a hypothesis or to generate an estimate of the precision of a statistic, can be placed within the broader category of resampling statistics or sample reuse. Such methods include the bootstrap, jackknife, randomization and permutation tests, and exact tests (for example, Fisher's exact test for 2 × 2 contingency tables). Given the incorporation of these techniques into conveniently available software, such as R, there are now many options for gauging the stability of the results of one's data analysis.

4.1 The Jackknife

An idea similar to the "hold-out-some(one)-at-a-time" is Tukey's Jackknife.

This was devised by Tukey to obtain a confidence interval on a parameter (and indirectly to reduce the bias of an estimator that is not already unbiased).

In Psychology, there is an early discussion of the Jackknife in the *Handbook of Social Psychology (Volume II)* (Lindzey and Aronson; 1968) by Mosteller and Tukey: Data Analysis — Including Statistics.

General approach for the Jackknife:

suppose I have n observations $X_1, \ldots, X_n$ and let θ be an unknown parameter of the population.

We have a way of estimating θ (by, say, $\hat{\theta}$) –

Group the n observations into t groups of m; thus, $n = tm$:

$\{X_1, \ldots, X_m\}, \ldots, \{X_{(t-1)m+1}, \ldots, X_{tm}\}$

Let $\hat{\theta}_{-0}$ be the estimate based on all groups;

Let $\hat{\theta}_{-i}$ be the estimate based on all groups except the i^{th}

Define new estimates of θ, called "pseudo-values" as follows:

$\hat{\theta}_{*i} = t\hat{\theta}_{-0} - (t-1)\hat{\theta}_{-i}$, for $i = 1, \ldots, t$

The Jackknife estimate of θ is the mean of the pseudo-values:

$\hat{\theta}_{*\cdot} = \sum_{i=1}^{t} \frac{\hat{\theta}_{*i}}{t}$

An estimate of its standard error is

$s_{\hat{\theta}_{*\cdot}} = [\sum_{i=1}^{t} \frac{(\hat{\theta}_{*i} - \hat{\theta}_{*\cdot})^2}{t(t-1)}]^{1/2}$

Approximate confidence interval:

$\hat{\theta}_{*\cdot} \pm s_{\hat{\theta}_{*\cdot}} t_{\frac{\alpha}{2},t-1}$

We act as if the t pseudo-values $\hat{\theta}_{*1}, \ldots, \hat{\theta}_{*t}$ are independent and identically distributed observations.

We also reduce some bias in estimation if the original estimate was biased.

An example:

suppose I want to estimate μ based on $X_1, \ldots, X_n$

Choose $t = n$

$\hat{\theta}_{-0} = \frac{1}{n} \sum_{j=1}^{n} X_j$

$\hat{\theta}_{-i} = \frac{1}{n-1} \sum_{j=1, i \neq j}^{n} X_j$

$\hat{\theta}_{*i} = n(\frac{1}{n} \sum_{j=1}^{n} X_j) - (n-1)(\frac{1}{n-1} \sum_{j=1, i \neq j}^{n} X_j) = X_i$

Thus, $\hat{\theta}_{*\cdot} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$

$s_{\hat{\theta}_{*\cdot}} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2} =$

$\sqrt{s_X^2/n}$, where s_X^2 is an unbiased estimate of σ^2

Confidence interval:

$\bar{X} \pm (\sqrt{s_X^2/n})\, t_{\frac{\alpha}{2},t-1}$

4.2 The Bootstrap

Population (“Theory World”): the pair of random variables X and Y are, say, bivariate normal

Sample (“Data World”): n pairs of independent and identically distributed observations on (X, Y):

$(X_1, Y_1), \ldots, (X_n, Y_n)$; these could be used to give r_{XY} as an estimate of ρ_{XY}

Now, make Data World the Theory World Population:

$(X_1, Y_1), \ldots, (X_n, Y_n)$, and each occurs with probability $\frac{1}{n}$

Sample this Theory World Population (with replacement) to get one “bootstrap” sample (with possible repeats):

$(X_1^{'}, Y_1^{'}), \ldots, (X_{n'}^{'}, Y_{n'}^{'})$ (usually, n equals $n^{'}$)

Get B bootstrap samples and compute the correlation for each: $r_{XY}^{(1)}, \ldots, r_{XY}^{(B)}$

This last distribution could be used, for example, to obtain a confidence interval on ρ_{XY}

4.2.1 Permutation tests for correlation measures

We start at the same place as for the Bootstrap:

Population (“Theory World”): the pair of random variables X and Y are, say, bivariate normal

Sample (“Data World”): n pairs of independent and identically distributed observations on (X, Y):

$(X_1, Y_1), \ldots, (X_n, Y_n)$; these could be used to give r_{XY} as an estimate of ρ_{XY}

Now, to test H_o : X and Y are statistically independent.

Under H_o, the X's and Y's are matched at random; so, assuming (without loss of generality) that we fix the X's, all $n!$ permutations of the Y's against the X's are equally likely to occur.

We can calculate a correlation for each of these $n!$ permutations and graph:

the distribution is symmetric and unimodal at zero; the range along the horizontal axis obviously goes from -1 to $+1$

p-value (one-tailed) = number of correlations as or larger than the observed correlation$/n!$

Also, as an approximation, $r_{XY} \sim N(0, \frac{1}{n-1})$;

Thus, the standard error is close to $\frac{1}{\sqrt{n}}$; this might be useful for quick "back-of-the-envelope" calculations

4.3 An Introductory Oddity: The 1954 Salk Polio Vaccine Trials

The 1954 Salk polio vaccine trials was the biggest public health experiment ever conducted. One field trial, labeled an observed control experiment, was carried out by the National Foundation for Infantile Paralysis. It involved the vaccination, with parental consent, of second graders at selected schools in selected parts of the country. A control group would be the first and third graders at these same schools, and indirectly those second graders for whom parental consent was not obtained. The rates for polio contraction (per 100,000) are given below for the three groups (see Francis et al., 1955, for the definitive report on the Salk vaccine trials).[1]

Grade 2 (Vaccine): 25/100,000;
Grade 2 (No consent): 44/100,000;
Grades 1 and 3 (Controls): 54/100,000.

The interesting observation we will return to below is that the Grade 2 (No consent) group is between the other two in the probability of polio contraction. Counterintuitively, the refusal to give consent seems to be partially protective.

The second field trial was a (double-blind) randomized controlled experiment. A sample of children were chosen, all of whose parents consented to vaccination. The sample was randomly divided into two, with half receiving the Salk vaccine and the other half a placebo of inert salt water. There is a third group formed from those children with no parental consent and who therefore were not vaccinated. We give the rates of polio contraction (per 100,000) for the three groups:

Vaccinated: 28/100,000;
Control: 71/100,000;
No consent: 46/100,000.

Again, not giving consent appears to confer some type of immunity;

[1] The interpretation of results and the source of the information given in this section, *An Evaluation of the 1954 Poliomyelitis Vaccine Trials*, is by Thomas Francis, Robert Korns, and colleagues (1955) (in particular, see Table 2b: Summary of Study of Cases by Diagnostic Class and Vaccination Status; p. 35).

the probability for contracting polio for the “no consent” group is between the other two.

The seeming oddity in the ordering of probabilities, where “no consent” seems to confer some advantage, is commonly explained by two “facts”: (a) children from higher-income families are more vulnerable to polio; children raised in less hygienic surroundings tend to contract mild polio and immunity early in childhood while still under protection from their mother’s antibodies; (b) parental consent to vaccination appears to increase as a function of education and income, where the better-off parents are much more likely to give consent. The “no consent” groups appear to have more natural immunity to polio than children from the better-off families. This may be one of the only situations we know of where children growing up in more resource-constrained contexts are conferred some type of advantage.

References

[1] Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). New York: Chapman & Hall / CRC.

[2] Fisher, R. A. (1971). *The design of experiments* (9th ed.). New York: Hafner.

[3] Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

[4] Hubert, L. J. (1987). *Assignment methods in combinatorial data analysis*. New York: Marcel Dekker.

[5] Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*, 1–24.

[6] Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

[7] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

[8] Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

[9] Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.