Statistical Science
2001, Vol. 16, No. 1, 35–57

Statistical and Legal Aspects of the Forensic Study of Illicit Drugs

Alan Julian Izenman

Abstract. Prosecuting those arrested for the unlawful possession, distribution or importation of illicit drugs, such as powder cocaine, crack cocaine, heroin and LSD, is usually both time-consuming and expensive, primarily because of the need to determine "beyond a reasonable doubt" the total amount of drugs seized in each case. Accuracy is important since penalties (imprisonment or fines) depend upon the quantity seized. Substantial backlogs in processing drug cases often develop as a result. In some jurisdictions, complete testing and analysis of all substances seized from a defendant are customary in support of a case, while in other jurisdictions random sampling of drugs and the subsequent presentation of an estimate of the total amount seized have been used for many years. Due to pressure from crime laboratories and prosecutors, who point to major increases in their caseloads as well as a trend toward decreasing funding and staffing for the crime laboratories, jurisdictions which currently carry out a complete census of all seized evidence are now seriously considering a change in their methodology with a view to instituting new guidelines for the scientific sampling of evidence. In this article, we discuss the statistical and legal issues that have arisen in cases involving illicit drugs.

Key words and phrases: Composite sampling, controlled substances, federal sentencing guidelines, forensic statistics, homogeneity, multistage sampling, random sampling, sample size, standards of proof, statistics and the law.

1. INTRODUCTION

The quantity of drugs may reflect significant facts about the defendant. It may demonstrate an individual's high level of trust within the drug distribution network. It may reflect how long he has been involved. It may correlate with the amount of money the offender may earn from the crime. A prosecutor's traditional allocution at the sentencing of a drug distributor with a large quantity of drugs who had no prior record is,

> "He may not have any prior arrests your honor, but the quantity of drugs alone suggests that he is no novice to drug dealing" [*From Young*, 1990, page 63].

Many books (e.g., DeGroot, Fienberg and Kadane, 1986; Kaye and Aickin, 1986; Gastwirth, 1988; Fienberg, 1989; Finkelstein and Levin, 1990; Aitken, 1995; Zeisel and Kaye, 1997; Gastwirth, 2000) and articles have appeared in recent years dealing with the impact of statistical arguments on the judicial process. In addition, a very useful collection of essays on scientific evidence in the courts can be found in the reference manual for judges produced by the Federal Judicial Center (1994). These books and articles show that statistical arguments are becoming increasingly relevant in employment discrimination, antitrust and environmental law cases, and that statisticians are participating in the legal process with increasing frequency.

The present article explores a different part of the statistics–law interface, specifically, statistical

Alan Julian Izenman is Professor of Statistics and Senior Research Fellow, Fox School of Business and Management, Department of Statistics, Temple University, Speakman Hall, Philadelphia, Pennsylvania 19122 (e-mail: alan@sbm.temple.edu).

issues that arise in sentencing those convicted of drug trafficking (i.e., the unlawful possession, distribution or importation of illicit drugs).

1.1 Background

Although a national "drug problem" had been present in the United States for some time, it mushroomed following the emergence of crack cocaine in 1981. By 1985 the cocaine problem had became widespread in East and West Coast cities. In particular, the use of cocaine changed from primarily casual use to addictive use and was aggressively marketed by drug traffickers to disadvantaged residents of the inner cities. This development led to the widespread proliferation of street sellers who distributed illicit drugs under the direction of professional organizations. To combat this turn of events, a massive expansion in drug enforcement took place and led to record numbers of defendants being arrested and convicted of drug offences. As one would expect, the resulting drug caseloads became an enormous burden to prosecutors (Boland and Healy, 1993).

In 1984, Congress enacted the Sentencing Reform Act as part of the Comprehensive Crime Control Act. The goals of the Sentencing Reform Act were "to reduce unwarranted disparity, increase certainty and uniformity, and correct past patterns of undue leniency for certain categories of serious offenses" (United States Sentencing Commission, 1991). In 1985, the United States Sentencing Commission was appointed as an independent, permanent agency within the judiciary to achieve these goals. Guidelines for determining appropriate types and lengths of sentences for every federal offense were first submitted to Congress in April 1987 and became law in November 1987. Following delays due to constitutional challenges, the federal sentencing guidelines were implemented nationally in January 1989. The sentencing guidelines are considered evolutionary, and modifications and revisions may be submitted each year to Congress.

Another part of the 1984 Comprehensive Crime Control Act was the Controlled Substances Penalties Amendments Act, which for the first time made punishment dependent upon the amount of illicit drugs seized. Congress also enacted several statutes imposing mandatory minimum sentences, especially for cases involving possession of significant amounts of illicit drugs. In particular, the 1986 Anti–Drug Abuse Act set up a new system of nonparolable, mandatory minimum sentences for drug trafficking where the minimum penalty was tied to the quantity of drugs involved in the offense. Later versions of this legislation extended mandatory minimum penalties to different aspects of drug violations. Relationships between the sentencing guidelines and mandatory minimum penalties are discussed in great detail in the Special Report to Congress by the United States Sentencing Commission (1991).

The last few years have seen a concern on the part of federal and state law enforcement agencies, crime laboratories and prosecutors regarding the intensive effort needed to provide detailed testimony in drug cases. Procedures that had been used for decades are now being questioned as to their appropriateness and efficiency. Caseloads are increasing at the same time as decreases occur in funding and staffing of crime laboratories. This point was made quite forcefully by Beaupre and Eisler (1996) in a newspaper report on the present state of crime laboratories. The escalation in numbers of arrests of those suspected of trafficking (possession with intent to distribute) in illicit drugs has prompted a reevaluation of policies for determining drug amounts found in possession of defendants. The quantity of illicit drugs is often the principal determinant of the sentence the defendant will receive upon conviction. In many jurisdictions, the greater the quantity of drugs seized the more severe the sentence.

Rather than require a crime laboratory to analyze every gram of suspected drugs to determine the total quantity of illicit drugs seized from a defendant, many jurisdictions will now accept scientific sampling of the evidence and, with it, the use of statistical estimates of the total quantity of drugs possessed by the defendant. The practice of sampling has long been accepted by the courts in a wide variety of criminal and civil cases. Larsen (1964) annotates over a hundred civil cases where "samples" were submitted in evidence of actions for recovery of damages by breaches of contract, warranty or covenant and by negligence and personal injury or death. More recently, Walker and Monahan (1999) studied the use of sampling to prove causation in lawsuits brought by states seeking reimbursement from tobacco companies for Medicaid payments attributable to tobacco-related diseases. The results of specially commissioned surveys and opinion polls have also been used to help decide damages in trademark infringement cases, antitrust cases, accounting cases and motions for a change in venue in criminal cases. See Diamond (1994), Strong (1992, Section 208) and Walker and Monahan (1998). For an excellent discussion of samples used in evidence, see Gastwirth (1988, Chapter 9).

All federal district and appellate courts have accepted sampling in drug cases, but state courts

may differ in their sampling policies, even within the same state. Although New York State allows sampling in drug cases, the crime laboratories around the state (with the notable exception of New York City) currently engage in 100% inspection and testing of the evidence because caseloads statewide are not very large. The New York City Police Department, however, which has a huge narcotics caseload (138,199 cases received in 1996, of which 66,927 were analyzed), has been consulting with statisticians since approximately 1996 to develop formal sampling plans.

1.2 Outline of This Article

In Section 2, we outline the legal standards used in cases of drug violations, including drug statutes, sentencing guidelines and the different standards of proof required for sentencing. In Section 3, we describe and comment upon the course of typical testimony by an expert witness regarding the laboratory practice of analyzing drug evidence to estimate Q, the total quantity (or weight) of the drugs seized from a defendant. We discuss scientific sampling and chemical testing of such evidence and the problem of determining the extent of homogeneity of substances found in multiple containers. In Section 4, we consider the various statistical issues that tend to arise in drug cases. These include how large a sample to choose when an unknown portion of the evidence may not be drugs, and we trace the origins of the popular, but theoretically unjustified "square-root rule." We describe a composite sampling method for estimating drug purity and then use it for assessing homogeneity. We also indicate how such results may be translated into probabilities for standards of proof in sentencing decisions. Concluding thoughts are given in Section 5.

2. LEGAL STANDARDS IN ILLICIT DRUG CASES

2.1 Drug Statutes and Sentencing

2.1.1 *Federal cases.* Federal statutes and associated penalties for drug violations are detailed in Title 21 of the United States Code (U.S.C.), Section 841 *et seq.*, popularly known as the Controlled Substances Act. According to these statutes, controlled substances are divided into five different "schedules," according to their potential for abuse. The two highest schedules are Schedule I for those drugs (or other substances), such as lysergic acid diethylamide (LSD), heroin and marijuana, that have high potential for abuse, that have no currently accepted medical use in treatment in the United States and for which there is a lack of an accepted safe dose for use of the drug under medical supervision, and Schedule II for those drugs, such as powder and crack cocaine, that have high potential for abuse, that have a currently accepted medical use in treatment in the United States or a currently accepted medical use with severe restrictions and for which abuse of the drug may lead to severe psychological or physical dependence. The Controlled Substances Act has two separate subsections that define (a) the "prohibited acts" of trafficking in illicit drugs and (b) the resulting "penalties." Until recently, subsection (b) was considered to be *independent* of subsection (a) and was regarded as relevant only after a defendant had been convicted of a crime under subsection (a). This interpretation has since been rejected by the Supreme Court (*Apprendi v. New Jersey*, 2000).

The present federal mandatory minimum penalties for first-time drug violations in the cases of cocaine, heroin and LSD range from 5 to 20 years. Such mandatory minimum sentences are doubled for second-time offenders. For third-time offenders, the penalty is life. Federal criminal sentences are also circumscribed by the United States Sentencing Guidelines (U.S.S.G. or "Federal Sentencing Guidelines"); see United States Sentencing Commission (1995, Section 2D1.1(c)). The guidelines were developed using "statistical analyses" of about 40,000 convictions and 10,000 augmented presentence reports, with a goal of tying the sentencing ranges to actual past practice. All federal crimes are governed by the Federal Sentencing Guidelines and federal judges are required to follow these guidelines. Sentencing decisions have to be documented in great detail and may be subject to rigorous appellate review. The United States Sentencing Commission (1995, Section 5G1.1), however, points out that "a maximum sentence set by statute trumps a higher sentence set forth in the Guidelines."

Sentencing for federal crimes is determined by two factors, which are arranged as the rows and columns of a two-way, 43-row by 6-column, "sentencing table" (U.S.S.G., Section 5A). The rows express the magnitude of the crime in terms of an "offense level," which is an adjusted version of a "base offense level," or BOL, numbered from 1 to 43. In general, the more serious the crime, the higher the BOL and the longer the sentence. See Izenman (2000a) for descriptions of how the BOL is determined in cases involving fraud and trafficking in cyberporn. In illicit drug cases, the BOL is determined primarily from a "drug quantity table" (U.S.S.G., Section 2D1.1(c)) by the total weight Q, say, of all illicit drugs seized from the

defendant. Upward or downward departures from a defendant's BOL can be made, depending upon the circumstances of the crime and any "relevant conduct" to arrive at the "offense level" for that crime. The columns of the sentencing table represent the defendant's history of recidivism, which the Supreme Court has described as "a traditional, if not the most traditional, basis for a sentencing court's increasing an offender's sentence" (*Almendarez-Torres v. United States*, 1998). Prior criminal history is listed as ordered categories ranging from I (no prior criminal history) to VI. Each cell in the table then gives a sentencing range from which the judge has to pick a sentence. For example, a conviction on possession of 50–150 grams of crack cocaine carries a BOL of 32; if the defendant has no criminal history (category I) and there are no adjustments to the BOL, the sentencing table gives a sentence of 121–151 months in prison.

Congress has decreed (18 U.S.C., Section 841 (1998)) that inactive ingredients, such as the dilutant, cutting agent or carrier medium, that are combined with heroin and cocaine also be included in the calculation of Q for sentencing purposes. The Federal Sentencing Guidelines did likewise (U.S.S.G., Section 2D1.1(c), Note (A)) in their BOL calculations of drug quantity. In the case of LSD, the Supreme Court in *Chapman v. United States* (1991) (and reaffirmed in *Neal v. United States*, 1996) held that if the carrier medium, such as blotter paper, cannot easily be distinguished and separated from the LSD, then it should be included when determining the weight of LSD. For a recent article on the effect of *Chapman*, see Gonyer (1998).

Often, Q falls very slightly over some threshold amount in the Drug Quantity Table, and the defendant receives a higher sentence than he or she feels is justified. In *United States v. Chapple* (1993), for example, the total weight of cocaine found in five packages was 5,003 grams, which was "just over the crucial five-kilogram line." At issue in the appeal was that under the Federal Sentencing Guidelines a total weight of just under five kilograms of cocaine (a BOL of 30) would mean a sentence of 120–151 months rather than the 170 months the defendant received for just over five kilograms of cocaine (a BOL of 32). In such cases, the original sentence is usually affirmed.

2.1.2 *State cases.* Individual states have their own statutes and penalties for illegal drug possession and trafficking. According to a recent survey by the Bureau of Justice Assistance (1996), 16 states had implemented, or were about to implement, sentencing guidelines similar to those of the federal government, and a further 5 states have guidelines under study. Most states, however, have not yet adopted sentencing guidelines, and several others have tried but failed to implement such guidelines. All states make use of some version of mandatory minimum sentencing laws, especially for drug trafficking. Whether following guidelines or not, penalties of different states can vary dramatically and depend also on the year in which they were enacted; typically, they will differ from the federal penalties.

2.2 Standard of Proof Required

Because of the obvious need to persuade a court and the subsequent impact on sentencing in drug cases, until recently, different standards of proof were used by state and federal courts in assessing the legal weight to give the value of Q provided by the government. The two primary standards of proof used by the courts in drug cases have been "beyond a reasonable doubt," which is constitutionally mandated for elements of a criminal offense, and "preponderance of the evidence" (also sometimes called "the balance of probabilities"), which is the general rule in civil suits. There is another standard of proof, namely, "clear and convincing evidence," which falls somewhere between the preponderance and reasonable doubt standards.

In civil cases, where the preponderance standard is the rule, it is not unusual to see judges quantifying its meaning without it being regarded as improper. In Pennsylvania, for example, the standard civil jury instructions (Pennsylvania Bar Institute, 1981, Section 5.50) describe the preponderance standard in terms of an ordinary balance scale, with a pan on each side, where evidence favorable to the plaintiff is placed upon one pan and evidence favorable to the defendant upon the other, and if, in considering the comparable weight of the evidence, the scales tip, "ever so slightly or to the slightest degree," in favor of the plaintiff (defendant), the verdict must be for the plaintiff (defendant). Model federal jury instructions in civil cases (Devitt, Blackmar, and Wolff, 1987, Section 72.01) describe the preponderance standard as trying to prove that something is "more likely so than not so." Walker (1994) noted that "[a] number of [federal and state] courts have ... interpreted or explained the meaning of preponderance ... using the quantitative terminology of mathematical probability. These courts have held that a 'preponderance' of evidence means having a probability of truth greater than 0.5, or having a 'greater than 50% chance' of being true."

In *United States v. Fatico* (1978), Judge Jack Weinstein took the opportunity to survey 10 district judges from the Eastern District of New York as to their probabilistic interpretations of the various standards of proof. Only the preponderance standard yielded consistent responses and was associated by all of them with a probability of just over 50% (with a standard deviation of only about 0.4%). Their interpretations of the clear and convincing standard (averaging about 67%) and the reasonable doubt standard (averaging about 86%) were much more variable (each having a standard deviation of about 6%) and were generally below the probabilities suggested by Weinstein himself. See Fienberg (1989, pages. 201–204) Gastwirth (1988, pages 700–702) and Gastwirth (1992) for further discussion.

In state courts, the value of Q in a drug trial is defined to be an essential element of the possession charge, especially when there is a lesser included offense for possessing a smaller amount. As such, the value of Q must be proved by a jury beyond a reasonable doubt.

In federal court, sentencing in drug cases has been quite different. Until very recently, every federal appeals court had held that factual determination of Q was *not* included under 21 U.S.C. Section 841 as an essential element of the offense which had to be proved beyond a reasonable doubt. If the defendant either pleaded guilty or was found guilty beyond a reasonable doubt after a federal trial, Q became an issue only at sentencing. In many jurisdictions, the value of Q was often not even mentioned in the indictment. At the time of sentencing, a probation officer would usually prepare a presentence report (including a determination of Q) on the defendant to be used by the court. Disputed facts at sentencing would then be established by the sentencing judge using the preponderance standard; judges would be allowed to consider even noncharged criminal activity (termed "relevant conduct") in computing a defendant's sentence (Greenwald, 1994). For sentencing purposes in all federal courts, the value of Q was an issue of fact to be established only by a preponderence of the evidence.

In June 2000, the Supreme Court decided *Apprendi v. New Jersey* (2000), with an opinion that dissenter Justice O'Connor termed "a watershed change in constitutional law." Although Apprendi had pleaded guilty to second-degree possession of a firearm for an unlawful purpose, after conviction the prosecutor filed a motion to enhance the sentence on the grounds that the crime was racially motivated. The sentencing judge, by a preponderance of the evidence, found the defendant's crime to be a hate crime and enhanced the sentence to a period longer than the maximum the defendant could have received based upon the jury's verdict. Using recent decisions, the Supreme Court reversed, and announced a constitutionally based rule:

> Other than the fact of a prior conviction, any fact that increases the penalty for a crime beyond the prescribed statutory maximum must be submitted to a jury, and proved beyond a reasonable doubt.

Most legal commentators (e.g., Liman, 2000; Mauro and Ringel, 2000) have inferred that this ruling will have a far-reaching effect on all criminal prosecutions, including drug cases charged under 21 U.S.C. Section 841. The 8th Circuit, the first federal appellate court to address the issue (*United States v. Sheppard*, 2000; *United States v. Aguayo-Delgado*, 2000), has agreed. It is not yet clear how other federal Circuits (and the Sentencing Commission) will interpret *Apprendi* or what its ultimate import will be. One result may be, however, that prosecutors choose to charge a range of possible values for Q in the indictment rather than a single value. It is likely that federal prosecutors will assume that drug quantity Q is an element of the crime, that it must be charged in the indictment and that it must be proved beyond a reasonable doubt to a jury. The *Apprendi* case has dramatically changed the landscape of federal drug prosecutions. Its impact on states' prosecutions will be minimal because most states have been implementing the new Supreme Court rule for some time.

3. COMMON PRACTICES IN THE COURTROOM AND LABORATORY

3.1 Total Quantity Q of Illicit Drugs

The typical scenario is that a defendant will be found in possession of $K \geq 1$ containers (e.g., plastic bags, ziplock "baggies," glassines, paper packets or envelopes, vials or even bottles) suspected of containing illicit drugs. In some cases, these K containers may further be packaged into, say, $B \geq 1$ batches, where the bth batch consists of K_b containers and $K = \sum_{b=1}^{B} K_b$. For example, in *People v. Argro* (1975), $B = 10$ and $K = 148$, and in *State of North Carolina v. Hayes* (1976), $B = 2$ and $K = 19$. For convenience in this exposition, we shall henceforth assume that $B = 1$.

Suppose that the weight of the tth item (e.g., pill, tablet, capsule, rock of crack cocaine) in the jth container is recorded as X_{jt} grams if the item tests positive; if that item tests negative, the

weight is recorded as $X_{jt} = 0,\ j = 1, 2, \ldots, K, t = 1, 2, \ldots, N_j$. (If the jth container holds loose heroin or cocaine powder, then $N_j = 1$.) The total weight of the drugs seized is

$$Q = \sum_{j=1}^{K} \sum_{t=1}^{N_j} X_{jt}. \tag{3.1}$$

The calculation of Q became necessary when differential penalties depending upon the amount seized were first spelled out by the 1988 Amendment to the Controlled Substances Act. If Q falls into a specific range or interval given in the statutes, then the penalty is determined as so many years of imprisonment. [In the special case of perforated sheets of 100 "one-dose" squares of a carrier medium (usually blotter paper) impregnated with LSD for oral ingestion, the problem is to determine how many of those squares contain LSD, a number which in some jurisdictions is then converted into a total weight measurement Q.]

Current practice in New York State is to examine in its entirety every substance seized from a defendant. To comprehend the magnitude of the work carried out, we note that in 1996 the New York City crime laboratory analyzed the contents of 792,305 vials, bags, packets etc. This can be a lengthy process, very labor intensive and costly to perform. Indeed, it is not uncommon in a single seizure for a police department crime laboratory to test the contents of several thousand bags, with a chemist taking two or three weeks to carry out the required tests. In *People v. Peneda* (1995), for example, a criminalist for the Orange County Sheriff's crime laboratory testified that, in a previous case, he had tested and analyzed a full 100 pounds of cocaine to testify that there was actually 100 pounds of the substance; the testing had taken him "two weeks of constant work" before he reached the 100-pound limit. In such cases, court delays are common while examinations are conducted. Even so, there is no guarantee that a 100% inspection (or "census") will yield exact results, especially when the total number of items $N = \sum_j N_j$ is large and errors in weighing and counting are possible.

Furthermore, because chemical testing destroys evidence in drug cases, 100% inspection and testing would, therefore, destroy all the evidence. This would limit the ability of prosecutors to display the substances in court and, in the interests of fairness, for the defense to challenge the prosecution's findings as to the value of Q. It is also well known (Le, Taylor, Vidal, Lovas and Ting, 1992) that testing huge quantities of drugs exposes forensic scientists who are carrying out such tests to chemicals that may create potential health hazards.

Together, these issues provide powerful arguments in favor of randomly sampling the seized evidence and thereby estimating Q. Both the crime laboratories and the courts would benefit from having a modest-sized sample carefully tested for drugs; a more accurate estimate of Q would then be obtained rather than that from possibly low-quality testing of every item in a huge drug seizure. In *United States v. Scalia* (1993), the court noted that "statistically based drug-quantity extrapolations predicted on random test samples" were acceptable, that "drug-quantity estimates need not be statistically or scientifically precise" and that, while case law permits extrapolation in general, the government has to be able to demonstrate "an adequate basis in fact for the extrapolation and that the quantity [has to be] determined in a manner consistent with the accepted standards of [reasonable] reliability."

3.2 Expert Forensic Testimony

When prosecuting a defendant in a drug trial, the government usually calls upon a forensic chemist to justify its estimate of Q. The following sequence of events constitutes the most important part of the forensic chemist's testimony.

3.2.1 *Homogeneity of substances.* Through visual examination, the chemist first determines that the contents of the K containers appear homogeneous. The concept of homogeneity is linked directly to the desire to do random sampling. If every item were tested completely for contraband, there would be no need to raise questions of homogeneity of those units. As the court noted in *People v. Kaludis* (1986), "[The] sole question to be answered before expert opinion based on random sampling would be allowed was whether [the] substance was homogeneous." This is usually done through a visual examination of the evidence. The different types of illicit drugs have provided courts with opportunities to decide on the homogeneity issue by using specific arguments to fit the substance in question.

Certain courts have held that individual units of particular types of substances are so alike that homogeneity of the seized material may not be difficult to determine, even by visual inspection. These include pills, tablets and capsules. In *Kaludis*, for example, a forensic chemist testified that

> the tablets [suspected of containing methaqualone, a Schedule II controlled substance commonly known as Quaaludes or ludes] were all the same size, diameter, roundness and thickness.

> The tablets also were all off-white in color and exhibited the same hardness. All tablets had the same markings "Lemmon 714," which were markings used by a particular pharmaceutical company . . . the tablets exhibited identical lettering characteristics, bevelling and scoring,

leading the chemist to conclude that, although the tablets were judged to be counterfeit since they were not manufactured by the Lemmon Company, they were all "manufactured on the same tablet press with the same set of dies."

The court in *Bond v. State of Florida* (1989) agreed that rocks of crack cocaine can also be regarded as homogeneous because they resemble similar-looking tablets rather than powder cocaine. The decision in *Bond* has since been followed in other jurisdictions. In *State of Missouri v. Gibson* (1993), the homogeneity of numerous off-white rocks of crack cocaine that were found in the same container was successfully established by a criminalist who explained that crack cocaine is made "by mixing cocaine powder with baking soda and water and cooking it. The resulting crystalline will fall to the bottom and is cut up into smaller chunks. . . [that are] uniform in color and consistency." This explanation helped the court determine that the rocks in question were all cut from a single "batch" of rock cocaine, rather than from separate sources. See also *Gabriel v. State of Texas* (1995). Similar arguments about the visual homogeneity of kilogram bricks of crack cocaine have appeared in testimony (*People v. Peneda*, 1995; *United States v. Pirre*, 1991).

Another illicit drug for which it is not difficult to establish visual homogeneity is LSD on blotter paper. In *People v. Saldana* (1986), for example, the evidence consisted of 20 sheets of paper, each sheet containing "orange-colored flying saucer symbols" and approximately one hundred "hits" of LSD. A Drug Enforcement Administration (DEA) forensic chemist testified that

> in his scientific opinion, based on a reasonable degree of scientific certainty, all 20 sheets had LSD evenly distributed over the surfaces. . . . Each piece of blotting paper was perforated so that there were 20 squares approximately one quarter inch on each side on each paper so that the total number of dosage units was 2,000. . . . Since each paper was uniformly coated, it appeared that the paper had been dipped in LSD.

This analysis was sufficient for the court to accept the homogeneity of the sheets of LSD.

When loose powder is found wrapped in some packaging medium, such as small plastic bags, and where the type of packaging gives no visual clues of homogeneity, then more detailed testing is necessary. Indeed, cases have often been reversed because of a lack of careful testing of the contents of every package. For example, the court in *Ross v. State of Florida* (1988) found a chemist's testimony at trial that 90 untested packets of white powder "looked alike from a visual inspection" to be insufficient for conviction, arguing that "the white powder contained therein may be milk sugar or any one of a vast variety of other white powdery chemical compounds not containing cocaine." In *State of Minnesota v. Robinson* (1994), a chemist noted in testimony that "there were numerous situations where she has tested questionable material and found it not to be cocaine. Some of the common substitutes for cocaine were baking soda, powdered sugar and soap pieces." The court noted in its decision that "in the case of substances not homogeneously packaged, drug dealers are known to substitute placebos for the real thing."

In cases where drugs are smuggled into the country, homogeneity is easy to argue. For example, in *United States v. Shonubi* (1995), where 103 balloons carried internally were found to contain white powder, the court noted that homogeneity could be ascertained based upon the following assumptions about human behavior: "[N]o one carries sugar in balloons from Nigeria to New York. . . 4 balloons chosen at random out of 103, when there is no ostensible reason for variation, is a sufficient sample. . .the balloons were all filled from the same batch of heroin, using the same filling technique. . .[and it was highly likely that] the 103 balloons [contained] heroin in roughly similar amounts."

3.2.2 *Preliminary screening tests.* Next, using a preliminary screening test, the chemist determines whether each of the K containers is likely to include an illicit drug. Usually, prior to arresting an individual for drug possession or trafficking, law enforcement officers would use a preliminary screening or field test (e.g., a color or crystal test, using a cobalt thiocyanate reagent) on a sample of the substances to ascertain whether illicit drugs are likely to be present. These screening tests, whose error rates are at most 1% (comparable to that for the ELISA screening test for detecting HIV infection in blood samples; Gastwirth, 1987; Brookmeyer and Gail, 1994, Chapter 6), cannot

conclusively identify a substance to a "reasonable scientific certainty."

3.2.3 *Random sampling of the evidence.* The chemist next explains how a random sample of the ingredients was selected from the containers. Sampling in drug cases has been used in certain jurisdictions since the 1960s to detect and identify drugs seized from a defendant. However, early uses of sampling in drug cases were typically ad hoc. Sample sizes were conveniently chosen by the chemist and no information was given as to whether or not the sampling was carried out randomly. Appeals of subsequent convictions often depended less on the randomness of the sampling mechanism than on legal issues regarding measurement procedures. Today, sampling procedures, such as simple random sampling, stratified sampling, multistage sampling and composite sampling, are used in different jurisdictions, but the courts decide whether or not a particular procedure is appropriate for the case in question.

Courts have been very accepting of sample data even if there exist alternative methods of proof which may be difficult to carry out. In many instances, samples have been introduced into evidence to establish a particular point and to illustrate the condition, quality or nature of a large amount of material that is involved in the matter under litigation but is not accessible to the court. Wigmore (1979, Section 439) gave two requirements for a sample, presented to show the quality or condition of the entire lot or mass from which it was taken, to be admissible in evidence: (1) the mass should be substantially uniform with reference to the quality in question (the "homogeneity" requirement) and (2) the sample portion should be of such a nature as to be fairly representative (see Kruskal and Mosteller, 1979).

3.2.4 *Definitive tests.* The chemist then carries out a series of definitive chemical tests (e.g., infrared or ultraviolet spectrophotography, thin-layer chromatography and gas chromatography coupled with mass spectrometry) on the sample of ingredients. Certain of these tests produce a spectrum graph; the number of observed peaks in the graph, their positions, dimensions and shapes are then compared with the standard spectrum of a given illicit drug. These tests are routinely used, are highly reliable and are usually combined for conclusively identifying to "a reasonable degree of scientific certainty" that the substances seized contain an illicit drug.

3.2.5 *Purity of the drugs.* The test by gas chromatography determines the purity (in percent) of the drugs. Because drugs are diluted as they move down the chain of distribution, purity of drugs seized can reflect a defendant's position in this chain and thus may be relevant to the sentencing process (United States Sentencing Commission, 1995, Section 2D1.1(c), application note 9).

3.2.6 *Estimation of Q.* The chemist weighs the contents of only those containers from which the sampled ingredients *that were conclusively proved to contain an illicit drug* were obtained. Suppose k containers are randomly drawn from the K seized containers and then a subsample of size n_j is randomly drawn from the N_j items contained in the jth selected container. Note that there is the possibility that all the containers may be chosen. We also assume that subsampling of one container is independent of subsampling of all other containers. Each of the $n = \sum_j n_j$ sampled items is then tested for the presence of an illicit drug.

As before, let X_{jt} be the weight of the tth sample item from the jth selected container if that item tests positive for an illicit drug, and $X_{jt} = 0$ if it tests negative. Then the two-stage Horwitz–Thompson estimator,

$$\widehat{Q}_{2st} = \frac{K}{k} \sum_j \frac{N_j}{n_j} \sum_t X_{jt}, \tag{3.2}$$

is an unbiased estimator of the quantity Q defined in (3.1). Note that the summations in (3.2) are taken over the k selected containers (summation over j) and over the n_j sample items in the jth selected container (summation over t). It is not difficult to show (Särndal, Swensson and Wretman, 1992, Chapter 4) that the variance of (3.2) is given by

$$\begin{aligned} \operatorname{var}(\widehat{Q}_{2st}) &= K^2 \left(\frac{1 - \pi_1}{k} \right) S_1^2 \\ &\quad + \frac{K}{k} \sum_{j=1}^{K} N_j^2 \left(\frac{1 - \pi_{2j}}{n_j} \right) S_{2j}^2, \end{aligned} \tag{3.3}$$

where $\pi_1 = k/K$, $\pi_{2j} = n_j/N_j$, $S_1^2 = (K-1)^{-1} \times \sum_{j=1}^{K} (Y_j - \overline{Y})^2$ and $S_{2j}^2 = (N_j - 1)^{-1} \sum_{t=1}^{N_j} (X_{jt} - \overline{X}_j)^2$; $Y_j = \sum_{t=1}^{N_j} X_{jt}$, $\overline{Y} = \sum_{j=1}^{K} Y_j / K$ and $\overline{X}_j = \sum_{t=1}^{N_j} X_{jt}/N_j$. [If the jth container holds loose cocaine or heroin powder, then $n_j = N_j = 1$ in (3.2) and (3.3).] An unbiased estimator of (3.3) is given by $\widehat{\operatorname{var}}(\widehat{Q}_{2st})$, which is obtained by substituting $s_1^2 = (k-1)^{-1} \sum_j (y_j - \bar{y})^2$ for S_1^2 and $s_{2j}^2 = (n_j - 1)^{-1} \sum_t (X_{jt} - \bar{x}_j)^2$ for S_{2j}^2 in (3.3), where $y_j = (N_j/n_j) \sum_t X_{jt}$, $\bar{y} = \sum_j y_j / k$, $\bar{x}_j = \sum_t X_{jt}/n_j$,

and where summations are now taken over sample quantities as in (3.2). Similar results for estimators derived from three-stage sampling (taking account of possible batching of containers) can be easily obtained, and the above expressions can also be specialized for stratified sampling and simple random sampling situations.

We note that in New York (see, e.g., *People v. McLaurin*, 1993), the figure that gets reported to the court and determines the defendant's responsibility is a total "pure weight," which is determined by multiplying $\widehat{Q}$ and the estimated purity.

If convicted of drug trafficking, the defendant typically challenges the value of $\widehat{Q}$ that is used for sentencing, arguing that the prosecution failed to prove beyond a reasonable doubt that he or she possessed more than a statutory threshold amount of the illicit drug. Usually, at issue is the legal weight to give the results of random sampling in such cases because not all units are tested. Appeals courts often note that, although the defendant had the chance at trial to carry out his or her own sampling of the evidence and thereby provide an alternative estimate of Q, the defendant never took that opportunity.

3.3 The Courts' Views of Random Sampling in Drug Cases

The practice of random sampling in drug cases has been described in *People v. Games* (1981) as "a rule of reason and practicality" and in *State of Louisiana v. Ballom* (1990) as "reasonable and reliable." In *United States v. Shonubi* (1995), the court noted that "sampling. . .is common in drug cases. In the Second Circuit,. . .sampling is often the basis of drug quantity determinations." In *United States v. McCutchen* (1993), a police chemist testified that it was accepted practice within the Philadelphia police department to test a representative sample, rather than all of the seized drugs. When asked whether it was common to test each and every vial that was contained in a package, the chemist replied, "No, it's not common. It would be a very rare instance when we tested every vial in a submission this large [119 vials]. It would be way too time-consuming. We're very busy with a lot of cases. And what we do is we take a representative sample and then we use that to project what is probably in the rest of them."

In accepting the admissibility of random sampling for drugs, the courts have explained their opinions in very similar ways: "It is well established that the prosecution need not prove that every grain of white powder seized in a single shipment *is* cocaine as long as cocaine is present within the single unit" (*People v. Peneda*, 1995); "To require the State to test each tablet of the 1,000 tablets delivered by appellant and prove that enough tablets contained enough pure methaqualone to satisfy the statutory amount is patently unreasonable" (*Asmer v. State of Florida*, 1982); "The law does not require each molecule, grain, leaf or stalk to be tested separately. Neither does the law prohibit an expert from arriving at a logical conclusion based upon a random sampling" (*Wright v. State of Florida*, 1977); "[The State] need not necessarily count, analyze, and weigh each individual capsule from every bag recovered from the defendant" (*People v. Yosell*, 1977); "A chemist is permitted to analyze a small amount of substance and give an opinion as to the whole" (*People v. Kline*, 1976); and "We do not believe that our statute requires that each and every leaf fragment must be tested individually to determine whether it is marihuana. The law does not demand impossibilities" (*State of Missouri v. Edwards*, 1975). Similar remarks about not having to test "every particle of white powder in every capsule" for LSD can be found in *People v. Hering* (1975).

Certain jurisdictions have decided that it is crucial whether the substances seized from a defendant are found in a single container or in multiple containers. Recent cases, however, have indicated that sampling of multiple containers has now become a controversial topic. In this section, we describe the various situations that have occurred.

3.3.1 *Single container.* In general, examination by a chemist of a random sample of a substance seized within a single bag or container has been accepted by the courts to prove the identity of the remainder of the substance in the container. For example, in *Kaludis*, the court affirmed the conviction in which a chemist randomly selected three tablets from a bag containing 100 homogeneous tablets found in the defendant's possession, tested the three tablets positively for methaqualone, and concluded that all 100 tablets contained methaqualone.

3.3.2 *Multiple containers.* Where the suspected contraband is found in multiple bags, containers or receptacles, certain jurisdictions have specifically distinguished the sampling situation from that of sampling similar-looking pills found in a single container.

The first specific sampling policy for multiple containers was formulated by the Illinois appellate courts. When a defendant can be charged with the lesser included offense of possession of a smaller amount of an illicit drug, the weight of the seized drug is an essential element of the crime

and must be proved beyond a reasonable doubt. When homogeneity of the various containers has not been proved definitively, therefore, the rule is as follows:

1. at least a sample from *each* container should be conclusively tested for the presence of an illicit drug;
2. the contents (and weight) of untested containers may *not* be considered in determining the severity of the offense;

and

3. if the total weight of the sampled containers that conclusively tested positive for an illicit drug exceeds the minimum amount needed to prove the charge against the defendant, then (1) can be ignored.

In essence, then, the courts regarded the containers as strata. We shall henceforth find it convenient for exposition purposes to refer to this as the "stratified sampling rule."

This sampling policy was initially based upon a series of three cases (*People v. Games*, 1981; *People v. Ayala*, 1981; *People v. Hill*, 1988), each of which used $\widehat{Q}_{2st}$ in (3.2) to estimate Q. In each case, a chemist first proved that a few of the seized bags each contained drugs, and then estimated the total weight of *all* bags by assuming the remaining untested bags also contained drugs. Because Q was an element of the crime and its value had to be proved beyond a reasonable doubt, the courts in all three cases refused to accept the inference that the untested bags contained drugs; the defendants were convicted of lesser included offenses based upon lower statutory thresholds, and they received reduced sentences. Then, in *People v. Black* (1994), an Illinois appellate court decided to allow random sampling of containers (and, hence, the estimate $\widehat{Q}_{2st}$) by formally extending the *Kaludis* decision so that multiple "similar" containers could be treated for sampling purposes as if they were pills or capsules. This situation did not last long, however. In *People v. Jones* (1995), an Illinois appellate court found the *Black* opinion to be "overly broad" and reinstated the stratified sampling rule, saying that two-stage sampling "represents an unwarranted departure from the long-standing rule requiring the State to test an adequate number of samples with a sufficient combined weight to establish the elements of the offense." The *Jones* opinion was affirmed by the Illinois Supreme Court, which offered no new insight or principle, but just stated its support for the stratified sampling rule. The two justices who dissented, including Chief Justice Bilandic, noted that

> There is no reason to require the State in cases such as this to test the contents of each of the items the defendant has in his possession. Random sampling can provide circumstantial evidence of guilt, the strength of which will vary from case to case. Today's decision simply imposes an unnecessary burden on the State, making more difficult the prosecution of offenders who are found with contraband divided among multiple bags, packets, or other containers that, under the majority's rule, must now be tested individually.

Florida and Minnesota are the only other states that have since adopted the Illinois sampling rule in drug cases (*Ross v. State of Florida*, 1988; *State of Minnesota v. Robinson*, 1994). Most other jurisdictions appear to pay little or no attention to this issue.

Prior to *Apprendi v. New Jersey* (2000), all federal courts accepted multistage sampling because the standard of proof for sentencing decisions was preponderance of the evidence rather than beyond a reasonable doubt. The effect, if any, of *Apprendi* on how forensic scientists sample the evidence in federal drug cases to estimate Q has yet to be determined.

In the special case of rocks of crack cocaine, courts that have addressed the specific question have held that only a random sample of rocks needs to be tested to establish the required amount, whether found in one container or in individual packets. In *Bond v. State of Florida* (1989), where rocks of crack cocaine were equated to tablets rather than powder cocaine, the court held that random testing of one bag was appropriate to determine beyond a reasonable doubt that all 139 small plastic bags contained rocks of crack cocaine. See also *Carter v. State of Florida* (1994) and *State of Florida v. Meeks* (1989).

Federal courts have similarly affirmed the use of random sampling of containers with rocks of crack cocaine inside. Sometimes, however, sampling is carried out by convenience only. For example, a Philadelphia police chemist testified at the sentencing hearing in *United States v. McCutchen* (1993) that all 119 vials seized from the defendant were projected to contain crack cocaine based upon his analysis that 15 of the vials contained crack cocaine; when asked how the 15 vials were selected, the chemist replied that "Well, the original four, I did select two which had been field tested. And I tested two others at random. The 11 that I did

today, I just picked the first 11 that I got out of the bag." Although the sentence was affirmed, the court was unhappy with the sampling procedure and recommended that in future such cases more details should be presented to show that a representative sample had been selected.

3.3.3 *Composite sampling*. With the enormous costs incurred by forensic laboratories in analyzing the contents of hundreds, if not thousands, of multiple containers for illicit drugs, many laboratories have tried to reduce their workloads and costs by adopting alternative protocols, such as composite sampling. Federer (1984), in his review of cutting-edge issues in biometry, referred to composite sampling as a method that should be utilized more in laboratory analyses. Composite sampling, which has long been required by statute for government inspection and certification of certain agricultural products, has also been used in many other areas, including environmental management, industrial quality control and reliability, public health and disease monitoring, geochemistry and remote sensing and multiple-access communication problems. A recent review of composite sampling has been provided by Lancaster and Keller-McNulty (1998) and an annotated bibliography by Boswell, Gore, Lovison and Patil (1996).

When faced with materials from a number of separate sources, the composite sampling algorithm is to take a "core sample" from each source, thoroughly mix the samples and then subsample from the mixture. The mixture of core samples is called a "composite sample." Composite sampling is most cost-effective when the trait of interest (e.g., HIV antibodies) has a low prevalence rate in the population, where a large amount of source material is to be tested for that trait and where the cost of measurement is high. When core samples from each source are mixed and if the trait of interest is present in at least one of the sources prior to mixing, then a subsample from that mixture will, with high probability, test positive for that trait. If the subsample tests negative for the trait, then no more sampling or testing would be needed and all sources would be declared free of the trait. If, on the other hand, the subsample tests *positive* for the trait, then all that can be claimed from the results of that test is that *at least one source possesses the trait*. When the object is to identify exactly which of the sources possess the trait, then further sampling and repeated "group testing" would be necessary.

Our interest centers on mixtures of solid (rather than liquid) substances. Core samples are physically mixed or blended into a single homogeneous composite sample, which, if necessary, can then be reduced in size (possibly through subsampling) to a form suitable for laboratory testing. When mixing solid substances, the object of the sampling is to estimate the mean of some trait of interest (such as mass, weight or volume) usually related to quality of the material sampled, and also to determine the variability of that trait over all sources that made up the composite sample. Interest in the mean relates to marketing, pricing and quality assessment and control strategies for the materials in question. Duncan (1962) cautioned that "with random selection [without mixing] we know the risks that are involved, whereas until we have positive knowledge of what a given mixing process does we are running unknown risks."

In illicit drug cases, a forensic chemist typically forms a composite sample as follows (see, e.g., *People v. McLaurin*, 1993). A random sample of k out of the K seized containers is chosen; from the contents of each sample container, c "core samples" are removed and emptied into a mortar; the small fragments of powder in the mortar from all kc core samples are next ground with a pestle into a very fine powder (the "composite sample"); the powder is mixed by repeated tossing and stirring to look as homogeneous as possible; and then, using a spatula, R subsamples are scooped from around the mixture and placed into weighing dishes, where they are tested for the presence of an illicit drug. In practice, c is taken to be four or five and R is taken to be one or two.

A question that is sometimes raised at trial relates to how the law should treat the oversensitivity of the tests to small drug quantities in the mixture. This issue was raised in *Ross v. State of Florida* (1988), when the jury enquired whether "it was sufficient to find an entire substance to be cocaine so long as there is any trace amount in the mixture," a question the court did not answer directly. In *Robinson*, a chemist testified under cross-examination that "if one of the packets had contained 'a lot of cocaine' and three or four other packets contained only baking soda, the test would still have been positive." Actually, the main effect of mixing a certain amount of an illicit drug with other substances ("adulterants") is that although tests of the mixture will probably test positive, it is the purity measurement that is really affected; the greater the proportion of adulterants in the mixture, the lower the purity.

If composite sampling is used in an illicit drug case, the forensic chemist will have several objectives to accomplish if he or she has to testify about the nature of the contents of the K seized containers. These objectives include many of those

mentioned above. In particular, it is necessary to determine whether composite sampling can be used to (1) estimate the proportion of seized containers that contain an illicit drug, (2) investigate the issue of drug homogeneity, and (3) provide an accurate estimate of p, the overall purity of the illicit drug in question. Testimony by a forensic chemist on these issues will have important repercussions in determining the total amount Q of drugs seized from the defendant and, hence, the sentence imposed.

The first objective has been treated lightly (and incorrectly) by the courts, which readily accept testimony by forensic chemists that there are illicit drugs in *all* K containers when only a *single* subsample ($R = 1$) from a composite sample tests positive. As we noted previously, a positive result only implies that at least one of the core samples contains an illicit drug. In fact, appeals, which are taken precisely on this issue, usually succeed or fail not because of the poor inference that is made, but rather because the court focuses solely on the mechanism by which the composite sample was constructed. If a jurisdiction accepts only the stratified sampling rule, then it will accept such a poor inference from composite sampling provided that a portion of *every* container ($k = K$) was incorporated into the composite sample.

There do not appear to be uniform guidelines by which composite samples are constructed in testing for illicit drugs. At one extreme, some forensic laboratories combine *all* the contents of *each* container prior to sampling and testing (e.g., *Mello v. State of Texas*, 1991; *People v. Little*, 1986; *People v. Jackson*, 1985; *State of North Carolina v. Clark*, 1973). This strategy would certainly satisfy the stratified sampling rule. A problem with a composite sampling protocol that mixes the entire contents of all containers is that if the mixture tests positive, then retesting procedures to discover exactly which and how many of the containers possess the drug become impossible to carry out. A less extreme strategy, but one which would still satisfy the stratified sampling rule, is to form a composite sample by mixing *portions* of the contents of each of the seized containers (e.g., *Pugh v. State of Georgia*, 1995).

In some recent cases, forensic laboratories have shown a greater willingness to extend the use of composite sampling by employing full two-stage procedures, and some courts have shown a willingness to affirm convictions based upon such procedures (e.g., *People v. Rodriguez*, 1993; *State of North Carolina v. Harding*, 1993). Not all of the sentences of relevant cases were affirmed, however, with certain jurisdictions maintaining their strong hold on the stratified sampling rule and reducing a sentence to that of a lesser included offense if that rule was not followed (e.g., *State of Minnesota v. Robinson*, 1994).

It should not be surprising to note that composite sampling applied to illicit drug cases is regarded as a controversial procedure. The best example of this controversy appeared in a dissenting opinion in *Little*, which also criticized the decision in *Jackson*. In that opinion, the judge disagreed with the whole notion of composite sampling, saying that "I cannot believe that the majority here would have allowed the police to combine one of those manila envelopes with, say, a five pound bag of flour or sugar, had such an item also been in the defendant's possession. That, however, is where their logic leads them."

4. STATISTICAL ISSUES AND METHODS

Because of the need by courts to accumulate quantitative evidence for sentencing purposes in drug trials, more of the courts' attention has been focussed on statistical issues. As we have seen, the main issues include which sampling plan to implement in a given seizure, how many of the K containers should be randomly chosen for drug testing if the courts admit two-stage sampling procedures, how to assess homogeneity of the contents of seized containers, how to estimate purity p and total drug quantity Q over all containers and how to translate the quantitative evidence into posterior probabilities for standards of proof.

4.1 Choice of Sample Size

The choice of sample size to be used in selecting items randomly for testing from a population of containers has rarely been addressed by the courts in drug cases. However, attempts have been made by statisticians and forensic scientists to develop guidelines (and sometimes) formal rules for determining sample sizes in drug cases. These rules are similar to those developed in the sampling inspection literature (e.g., Wetherill and Brown, 1991, Chapters 11 and 12), which differentiates between inspection by variables and inspection by attributes.

There are situations in which seized items or containers appear to be visually homogeneous and little or no reason exists to suspect their contents of being anything but illicit drugs. A prime example is international drug smuggling (Izenman, 2000b, c). In such instances, the characteristic of interest is a continuous variable, the weight X of the contents of a container, and a sample size n is derived based upon a desired precision of the classical confidence intervals on the mean of X; see Cochran (1977, Chapter 4) and Desu and Raghavarao (1990)

for details. Some popular but nonoptimal rules that have been used for determining sample size in drug cases are described in Section 4.1.1.

More likely, it will not be known prior to laboratory testing whether the contents of the containers are homogeneous or not, or whether they even contain drugs. Such situations occur in drug seizures in the inner cities, where an unknown number of the seized containers may contain only look-alike substances, such as sugar, flour or pancake mix. It would be more appropriate in these latter instances to view the problem in attribute sampling terms in which each container is classified according as its contents test positive or negative for drugs. In that case, to satisfy certain statutory thresholds for sentencing, a sample size would have to be employed that would take into consideration the possible presence of containers whose contents might test negative for drugs. See Sections 4.1.2 and 4.1.3 for proposed approaches to this problem.

4.1.1 *The square-root and other popular rules.* A worldwide survey of sampling practices and choices of sample sizes for forensic drug analysis (Colon, Rodriquez and Diaz, 1993) found that the most popular rule for deciding how many containers or items, whether homogeneous or not, to sample for drug testing was *not* a statistically motivated one. Instead, the most popular rule was the *square-root rule*, $k = \sqrt{K}$ for containers and $n = \sqrt{N}$ for items, used by laboratories in Australia, Austria, Canada, England, New Zealand, Hong Kong, Northern Ireland, the United States and U.S.A. Army–Europe. One would assume, therefore, that the square-root rule would be an accepted part of sampling practice. Yet, in an informal, but extensive, survey of sampling practioners, we found that most sampling experts had never encountered the square-root rule and no textbook on sampling theory or practice nor review of the field (Stephan, 1948; Hansen, 1987; Bellhouse, 1988) even refers to it. Attempts have been made to ascribe optimal properties to the rule, but none are really convincing. Because the rule is popularly used by several U.S. government agencies, it is worthwhile setting down its obscure origins for the record.

The square-root rule apparently originated in the 1920s from a need to provide agricultural regulatory inspectors (specifically, those who knew how to extract a square-root) with a convenient, memorizable rule for sample size determination. In 1925, the Association of Official Agricultural Chemists (AOAC), which enjoyed a strong relationship with the U.S. Department of Agriculture, set up a Committee on Sampling (1925) to study all aspects of the sampling problem for agricultural research as well as for regulatory activities. The Report of that Committee (Blanck, 1927) formally adopted a proposal made in an unpublished 1919 report by the AOAC to use the square-root rule for sampling certain classes of foods. A companion article by Paul (1927) also recommended the square-root rule for sampling of bulk or powder drugs in small packages. Although Runkel (1926), in a separate AOAC survey of methods for sampling sacks of flour, declared that the square-root rule used by some baking chemists and millers had "no theoretical significance," his warning was basically ignored. Munch and Bidwell (1928) then provided a confused attempt to justify the square-root rule based upon inverting the formula for the probable error of a correlation coefficient. In the next 40 years, the AOAC's square-root rule operated as the standard for sampling such agricultural lots as wheat flour, feeding stuffs, boxed dried fruit and sacked cacao nibs. Runkel's warning was later repeated by Quackenbush and Rund (1967), but again this appears to have had no affect on AOAC sampling policy. In fact, in a December 1968 report of the Salmonella Committee of the National Research Council (Foster, 1971), the rule was referred to as a "commonly used [sampling] plan," and noted that "[o]n statistical grounds, the square-root plan is simply a practical rule-of-thumb guide."

The square-root rule has since been adopted by many federal regulatory agencies, most importantly by the DEA. See, for example, *People v. Hill* (1993) and *In re Lemons* (1991). Recommendations for drug testing are also published by the United Nations (1987), which recommends the square-root rule for composite sampling. The popularity of the square-root rule, despite the lack of theoretical support for this rule, shows how an unfounded rule-of-thumb can be established in the practice of a particular field.

Colon, Rodriguez and Diaz (1993) also noted that a 10% sample size rule is used by laboratories in Australia, Canada and the United States. See, for example, *State of Minnesota v. Robinson* (1994). As far as we are aware, the 10% rule can be traced back at least to Deming (1954). Furthermore, a 4% rule is used in England, half the square-root rule is used in Switzerland and, in the United States, some laboratories use a fixed sample size rule (1, 4 or 15) regardless of the amount of seized contraband. Based upon their survey, Colon, Rodriguez and Diaz recommended the 10% rule for use in Puerto Rico when the number of drug units is at least 20; for less than twenty, they recommended that all units be individually analyzed. [Dr. Shaun Burke, a consultant statistician for LGC (Teddington) Ltd, informed

us that an unusual choice of sample size is given by British Standard BS 5309 (1993), which uses the formula $n = 3 \times N^{1/3}$ to sample chemical products.] Funding limits have also been used to determine when to stop testing (*Gabriel v. State of Texas*, 1995).

4.1.2 *When not all items contain drugs.* When a seizure is made in which some of the containers or items may not contain illicit drugs, a different approach to sample size determination needs to be used. This problem has been considered by Frank, Hinkley and Hoffman (1991), Tzidony and Ravreby (1992) and Hedayat, Izenman and Zhang (1996).

Suppose the number N of seized items (or containers) is sufficiently small ($N \le 50$) that sampling cannot be assumed to be with replacement. Suppose also that out of the N seized items an unknown number N_0 would test negative for drugs while the remaining $N_1 = N - N_0$ would test positive. If we take a random sample of size n from all items, the probability distribution of Y, the observed number of negative items in the sample, is hypergeometric,

$$\text{Prob}[Y = y|N_0, N_1, n] = \frac{\binom{N_0}{y}\binom{N_1}{n-y}}{\binom{N}{n}}, \tag{4.1}$$

where $\max\{0, n - N_1\} \le y \le \min\{n, N_0\}$. Schuster (1991) uses the hypergeometric distribution (4.1) to choose a sample size in an anonymous Florida drug case that involved a "reverse cocaine sting" set up by the police department, where the defense argued that both positive and negative items could have been present in the seizure.

In forensic applications, an appropriate sample size n can be derived by considering the following testing problem:

$$\mathscr{H}_0\colon N_1 < k \quad \text{versus} \quad \mathscr{H}_1\colon N_1 \ge k,$$

where k is a value chosen by the forensic scientist. For example, k might be the minimum number of positive items in the entire seizure for the defendant to be sentenced to a given term of imprisonment. Frank, Hinkley and Hoffman (1991) set $k = \theta_0 N$, $0 < \theta_0 < 1$. Thus, with specified probability $1 - \alpha$ ($0 < \alpha < 1$), we wish to declare that the available evidence contradicts the possibility that $\mathscr{H}_0$ may be true for some prespecified k. In particular, if $\alpha = 0.05$ and $\theta_0 = 0.9$, then we would like to find a sample size n such that we can declare, with 95% confidence, that at least 90% of all N seized items contain illicit drugs.

The uniformly most powerful test rejects the above $\mathscr{H}_0$ in favor of $\mathscr{H}_1$ for the hypergeometric distribution (4.1) when Y is too small (Lehmann, 1986, page 80). We draw a random sample of size n from N and test all sample items. Let m_0 denote the number of items in the sample that the forensic scientist would allow to test negative for drugs. There is a small probability, $100\alpha\%$, that a Type I error (rejecting $\mathscr{H}_0$ because Y is observed to be smaller than m_0, while N_1 is actually smaller than k) will be made; that is,

$$\max_{N_1 < k} \text{Prob}[Y \le m_0|N_1] \le \alpha.$$

This is equivalent to

$$\max_{N_1 \le k-1} \text{Prob}[Y \le m_0|N_1] \le \alpha,$$

which, because $\text{Prob}[Y \le m_0|N_1]$ is decreasing in N_1, is satisfied if

$$\text{Prob}[Y \le m_0|N_1 = k - 1] \le \alpha.$$

This last expression is a cumulative hypergeometric probability, and so the required sample size n can be obtained from the inequality,

$$\sum_{i=n-m_0}^{n} \frac{\binom{k-1}{i}\binom{N-k+1}{n-i}}{\binom{N}{n}} \le \alpha, \quad 0 < \alpha < 1 \tag{4.2}$$

(Frank, Hinkley and Hoffman 1991). Of the five variables in (4.2), N is fixed and m_0 is chosen by the forensic scientist. Of the remaining three variables, knowing any two of them will yield the nearest value of the third for (4.2) to hold. Also, because k and n are integers, α cannot take all real values between 0 and 1. A computational algorithm would loop through possible values of k and n until it finds the largest value of k and the smallest value of n to satisfy (4.2). If we set $m_0 = 0$, then (4.2) reduces to the inequality

$$\frac{(k-1)!(N-n)!}{N!(k-n-1)!} \le \alpha. \tag{4.3}$$

For example, if we have $N = 150$ suspected drug items, then in order to claim with 95% confidence that the number of items N_1 that would test positive for drugs is at least $k = 0.9N = 135$, and allowing for no negative items in the sample ($m_0 = 0$), a sample size of $n = 25$ would be required based upon solving (4.3). If we allow for one negative in the sample ($m_0 = 1$), the required sample size increases to 39; allowing for two negatives ($m_0 = 2$), the required sample size becomes 50 (Frank, Hinkley and Hoffman, 1991, Table 2).

Let m denote the number of negative items out of the n. If $m \le m_0$, we are done and, so, with probability $1 - \alpha$, we reject $\mathscr{H}_0$. If, on the other hand, $m > m_0$, then a sequential (multistage) operation can be introduced into the decision procedure. Such an idea was hinted at by Frank, Hinkley and Hoffman (1991). A statistical problem is to design

such a multistage sampling procedure while preserving the α-level of significance. See Hedayat, Izenman and Zhang (1996) for further comments. In a private communication, Hedayat reports that a beta version of a software program for forensic drug sampling based upon solving (4.2) is available.

4.1.3 *A Bayesian approach.* A Bayesian approach to the sample size problem considered in Section 4.1.2 was studied by Aitken (1999, 2000). Consider a large seizure of N items suspected of containing drugs. As before, Y is the number of items out of a random sample of size n from N that do not contain drugs. Assume that the seizure is large enough that sampling can be regarded as being with replacement. Then Y has the binomial, $\text{Bin}(n, 1-\theta)$, probability distribution, where θ is the chance that any individual item contains drugs. In a Bayesian approach, the likelihood function,

$$L(y|\theta, n) = \binom{n}{y}\theta^{n-y}(1-\theta)^{y}, \quad 0 < \theta < 1, \tag{4.4}$$

is used to update prior knowledge about θ, which can conveniently be expressed in terms of a $\text{Beta}(a, b)$ conjugate prior density,

$$f(\theta|a, b) = [B(a, b)]^{-1}\theta^{a-1}(1-\theta)^{b-1}, \quad 0 < \theta < 1,\ a, b > 0, \tag{4.5}$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, and $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt = (a-1)!$ is the gamma function. The resulting posterior distribution,

$$f(\theta|y, n, a, b) = L(y|\theta, n) \times f(\theta|a, b), \tag{4.6}$$

reduces to a $\text{Beta}(n-y+a,\ y+b)$ distribution.

Suppose, now, we wish to determine n such that we will be $100(1-\alpha)\%$ certain that at least $100\theta_0\%$ $(0 < \theta_0 < 1)$ of the seized items contain drugs if every one of the n sample items tests positive for drugs. That is, if $x = 0$. Then, we write that

$$[B(n+a, b)]^{-1}\int_{\theta_0}^{1}\theta^{n+a-1}(1-\theta)^{b-1}d\theta = 1-\alpha. \tag{4.7}$$

A choice of n can be determined by substituting specified values of a, b, θ_0 and α into (4.7), evaluating the resulting integral and then solving for n using a trial-and-error strategy. For example, suppose $\alpha = 0.05$ and we set $a = b = 1$ to represent maximum uncertainty about θ. If we also set $\theta_0 = 0.5$, the scientist would only have to examine $n = 4$ items, while, for $\theta_0 = 0.9$, the required sample size increases to $n = 28$. If we allow one item in the sample to test negative for drugs $(x = 1)$, then solving (4.7) with $\theta_0 = 0.5$, it can be shown that n has to be at most 7. In other words, if one out of the four randomly selected items tests negative for drugs, we can select an additional three items from the seizure; if all three additional items test positive (making 6 out of 7), then there is a 96% chance that at least 50% of the items in the entire seizure contain drugs. Further examples can be found in Aitken (2000).

4.2 Group Testing

One might think that the various algorithms for group testing should apply also to composite samples formed from mixtures of powdery substances. We have already noted that an essential element of the group testing methodology is that the trait is assumed to be rare in the population being studied. In any given drug seizure, however, the trait—presence of a drug—has an extremely high probability of being detected in any composite sample. Indeed, as we argued above, the chance would be high even if some components of the composite sample were not drugs. This situation is contrary to the usual group-testing setup. To try to shoehorn the problem to make it fit a group-testing scenario is even more difficult. If we try to reverse the problem by designating, instead, the absence of a drug—the rare event—as the trait to be studied, this would not help us because the absence of a trait is impossible to detect through group testing. These considerations, therefore, make a composite sampling strategy for determining the proportion of seized containers that contain an illicit drug (and hence estimating Q) an impossible statistical problem.

4.3 An Estimate of Overall Purity

It is possible to use composite sampling to estimate the overall purity p of the substances in the K containers. Let $0 \le P_{ij} \le 1$ be the purity of the entire ith core sample that was extracted from the jth sample container (henceforth termed the "(i, j)th core sample"), $i = 1, 2, \ldots, c$, $j = 1, 2, \ldots, k$. Let g_{rij} be the amount (or weight) of the (i, j)th core sample that is incorporated into the rth subsample, and set $G_r = \sum_i \sum_j g_{rij}$ to be the total weight of the rth subsample, $r = 1, 2, \ldots, R$. The purity $\hat{p}_r$ of the rth subsample can be modelled as a weighted sum of the P_{ij},

$$\begin{aligned}\hat{p}_r &= \sum_{i=1}^{c}\sum_{j=1}^{k}\frac{g_{rij}}{G_r}P_{ij} \\ &= \sum_{i=1}^{c}\sum_{j=1}^{k} w_{rij}P_{ij}, \quad r = 1, 2, \ldots, R,\end{aligned} \tag{4.8}$$

where the weight w_{rij} represents the proportion of the rth subsample that comes from the (i, j)th core

sample. Thus, $\sum_i \sum_j w_{rij} = 1$. If we assume that the weights $\{w_{rij}\}$ in (4.8) are fixed and known, and that each core sample from the ith sample container contributes the same amount of substance to the rth subsample, then $g_{rij} = g_r, i = 1, 2, \ldots, c, j = 1, 2, \ldots, k$, whence $w_{rij} = 1/kc$ and $\hat{p}_r = \bar{P} = \sum_i \sum_j P_{ij}/kc$. In this case, the estimator of overall purity based upon the rth subsample is the same as would be obtained by averaging the purities from each of the core samples that make up that subsample.

A more reasonable model, however, would consider the weights in (4.8) as random variables. The sampling properties of an estimate of p can then be determined using a random effects model, but where sampling is carried out from a finite population. Although there is an enormous literature on linear models in general and random effects models in particular, almost all of this work assumes that the random effects are drawn from infinite populations. The finite population case is relevant here because a court would only be interested in the contents of the specific K containers seized from a defendant. The few articles that deal with the finite population case (Gaylor and Hartwell, 1969; Searle and Fawcett, 1970) give rules for converting expectations of mean squares under infinite population models into expectations under finite population models.

Thus, the purity of the rth subsample can be modelled as the bilinear random form,

$$\hat{p}_r = \sum_{i=1}^{c} \sum_{j=1}^{k} W_{rij} P_{ij}, \quad r = 1, 2, \ldots, R, \tag{4.9}$$

where the weights $\{W_{rij}\}$ and the purities $\{P_{ij}\}$ are independent random variables. The weights are assumed to have equal means, $\mathscr{E}(W_{rij}) = \mu_W$, and equal variances, $\mathrm{var}(W_{rij}) = \sigma_W^2$. Brown and Fisher (1972) showed that σ_W^2 can be thought of as the variability due to nonperfect mixing of the composite samples. They also showed that if the core samples from each of the sample containers were perfectly blended together to form a composite sample, then $\sigma_W^2 = 0$, which means that σ_W^2 could be used to characterize the blending process. See also Elder, Thompson and Myers (1980). In (4.9), the $\{P_{ij}\}$ can be described by a random effects model,

$$P_{ij} = p + t_j + e_{ij}, \quad i = 1, 2, \ldots, c, \;\; j = 1, 2, \ldots, k, \tag{4.10}$$

where $0 \le p \le 1$ is the true purity of the substances over all K containers, t_j is the random effect due to the jth sample container with $\mathscr{E}(t_j) = 0, \mathrm{var}(t_j) = \sigma_t^2$ and e_{ij} is a random error component with $\mathscr{E}(e_{ij}) = 0, \mathrm{var}(e_{ij}) = \sigma_e^2$. We further assume that $\mathrm{cov}(t_j, t_{j'}) = \mathrm{cov}(e_{ij}, e_{ij'}) = \mathrm{cov}(t_j, e_{ij}) = 0, \;\; j \ne j'$.

Our estimate of purity p is then the average of (4.9) over all R subsamples,

$$\bar{p} = \frac{1}{R} \sum_{r=1}^{R} \hat{p}_r, \tag{4.11}$$

whence $\mathscr{E}(\bar{p}) = p$ and

$$\mathrm{var}(\bar{p}) = \frac{kc}{R}\sigma_W^2\sigma_e^2 + \frac{kc^2(k-1)}{R(kc-1)}\sigma_W^2\sigma_t^2 + \frac{\sigma_e^2}{kc} + \frac{\sigma_t^2}{k}. \tag{4.12}$$

A proof of (4.12) is given in the Appendix together with a generalization of this result to the case in which a different number of core samples are taken from each container. Rohde (1976) notes that the usual definition of sample variance, $\sum_{r=1}^{R}(\hat{p}_r - \bar{p})^2/(R-1)$, is a biased estimate of $\mathrm{var}(\bar{p})$ because the $\{\hat{p}_r\}$ are correlated.

To calculate (4.12), we need to determine σ_W^2, σ_t^2 and σ_e^2. Following suggestions of Brown and Fisher (1972), we can take additional core samples from each sample container, randomly allocate them into two groups and then, by measuring the purity of every core sample in the first group, use an analysis of variance technique to estimate the variance components σ_t^2 and σ_e^2, and use blended core samples from the second group to estimate σ_W^2 based upon the estimates of σ_t^2 and σ_e^2. An alternative method of estimating σ_W^2 was given by Rohde (1976), who argued for the Dirichlet distribution. Duncan (1962) discussed the advantages of preparing two composite samples, subsampling from each independently and then combining the resulting estimates.

4.4 A Strategy for Assessing Homogeneity of Substances

Recall that the question of homogeneity is of interest because of the need to argue at trial that the contents of the untested containers are similar to those of the tested ones and, hence, that any statistical inference resulting from the sample containers is valid for the entire seizure. If the current practice of identifying substance homogeneity through a visual screening only is to be improved, it is necessary to adopt a more scientific approach to the problem. Toward this end, define the contents of K containers to be "completely homogeneous" if they satisfy two criteria: (1) the substances found in every one of the K containers are identified through testing as the same drug and (2) the purity of the contents in every container is determined to be roughly the same. In most laboratory situations, and especially

if K is very large, it may be impossible to carry out all the required tests needed to confirm complete homogeneity. This leads us to consider strategies based upon composite samples.

Consider the following sampling procedure. First, divide the K containers randomly into $B \geq 1$ batches, where the bth batch is assigned K_b containers, $b = 1, 2, \ldots, B$, so that $K = \sum_{b=1}^{B} K_b$. The assignment should be made so that each batch has approximately the same number of containers. (We note that choice of B for composite sampling and testing has been studied within the group-testing literature; see, e.g., Hughes-Oliver and Swallow, 1994.) Next, take a random sample of size k_b of the containers within the bth batch. From each of the sample containers in each batch, take c "core samples" of its contents, mix the core samples thoroughly to provide a "batch composite sample" (or "batch-mixture") and then take $R \geq 2$ subsamples from each batch-mixture. This strategy would be appropriate both for jurisdictions that allow multistage sampling and for those that adhere to the stratified sampling rule (in which case, take $k_b = K_b, b = 1, 2, \ldots, B$). The idea of "batching" the containers is similar in spirit to the work of Gastwirth and Hammick (1989) and Hammick and Gastwirth (1994), who were interested in estimating the prevalence of a rare disease or trait in a population without having to measure any individuals.

The next step is to apply a preliminary screening test for an illicit drug to each of those BR subsamples. For jurisdictions that follow the stratified sampling rule, we sample the contents of all K containers, but only carry out BR tests. In general, BR should be much smaller than K. For example, we could divide $K = 100$ containers into $B = 5$ batches each of 20 containers, prepare 5 batch-mixtures by taking core samples from all containers in each batch and then take $R = 2$ subsamples from each batch-mixture. This results in 10 subsamples to be tested. Each subsample would then be given a screening test and a confirmatory test, and its purity determined. The multiple confirmatory tests on each batch-mixture operate as homogeneity checks on the container distribution of illicit drugs.

Let $0 \leq \hat{p}_{br} \leq 1$ be the estimate of purity of the rth subsample, $r = 1, 2, \ldots, R$, obtained as in Section 4.3 from the bth batch-mixture, $b = 1, 2, \ldots, B$. If the confirmatory test on the rth subsample from the bth batch-mixture is negative, set $\hat{p}_{br} = 0$. On the assumption that this will be a rare event, we ignore this possibility in our description of this procedure. The purity data $\{\hat{p}_{br}\}$, therefore, display the classic form of a one-way layout, with R observations being taken from each of B batch-mixtures, the only difference being that the R purity measurements within each batch-mixture are not independent observations. To carry out an exploratory check of the homogeneity of the substances in the containers of the bth batch, we can compare the purities of the subsamples from the bth batch-mixture. Those purities should be similar to one another since they are correlated.

We can also compare the homogeneity of substances across the B batch-mixtures. Suppose p_b is the true purity of the bth batch-mixture. A test for homogeneity would have as null hypothesis

$$\mathscr{H}_0\colon\ p_1 = p_2 = \cdots = p_B \tag{4.13}$$

and alternative hypothesis that at least one of the batch-mixture purities differs from the rest. If $\mathscr{H}_0$ is rejected, one can argue that the substances in the K containers are not completely homogeneous.

We estimate p_b by the average of the R purities from the bth batch,

$$\bar{p}_b = \frac{1}{R}\sum_{r=1}^{R} \hat{p}_{br}, \quad b = 1, 2, \ldots, B, \tag{4.14}$$

and, if $\mathscr{H}_0$ is true, an unbiased estimate of the common purity p across all B batches (and, hence, across all K containers) is given by the pooled value

$$\bar{p} = \frac{1}{B}\sum_{b=1}^{B} \bar{p}_b. \tag{4.15}$$

To test $\mathscr{H}_0$, we assume, for fixed batch-mixture b, that $[\hat{p}_{b1}, \ldots, \hat{p}_{bR}]^\tau$ is a normally-distributed R-vector variate whose elements have equal variances and equal bivariate correlations. We also note that the $\{\bar{p}_b\}$ are independent. Then, it can be shown (e.g., Huynh and Feldt, 1970) that the usual F-statistic for a one-way layout,

$$\mathscr{F} = \frac{(R/(B-1))\sum_{b=1}^{B}(\bar{p}_b - \bar{p})^2}{[1/(B(R-1))]\sum_{b=1}^{B}\sum_{r=1}^{R}(\hat{p}_{br} - \bar{p}_b)^2}, \tag{4.16}$$

can be used to test $\mathscr{H}_0$, where $\mathscr{F} \sim F_{B-1,\,B(R-1)}$, and $\mathscr{H}_0$ is rejected for large values of $\mathscr{F}$. Although, for fixed b, the $\{\hat{p}_{br}\}$ are correlated, this neither affects the independence of the numerator and denominator of $\mathscr{F}$, nor the exact distributional theory.

If $\mathscr{H}_0$ is rejected, it may be because a purity value, $\bar{p}_{b_0}$, say, obtained from the b_0th batch-mixture is substantially smaller than the rest. This may indicate that the contents of one or more of the constituent containers of the b_0th batch either might contain drugs of lower purity than the other containers in that batch or might in fact test negative if individually tested. The next step in that case

would be to test individually all the containers in the b_0th batch for a possible lack of drugs. Sampling thus produces a determination of "statistical homogeneity," which should be sufficiently scientific for a court to accept *in lieu* of testing for complete homogeneity.

4.5 Standards of Proof and Q

There are often situations where 100% inspection (or even sampling) is out of the question and the court has to sentence a defendant based only upon incomplete or second-hand knowledge about his or her drug trafficking activities. "Incomplete" information can take different forms: (a) the defendant may have engaged in several suspected smuggling trips, but only the last one yielded real data (e.g., *United States v. Shonubi*, 1995), (b) evidence has been deliberately destroyed [either by the defendant, who wishes to hide the crime (e.g., *United States v. Hilton*, 1990), or by the government, which may need to obtain more storage space (e.g., *United States v. Pirre*, 1991)]. "Second-hand" information often occurs when evidence of a "continuing criminal enterprise" is not directly available and instead has to be pieced together based upon witness testimony of the defendant's drug-buying or drug-selling activities (e.g., *United States v. Whiting*, 1994). In such situations, the courts have to determine, using an appropriate standard of proof, which of the proferred information is reliable and admit an estimate of Q accordingly.

How, then, should a court decide on a minimum value of Q that would satisfy the preponderance standard? Or the reasonable doubt standard? These questions have been considered by Aitken, Bring, Leonard and Papasouliotis (1997) and Bring and Aitken (1997) for situations in which it would be natural to believe that the contents of every container or item, if indeed they were available, would test positive for illicit drugs. These authors suggest several different approaches to these problems, each directed towards the specific circumstances of individual cases. In general, they propose plotting the reliability function,

$$\overline{F}(q) = 1 - F(q) = \text{Prob}[Q > q], \tag{4.17}$$

where F is the cumulative distribution function of Q, against q, and then defining the values q_R, q_C, and q_P ($q_R < q_C < q_P$) such that beyond a reasonable doubt translates as $\overline{F}(q_R) = 0.95$, clear and convincing evidence as $\overline{F}(q_C) = 0.70$ and preponderance of the evidence as $\overline{F}(q_P) = 0.50$. If, for example, the reasonable doubt standard were applied, the defendant would be sentenced for possessing q_R grams or more of illicit drugs. The shape of the reliability function (4.17) would depend upon an estimate of F as determined by the specific case in question.

In cases where all seized containers are available, but a sample of them is taken for estimating Q, Aitken et al. (1997) derived a predictive density $f(q|\mathbf{x})$ for Q given drug weights $\mathbf{x}$ from all sampled containers. They used a normal prior on the mean θ and an independent chi-squared prior on the variance ϕ of a normal distribution of a container's drug weight, and then computed a (predictive) reliability function $\overline{F}(q|\mathbf{x}) = \text{Prob}[Q > q|\mathbf{x}]$ for some specified threshold quantity q of drugs, thereby obtaining $q_R(\mathbf{x})$, $q_C(\mathbf{x})$ and $q_P(\mathbf{x})$. Extensions of some of these results to the multivariate case were provided by Izenman, Papasouliotis, Leonard and Aitken (1998), whose primary focus was on how laboratory measurement error affected the estimation of Q.

5. DISCUSSION

In this article, we have sought to understand the statistical and legal perspectives involved in the forensic study of illicit drugs. Whenever crime laboratories take samples of substances seized from a defendant, a point or interval estimate of Q is required to help the court in calculating the sentence imposed upon the defendant. In situations where seizures consist of several thousand bags of illicit drugs, many jurisdictions recognize that sampling is a more practical procedure for determining the value of Q than carrying out a 100% inspection of the contraband, given the shrinking sizes of budgets. A few jurisdictions still resist this trend, basing their arguments upon the fact that drug quantity Q has to be proved beyond a reasonable doubt. Law enforcement agencies in general, however, have been most effective in promoting the view that sampling and statistical inference techniques ought to be enjoined as legitimate weapons in the war on drugs.

This article has described the often unpredictable nature of the substances obtained from a drug seizure. The possibility of nonillicit substances being included in individually packaged containers has made some courts wary about blindly accepting the results from sampling, especially when homogeneity of the substances has not been conclusively established. As a result, those courts often fail to take advantage of major gains in terms of time, cost and effort from scientific sampling of the evidence. Even under the reasonable doubt standard, many courts in nonfederal jurisdictions have affirmed sentences based upon statistical estimates of Q from two-stage sampling of multiple containers.

Consequently, the courts should be made more aware of the importance and advantages of randomization in producing truly random samples as opposed to the convenience or judgment samples that still unwittingly pass as good sampling practice by many law enforcement agencies and crime laboratories. Nonprobability samples lead to problems of selection bias, which, in turn, may result in a sample that is not "representative" of the population being sampled and for which the underlying probability model will be misleading and incorrect. Such deficiencies then render any subsequent statistical inferences from the sample worthless. Some courts have not yet acknowledged a difference between probability and nonprobability sampling procedures.

In cases where a sample of containers is selected for determining Q, we suggest that a careful explanation of the randomization procedure used should be given at trial, together with reasons why such randomization is a potent scientific principle. This type of presentation should help the jury to be convinced beyond a reasonable doubt that a statistical estimate of Q is scientifically accurate *in lieu* of 100% inspection and testing of all containers. On the other hand, whenever nonprobability sampling has been used, courts would be correct in following a more cautious approach and should rely only on the weighings of the contents of the tested containers, without projecting similar results to the untested containers.

Composite sampling is used in many forensic laboratories to investigate multiple containers for illicit drugs. Using a composite sampling scheme to decide how many containers would individually test positive for illicit drugs is an impossible statistical problem, however. Yet, we have shown that if carried out correctly, and assuming that inspection and testing of individual containers is more expensive than composite sampling, we can use that type of procedure to obtain an unbiased estimate (with standard error) of drug purity across all containers, and a modification can be used to assess the extent of homogeneity of substances in the different containers.

The question of sample size determination is also relevant when illicit drugs are conveyed in multiple containers. How many containers need to be sampled from the total number seized to provide a reliable estimate of Q when the possibility exists that the contents of some of the containers would not test positive for illicit drugs? It is surprising to discover that the square-root rule, which apparently has no statistical logic to commend it, is currently the most widely used sampling rule in drug cases. Fortunately, several crime laboratories have recently taken steps to employ more optimal sampling plans in consultation with statisticians rather than blindly following the square-root rule.

Finally, quantifying the evidence and deriving a (predictive) distribution for Q enables us to provide threshold values of Q that correspond to standards of proof needed to resolve differences in sentencing the defendant. It has also been shown that such threshold values are sensitive to the quality of the data available and any (Bayesian?) distributional assumptions made.

APPENDIX: PROOF OF (4.14)

The first step is based upon methodology developed by Brown and Fisher (1972), Rohde (1976) and Elder, Thompson and Myers (1980), who assumed that the random effects derived from an infinite population. See also the survey article by Boswell and Patil (1987).

Fix the rth subsample and consider the case of mixing c core samples from a single ($k = 1$) sample container (which temporarily allows us to drop the subscript j). Then (4.9) can be written in matrix notation as

$$\hat{p}_r = \sum_{i=1}^{c} W_{ri} P_i = \mathbf{W}_r^\tau \mathbf{P}, \tag{A.1}$$

where $\mathbf{W}_r = [W_{ri}]$ and $\mathbf{P} = [P_i]$ are both c-dimensional column vectors, and $\mathbf{A}^\tau$ is the transpose of the matrix $\mathbf{A}$. We will use (A.1) to model the variation in the purity of samples drawn from a single container. Suppose $\mathscr{E}(\mathbf{W}_r) = \mu_W$, $\operatorname{cov}(\mathbf{W}_r) = \mathbf{\Sigma}_W$ and $\mathscr{E}(\mathbf{P}) = \mu_P$, $\operatorname{cov}(\mathbf{P}) = \mathbf{\Sigma}_P$. Then

$$\mathscr{E}(\hat{p}_r) = \mu_W^\tau \mu_P, \tag{A.2}$$

$$\operatorname{var}(\hat{p}_r) = \operatorname{tr}(\mathbf{\Sigma}_W \mathbf{\Sigma}_P) + \mu_W^\tau \mathbf{\Sigma}_P \mu_W + \mu_P^\tau \mathbf{\Sigma}_W \mu_P, \tag{A.3}$$

$$\operatorname{cov}(\hat{p}_r, \hat{p}_s) = \operatorname{tr}(\mathbf{\Sigma}_{WW} \mathbf{\Sigma}_P) + \mu_W^\tau \mathbf{\Sigma}_P \mu_W + \mu_P^\tau \mathbf{\Sigma}_{WW} \mu_P, \tag{A.4}$$

where $\mathbf{\Sigma}_{WW}$ is the covariance matrix of $\mathbf{W}_r$ and $\mathbf{W}_s$.

In the following, we denote by $\mathbf{1}$ the c-dimensional column vector whose elements are each unity, by $\mathbf{I}_c$ the $(c \times c)$ identity matrix, and by $\mathbf{J}_c = \mathbf{1}\mathbf{1}^\tau$ the $(c \times c)$ matrix all of whose entries are unity. Then $\mathbf{1}^\tau \mathbf{W}_r = \sum_i W_{ri} = 1$. Taking the expected value, we have that $\mathscr{E}(\mathbf{1}^\tau \mathbf{W}_r) = \mathbf{1}^\tau \mu_W = c\mu_W = 1$, whence $\mu_W = 1/c$. Furthermore, it is reasonable to assume that the covariance matrix of $\mathbf{W}_r$ can be written in intraclass form as $\mathbf{\Sigma}_W = \sigma_W^2[(1-\rho)\mathbf{I}_c + \rho \mathbf{J}_c]$, where ρ is the intraclass correlation coefficient, so that $\mathbf{\Sigma}_W$ has diagonal entries σ_W^2 and off-diagonal entries $\rho\sigma_W^2$.

Because $\mathrm{var}(\mathbf{1}^\tau \mathbf{W}_r) = \mathbf{1}^\tau \boldsymbol{\Sigma}_W \mathbf{1} = c(1+(c-1)\rho)\sigma_W^2 = 0$, we have $\rho = -1/(c-1)$. Thus the $(c\times 1)$ mean vector and $(c \times c)$ covariance matrix of $\mathbf{W}_r$ are

$$\mu_W = \frac{1}{c}\mathbf{1}, \quad \boldsymbol{\Sigma}_W = \frac{\sigma_W^2}{c-1}(c\mathbf{I}_c - \mathbf{J}_c), \tag{A.5}$$

respectively. Thus (A.5) indicates that the subsampling procedure is unbiased in that we expect the subsamples to consist of equal proportions of substances from each core sample that makes up a given composite sample.

Now, suppose that an equal number c of core samples from each of k sample containers are mixed to form a single composite sample (while still keeping r fixed). This balanced case would be appropriate when the containers appeared to be homogeneous and contained about equal amounts of substance. Let $\mathbf{W}_r = [W_{rij}]$ be the $(c \times k)$ matrix of random weights whose k columns $\{\mathbf{W}_{ri}\}$ are assumed to be independent and identically distributed random vectors, and let $\mathbf{P} = [P_{ij}]$ be the $(c \times k)$ matrix of random purity values. Then (4.9) can be written as

$$\hat{p}_r = (\mathrm{vec}(\mathbf{W}_r))^\tau \mathrm{vec}(\mathbf{P}), \tag{A.6}$$

where $\mathrm{vec}(\mathbf{A})$ denotes the $(mn \times 1)$ column vector formed by placing the columns of an $(m \times n)$ matrix $\mathbf{A}$ under one another successively. The $(kc\times 1)$ mean vector and $(kc \times kc)$ covariance matrix of $\mathrm{vec}(\mathbf{W}_r)$ are, therefore, given by

$$\mathbf{m}_W = \mathscr{E}(\mathrm{vec}(\mathbf{W}_r)) = \frac{1}{kc}\mathbf{1}, \tag{A.7}$$

$$\mathbf{C}_W = \mathrm{cov}(\mathrm{vec}(\mathbf{W}_r)) = \boldsymbol{\Sigma}_W \otimes \mathbf{I}_k, \tag{A.8}$$

respectively, where the Kronecker product of the $(m\times n)$ matrix $\mathbf{A}$ and the $(p\times q)$ matrix $\mathbf{B}$ is defined here as the $(mp \times nq)$ block matrix $\mathbf{A} \otimes \mathbf{B} = [\mathbf{A}b_{jk}]$, and $\boldsymbol{\Sigma}_W$ is given by (A.5).

The $(kc\times 1)$ mean vector and $(kc\times kc)$ covariance matrix of $\mathrm{vec}(\mathbf{P})$ are given by

$$\mathbf{m}_P = \mathscr{E}(\mathrm{vec}(\mathbf{P})) = p\mathbf{1}, \tag{A.9}$$

$$\mathbf{C}_P = \mathrm{cov}(\mathrm{vec}(\mathbf{P})) = \mathbf{V} \otimes \mathbf{I}_k, \tag{A.10}$$

respectively, where $\mathbf{V} = \sigma_e^2\mathbf{I}_c + \sigma_t^2\mathbf{J}_c$. In other words, the purities of core samples from the same container are correlated, while those from different containers are uncorrelated.

Rewriting (A.3) using (A.7), (A.8), (A.9) and (A.10), we have that the variance of $\hat{p}_r$ over the k sample containers is given by

$$\begin{aligned}\mathrm{var}(\hat{p}_r) &= \mathrm{tr}(\mathbf{C}_W\mathbf{C}_P) + \mathbf{m}_W^\tau \mathbf{C}_P \mathbf{m}_W \\ &\quad + \mathbf{m}_P^\tau \mathbf{C}_W \mathbf{m}_P^\tau,\end{aligned} \tag{A.11}$$

where

$$\begin{aligned}\mathrm{tr}(\mathbf{C}_W\mathbf{C}_P) &= \mathrm{tr}(\boldsymbol{\Sigma}_W\mathbf{V} \otimes \mathbf{I}_k) \\ &= \mathrm{tr}(\boldsymbol{\Sigma}_W\mathbf{V})\mathrm{tr}(\mathbf{I}_k) = k\,\mathrm{tr}(\boldsymbol{\Sigma}_W\mathbf{V}) \\ &= k\,\mathrm{tr}\left(\frac{\sigma_W^2\sigma_e^2}{kc-1}(kc\mathbf{I}_c - \mathbf{J}_c)\right. \\ &\quad \left. + \sigma_W^2\sigma_t^2\left(\frac{c(k-1)}{kc-1}\right)\mathbf{J}_c\right) \\ &= kc\sigma_W^2\sigma_e^2 + \frac{kc^2(k-1)}{kc-1}\sigma_W^2\sigma_t^2,\end{aligned} \tag{A.12}$$

$$\begin{aligned}\mathbf{m}_W^\tau \mathbf{C}_P \mathbf{m}_W &= \frac{1}{(kc)^2}\mathbf{1}^\tau(\mathbf{V}\otimes\mathbf{I}_k)\mathbf{1} \\ &= \frac{\sigma_e^2}{kc} + \frac{\sigma_t^2}{k},\end{aligned} \tag{A.13}$$

$$\mathbf{m}_P^\tau \mathbf{C}_W \mathbf{m}_P^\tau = p^2\mathbf{1}^\tau(\boldsymbol{\Sigma}_W \otimes \mathbf{I}_k)\mathbf{1} = 0. \tag{A.14}$$

Substituting (A.12), (A.13) and (A.14) into (A.11), we get that

$$\begin{aligned}\mathrm{var}(\hat{p}_r) &= kc\sigma_W^2\sigma_e^2 + \frac{kc^2(k-1)}{kc-1}\sigma_W^2\sigma_t^2 \\ &\quad + \frac{\sigma_e^2}{kc} + \frac{\sigma_t^2}{k}.\end{aligned} \tag{A.15}$$

Furthermore, from (A.4) and (A.14),

$$\begin{aligned}\mathrm{cov}(\hat{p}_r, \hat{p}_s) &= \mathbf{m}_W^\tau \mathbf{C}_P \mathbf{m}_W \\ &= \frac{\sigma_e^2}{kc} + \frac{\sigma_t^2}{k}.\end{aligned} \tag{A.16}$$

Thus, the $\{\hat{p}_r\}$ have equal variances and all pairs $\{(\hat{p}_r, \hat{p}_s), r \neq s\}$ have equal correlations.

In the second step, we convert the infinite population results (A.15) and (A.16) to the corresponding results in which the k containers are assumed to have been sampled at random without replacement from the finite population of K containers. If each of the containers carried the same number N of discrete, uniquely identifiable items (e.g., pills, tablets, capsules, rocks of crack cocaine, squares of blotter paper soaked with LSD), Searle and Fawcett's (1970) rule would replace the variance component σ_t^2 in the above formulas by $\sigma_t^2 - \sigma_e^2/N_e$, where $N_e = N$ is the size of the population of error terms, to change to a finite population model. Comparable results for the case where N_j items are found in the jth container, $j = 1, 2, \ldots, k$, are currently not available. Because of the enormous amount of tedious algebra needed for the general case, use of a symbolic algebra computation program might be appropriate.

If the contents of the containers are instead loose cocaine or heroin powder (the scenario of most interest here), and if the sampling proportion for each container does not exceed 5% (or possibly even 10%) of the entire amount found in that container (Cochran, 1977, page 25), an assumption which is entirely reasonable given the typical practice of forensic laboratories, then we take N_e to be infinite, in which case no changes to the formulas (A.15) and (A.16) would be necessary.

Our estimate of purity over all K containers is $\bar{p}$ in (4.11). If $\hat{p}_s = (\mathrm{vec}(\mathbf{W}_s))^\tau \mathrm{vec}(\mathbf{P})$, where $\mathbf{W}_s$ is another matrix of random weights, then, using (A.15) and (A.16),

$$\begin{aligned}\mathrm{var}(\bar{p}) &= \frac{1}{R^2}\left(\sum_{r=1}^{R}\mathrm{var}(\hat{p}_r) + \sum_{\substack{r=1\\ r\neq s}}^{R}\sum_{s=1}^{R}\mathrm{cov}(\hat{p}_r, \hat{p}_s)\right) \\ &= \frac{kc}{R}\sigma_W^2\sigma_e^2 + \frac{kc^2(k-1)}{R(kc-1)}\sigma_W^2\sigma_t^2 + \frac{\sigma_e^2}{kc} + \frac{\sigma_t^2}{k},\end{aligned} \tag{A.17}$$

which proves (4.12).

For the case in which different numbers of core samples are extracted from the different sample containers (possibly because the sample containers hold unequal amounts of substance), the derivations are more complicated because we cannot use Kronecker products and vec notation as in the equal c case above. We briefly indicate the extensions to this case, leaving the reader to fill in the details. Suppose c_j core samples are taken from the jth sample container, $j = 1, 2, \ldots, k$. Let $C = \sum_{j=1}^{k} c_j$. The mean vectors (A.7) and (A.9) become $\mathbf{m}_W = C^{-1}\mathbf{1}$ and $\mathbf{m}_P = p\mathbf{1}$, respectively, while the covariance matrices (A.8) and (A.10) become $\mathbf{C}_W = \mathrm{diag}\{\mathbf{\Sigma}_{W_j}\}$ and $\mathbf{C}_P = \mathrm{diag}\{\mathbf{V}_j\}$, respectively, where $\mathbf{\Sigma}_{W_j} = \sigma_W^2(C\mathbf{I}_{c_j} - \mathbf{J}_{c_j})/(C-1)$ and $\mathbf{V}_j = \sigma_e^2\mathbf{I}_{c_j} + \sigma_t^2\mathbf{J}_{c_j}$. Then, (A.15) and (A.16) become

$$\begin{aligned}\mathrm{var}(\hat{p}_r) &= C\sigma_W^2\sigma_e^2 + \left(\frac{C^2 - \sum_{j=1}^{k} c_j^2}{C-1}\right)\sigma_W^2\sigma_t^2 \\ &\quad + \frac{\sigma_e^2}{C} + \left(\sum_{j=1}^{k}\frac{c_j^2}{C^2}\right)\sigma_t^2,\end{aligned} \tag{A.18}$$

$$\mathrm{cov}(\hat{p}_r, \hat{p}_s) = \frac{\sigma_e^2}{C} + \left(\sum_{j=1}^{k}\frac{c_j^2}{C^2}\right)\sigma_t^2, \tag{A.19}$$

respectively, and (A.17) becomes

$$\begin{aligned}\mathrm{var}(\bar{p}) &= \frac{C}{R}\sigma_W^2\sigma_e^2 + \left(\frac{C^2 - \sum_{j=1}^{k} c_j^2}{R(C-1)}\right)\sigma_W^2\sigma_t^2 \\ &\quad + \frac{\sigma_e^2}{C} + \left(\sum_{j=1}^{k}\frac{c_j^2}{C^2}\right)\sigma_t^2.\end{aligned} \tag{A.20}$$

ACKNOWLEDGMENTS

The author thanks A. S. Hedayat for introducing him to this topic, Betty-Ann Soiefer Izenman, a former Assistant U. S. Attorney, for help with legal questions, Giljam Derksen, Joseph L. Gastwirth, Burt S. Holland, Karen Kafadar, David H. Kaye, G. P. Patil, D. Raghavarao, Peter Tillers, Sheila Willis and others for invaluable comments, suggestions, references, correspondence and encouragement, and Deputy Inspector Phil Pulaski, Commanding Officer of the NYPD Forensic Investigations Division, and Lt. William Schmidt and Richie Gaceta, NYPD Police Laboratory, for helpful information on crime laboratory practice. The author also thanks three anonymous referees, an Editor and Leon Jay Gleser for constructive comments and suggestions on this manuscript. An earlier version of this paper was presented by the author at the Third International Conference on Forensic Statistics held at the University of Edinburgh, Scotland, June 30–July 3, 1996.

REFERENCES

AITKEN, C. G. G. (1995). *Statistics and the Evaluation of Evidence for Forensic Scientists.* Wiley, New York.

AITKEN, C. G. G. (1999). Sampling—how big a sample? *J. Forensic Sciences* **44** 750–760.

AITKEN, C. G. G. (2000). Sample size determination and evidence interpretation. In *Statistical Science in the Courtroom* (J. Gastwirth, ed.) Springer, New York.

AITKEN, C. G. G., BRING, J., LEONARD, T. and PAPASOULIOTIS, O. (1997). Estimation of quantities of drugs handled and the burden of proof. *J. Roy. Statist. Soc. Ser. A* **160** 333–350.

BEAUPRE, B. and EISLER, P. (1996). Crisis in evidence. *USA Today* August 20, 1A, 2A, 7A; August 21 9A.

BELLHOUSE, D. (1988). A brief history of survey sampling. In *Handbook of Statistics* (P. R. Krishnaiah and C.R. Rao, eds.) **6** 1–15. North-Holland, Amsterdam.

BLANCK, F. C. (1927). Report of the Committee on Sampling. *J. Assoc. Official Agricultural Chemists* **10** 92–98.

BOLAND, B. and HEALY, K. M. (1993). *Prosecutorial Response to Heavy Drug Caseloads: Comprehensive Problem-Reduction Strategies*. U.S. Department of Justice, Washington, DC.

BOSWELL, M. T., GORE, S. D., LOVISON, G. and PATIL, G.P. (1996). Annotated bibliography of composite sampling. Part A. 1936–92. *Environmental and Ecological Statistics* **3** 1–50.

BOSWELL, M. T. and PATIL, G. P. (1987). A perspective of composite sampling. *Comm. Statist. Theory Methods* **16** 3069–3093.

BRING, J. and AITKEN, C. G. G. (1997). Burden of proof and estimation of drug quantities under the Federal Sentencing Guidelines. *Cardozo Law Review* **18** 1987–1999.

BRITISH STANDARD BS 5309 (1993). Methods for the sampling of chemical products. Part I. Introduction and general principles.

BROOKMEYER, R.and GAIL, M. H. (1994). *AIDS Epidemiology: A Quantitative Approach*. Oxford Univ. Press.

BROWN, G. H. and FISHER, N. I. (1972). Subsampling a mixture of sampled material. *Technometrics* **14** 663–668.

BUREAU OF JUSTICE ASSISTANCE (1996). *National Assessment of Structured Sentencing*, NCJ-153853. U.S. Department of Justice, Washington, DC.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

COLON, M., RODRIGUEZ, G. and DIAZ, R. O. (1993). Representative sampling of "street" drug exhibits. *J. Forensic Sciences* **38** 641–648.

COMMITTEE ON SAMPLING (1925). Report of the Committee on Sampling. *J. Assoc. Official Agricultural Chemists* **8** 287–290.

DEGROOT, M. H., FIENBERG, S. E. and KADANE, J. B., eds. (1986). *Statistics and the Law*. Wiley, New York. [Also reprinted (1994) as a Wiley Classics Library Edition.]

DEMING, W. E. (1954). On the presentation of the results of sample surveys as legal evidence. *J. Amer. Statist. Assoc.* **49** 814–825.

DESU, M. M. and RAGHAVARAO, D. (1990). *Sample Size Methodology*. Academic Press, New York.

DEVITT, E. J., BLACKMAR, C. B. and WOLFF, M. A. (1987). *Federal Jury Practice and Instructions, Civil,* **3**, 4th ed. West, St. Paul, MN.

DIAMOND, S. (1994). Reference guide on survey research. In *Reference Manual on Scientific Evidence* 221–271. Federal Judicial Center. U.S. Government Printing Office, Washington, DC.

DUNCAN, A. J. (1962). Bulk sampling: problems and lines of attack. *Technometrics* **4** 319–344.

ELDER, R. S., THOMPSON, W. O. and MYERS, R. H. (1980). Properties of composite sampling procedures. *Technometrics* **22** 179–186.

FEDERAL JUDICIAL CENTER (1994). *Reference Manual on Scientific Evidence*. U.S. Government Printing Office, Washington, DC.

FEDERER, W. T. (1984). Cutting edges in biometry. *Biometrics* **40** 827–839.

FIENBERG, S. E., ed. (1989). *The Evolving Role of Statistical Assessments as Evidence in the Courts*. Springer, New York.

FINKELSTEIN, M. O. and LEVIN, R. (1990). *Statistics for Lawyers*. Springer, New York.

FOSTER, E. M. (1971). The control of Salmonellae in processed foods: classification system and sampling plan. *J. Assoc. Official Analytical Chemists* **54** 259–266.

FRANK, R. S., HINKLEY, S. W. and HOFFMAN, C. G. (1991). Representative sampling of drug seizures in multiple containers. *J. Forensic Sciences* **36** 350–357.

GASTWIRTH, J. L. (1987). The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data. *Statist. Sci.* **2** 213–238.

GASTWIRTH, J. L. (1988). *Statistical Reasoning in Law and Public Policy:* (**1**) *Statistical Concepts and Issues of Fairness;* (**2**) *Tort Law, Evidence, and Health*. Academic Press, New York.

GASTWIRTH, J. L. (1992). Statistical reasoning in the legal setting. *Amer. Statist.* **46** 55–69.

GASTWIRTH, J. L., ed. (2000). *Statistical Science in the Courtroom*. Springer, New York.

GASTWIRTH, J. L. and HAMMICK, P. A. (1989). Estimation of prevalence of a rare disease, preserving anonymity of subjects by group testing: application to estimating the prevalence of AIDS antibodies in blood donors. *J. Statist. Plann. Inference* **22** 15–27.

GAYLOR, D. W. and HARTWELL, T. D. (1969). Expected mean squares for nested classifications. *Biometrics* **25** 427–430.

GONYER, T. E. (1998). Federal Sentencing in a post-*Chapman* world: what is a "mixture or substance." *Kansas Law Review* **46** 983–1011.

GREENWALD, L. (1994). Relevant conduct and the impact of the preponderance standard of proof under the Federal Sentencing Guidelines: a denial of due process. *Vermont Law Review* **18** 529–563.

HAMMICK, P. A. and GASTWIRTH, J. L. (1994). Group testing for sensitive characteristics: extension to higher prevalence levels. *Internat. Statist. Rev.* **62** 319–331.

HANSEN, M. H. (1987). Some history and reminiscences on survey sampling. *Statist. Sci.* **2** 180–190.

HEDAYAT, A. S., IZENMAN, A. J. and ZHANG, W. G. (1996). Random sampling for the forensic study of controlled substances (with discussion). In *Proceedings of the Physical and Engineering Sciences Section* 12–23. Amer. Statist. Assoc., Alexandria, VA.

HUGHES-OLIVER, J. M.and SWALLOW, W. H. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *J. Amer. Statist. Assoc.* **89** 982–993.

HUYNH, H. and FELDT, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. *J. Amer. Statist. Assoc.* **65** 1582–1589.

IZENMAN, A. J. (2000a). Statistical issues in the application of the Federal Sentencing Guidelines in drug, pornography, and fraud cases. In *Statistical Science in the Courtroom* (J. Gastwirth, ed.) 25–50. Springer, New York.

IZENMAN, A.J. (2000b). Introduction to two views on the *Shonubi* case. In *Statistical Science in the Courtroom* (J. Gastwirth, ed.) 393–403. Springer, New York.

IZENMAN, A. J. (2000c). Assessing the statistical evidence in the *Shonubi* case. In *Statistical Science in the Courtroom* (J. Gastwirth, ed.) 415–443. Springer, New York.

IZENMAN, A. J., PAPASOULIOTIS, O., LEONARD, T. and AITKEN, C. G. G. (1998). Bayesian predictive evaluation of measurement error with application to the assessment of illicit drug quantity. Unpublished manuscript.

KAYE, D. H. and AICKIN, M., eds. (1986). *Statistical Methods in Discrimination Litigation*. Dekker, New York.

KRUSKAL, W. and MOSTELLER, F. (1979). Representative sampling. I. Nonscientific literature. II. Scientific literature, excluding statistics. III. The current statistical literature. *Internat. Statist. Rev.* **47** 13–24, 111–127, 245–265.

LANCASTER, V. A. and KELLER-MCNULTY, S. (1998). A review of composite sampling methods. *J. Amer. Statist. Assoc.* **93** 1216–1230.

LARSEN, K. H. (1964). Admissibility in evidence of sample or samples of article or substance of which the quality, condition, or the like is involved in litigation. *ALR2d* **95** 681–712. [See also Later Case Service Supplementing **93–95** *ALR2d* (1993) 366–368.]

LE, S. D., TAYLOR, R. W., VIDAL, D., LOVAS, J. J. and TING, E. (1992). Occupational exposure to cocaine involving crime lab personnel. *J. Forensic Sciences* **37** 959–968.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

LIMAN, L. J. (2000). Initial thoughts on *Apprendi v. New Jersey*. *New York Law Journal* July 5 3.

MAURO, T. and RINGEL, J. (2000). In striking N.J. hate crimes law, Court may trigger sentencing review. *New Jersey Law Journal* July 3 1.

MUNCH, J. C. and BIDWELL, G. L. (1928). What constitutes an adequate sample? *J. Assoc. Official Agricultural Chemists* **11** 220–222.

PAUL, A. E. (1927). Sampling drugs. *J. Assoc. Official Agricultural Chemists* **10** 99–102.

PENNSYLVANIA BAR INSTITUTE (1981). *Pennsylvania Suggested Standard Civil Jury Instructions*.

QUACKENBUSH, F. W. and RUND, R. C. (1967). The continuing problem of sampling. *J. Assoc. Official Analytical Chemists* **50** 997–1006.

ROHDE, C. A. (1976). Composite sampling. *Biometrics* **32** 273–282.

RUNKEL, H. (1926). Report on methods for sampling flour. *J. Assoc. Official Agricultural Chemists* **9** 423–426.
SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.
SCHUSTER, J. J. (1991). The statistician in a reverse cocaine sting. *Amer. Statist.* **45** 123–124.
SEARLE, S. R. and FAWCETT, R. F. (1970). Expected mean squares in variance components models having finite populations. *Biometrics* **26** 243–254.
STEPHAN, F. F. (1948). History of the uses of modern sampling procedures. *J. Amer. Statist. Assoc.* **43** 12–39.
STRONG, J. W., ed. (1992). *McCormick on Evidence* **1**, 4 ed. West, St. Paul, MN.
TZIDONY, D. and RAVREBY, M. (1992). A statistical approach to drug sampling: a case study. *J. Forensic Sciences* **37** 1541–1549.
UNITED NATIONS (1987). *Recommended Methods for Testing Cannabis: Manual for Use by National Narcotics Laboratories*. United Nations, Division of Narcotic Drugs, Vienna.
UNITED STATES SENTENCING COMMISSION (1991). *Special Report to the Congress: Mandatory Minimum Penalties in the Federal Criminal Justice System*. U.S. Government Printing Office, Washington, DC.
UNITED STATES SENTENCING COMMISSION (1995). *Federal Sentencing Guidelines Manual*. West, St. Paul, MN.
WALKER, L. and MONAHAN, J. (1998). Sampling damages. *Iowa Law Review* **83** 545–568.
WALKER, L. and MONAHAN, J. (1999). Sampling liability. *Virginia Law Review* **85** 329–351.
WALKER, V. R. (1994). Direct inference in the last chance cases: factfinding constraints under minimal fairness to parties. *Hofstra Law Review* **23** 247–307.
WETHERILL, G. B. and BROWN, D. W. (1991). *Statistical Process Control: Theory and Practice.* Chapman and Hall, New York.
WIGMORE, J. H. (1979). *Evidence in Trials at Common Law* **2**. Little, Brown, Boston.
YOUNG, D. (1990). Rethinking the Commission's drug guidelines: courier cases where quantity overstates culpability. *Federal Sentencing Reporter* **3** 63–66.
ZEISEL, H. and KAYE, D. (1997). *Prove It with Figures: Empirical Methods in Law and Litigation*. Springer, New York.

CASES CITED

Almendarez-Torres v. United States (1998). 523 U.S. 224, 140 L.Ed.2d 350, 118 S.Ct. 1219.
Apprendi v. New Jersey (2000). 147 L.Ed.2d 435, 120 S.Ct. 2348, 2000 U.S. LEXIS 4304, 68 U.S.L.W. 4576.
Asmer v. State of Florida (1982). 416 So.2d 485 (Fla.App. 4 Dist.).
Bond v. State of Florida (1989). 538 So.2d 499 (Fla.App. 3 Dist.).
Carter v. State of Florida (1994). 646 So.2d 853 (Fla.App. 3 Dist.).
Chapman v. United States (1991). 501 U.S. 1270, 111 S.Ct. 1919.
Gabriel v. State of Texas (1995). 900 S.W.2d 721 (Tex.Cr.App.).
In re Lemons (1991). 77 Ohio.App.3d 691, 603 N.E.2d 315 (Ohio App. 8 Dist.).
Mello v. State of Texas (1991). 806 S.W.2d 875 (Tex.App.-Eastland).
Neal v. United States (1996). 516 U.S. 284, 116 S.Ct. 763, 133 L.Ed.2d 709.
People v. Argro (1975). 37 N.Y.2d 929, 379 N.Y.S.2d 840, 342 N.E.2d 601.
People v. Ayala (1981). 96 Ill.App.3d 785, 422 N.E.2d 127.
People v. Black (1994). 264 Ill.App.3d 875, 638 N.E.2d 233 (Ill.App. 1 Dist.).
People v. Games (1981). 94 Ill.App.3d 130, 418 N.E.2d 520.
People v. Hering (1975). 27 Ill.App.3d 936, 327 N.E.2d 583.
People v. Hill (1988). 169 Ill.App.3d 901, 524 N.E.2d 604, *appeal denied*, 122 Ill.2d 585, 530 N.E.2d 256.
People v. Hill (1993). 199 A.D.2d 111, 605 N.Y.S.2d 734.
People v. Jackson (1985). 134 Ill.App.3d 785, 481 N.E.2d 1222.
People v. Jones (1995). 276 Ill.App.3d 926, 659 N.E.2d 70 (Ill.App. 1 Dist.), *aff'd*, 174 Ill.2d 427, 221 Ill.Dec. 192, 675 N.E.2d 99 (Ill. 1996).
People v. Kaludis (1986). 146 Ill.App.3d 888, 497 N.E.2d 360 (Ill.App. 1 Dist.), *appeal denied*, 113 Ill.2d 564.
People v. Kline (1976). 41 Ill.App.3d 261, 354 N.E.2d 46.
People v. Little (1986). 140 Ill.App.3d 682, 489 N.E.2d 322.
People v. McLaurin (1993). 157 Misc.2d 783, 598 N.Y.S.2d 911.
People v. Peneda (1995). 32 C.A.4th 1022, 38 Cal.Rptr.2d 312 (Cal.App. 4 Dist.).
People v. Rodriguez (1993). 190 A.D.2d 527.
People v. Saldana (1986). 146 Ill.App.3d 328, 496 N.E.2d 757 (Ill.App. 2 Dist.).
People v. Yosell (1977). 53 Ill.App.3d 289, 368 N.E.2d 735.
Pugh v. State of Georgia (1995). 216 Ga.App. 253, 454 S.E.2d 538 (Ga.App.).
Ross v. State of Florida (1988). 528 So.2d 1237 (Fla.App. 3 Dist.); *review denied*, 537 So.2d 569 (Fla. 1988).
State of Florida v. Meeks (1989). 552 So.2d 328 (Fla.App. 3 Dist.).
State of Louisiana v. Ballom (1990). 562 So.2d 1073 (La.App. 4th Cir.), *writ denied*, 575 So.2d 386 (La. 1991).
State of Minnesota v. Robinson (1994). 517 N.W.2d 336 (Minn.).
State of Missouri v. Edwards (1975). 521 S.W.2d 474 (Mo.App.)
State of Missouri v. Gibson (1993). 856 S.W.2d 78 (Mo.App. E.D.).
State of North Carolina v. Clark (1973). 18 N.C.App. 473, 197 S.E.2d 81 (N.C.App.).
State of North Carolina v. Harding (1993). 110 N.C.App. 155, 429 S.E.2d 416 (N.C.App.).
State of North Carolina v. Hayes (1976). 291 N.C. 293, 230 S.E.2d 146 (N.C.).
United States v. Aguayo-Delgado (2000). 220 F.3d 926, 2000 U.S.App. LEXIS 17243 (8th Cir.).
United States v. Chapple (1993). 985 F.2d 729 (3d Cir.).
United States v. Fatico (1978). 458 F.Supp. 388 (E.D.N.Y.), *aff'd on other grounds*, 603 F.2d 1053 (2d. Cir. 1979), *cert. denied*, 444 U.S. 1073, 100 S.Ct. 1018, 62 L.Ed.2d 755 (1980).
United States v. Hilton (1990). 894 F.2d 485 (1st Cir.).
United States v. McCutchen (1993). 992 F.2d 22 (3rd Cir.).
United States v. Pirre (1991). 927 F.2d 694 (2nd Cir.).
United States v. Scalia (1993). 993 F.2d 984 (1st Cir.).
United States v. Sheppard (2000). 219 F.3d 766, 2000 U.S.App. LEXIS 17242 (8th Cir.).
United States v. Shonubi (1995). 895 F.Supp. 460 (E.D.N.Y.).
United States v. Whiting (1994). 28 F.3d 1296 (1st Cir.).
Wright v. State of Florida (1977). 351 So.2d 1127 (Fla.App. 1 Dist.).