J Forensic Sci, Sept. 2002, Vol. 47, No. 5
Paper ID JFS2001437_475
Available online at: www.astm.org

Colin G.G. Aitken,[1] *Ph.D. and David Lucy,*[1] *Ph.D.*

Estimation of the Quantity of a Drug in a Consignment from Measurements on a Sample

ABSTRACT: A consignment of individual packages is thought to contain illegal material, such as drugs, in some or all of the packages. A sample from the consignment is inspected and the quantity of drugs in each package of the sample is measured. It is desired to estimate the total quantity of drugs in the consignment. Sampling variation is present in the original measurements and it is not sufficient just to adjust the sample mean pro rata. An analysis is described which takes account of the uncertainty concerning the proportion of the packages that contain drugs and provides a probabilistic summary of the quantity of drugs in the consignment. In particular, a probabilistic lower bound for the quantity of drugs in the consignment is given, which is dependent on the required standard of proof. This is in contrast to the approach based on confidence intervals which assumes that in the long run, the interval will contain the correct quantity the appropriate proportion of the time, but gives no measure of uncertainty associated with the particular consignment under consideration.

KEYWORDS: forensic science, drugs, statistics, Bayesian inference, quantity estimation

According to procedures described in Tzidony and Ravreby (1), street doses of heroin in Israel are frequently packaged in small pieces of folded paper, which are further wrapped in plastic, which is heat-sealed, making it time-consuming to open and weigh the enclosed powder. Procedures are described in (1), and repeated here, for choosing a sample size from a consignment of doses and, following examination of the contents of the sample for estimating the total quantity of drugs in the consignment.

As reported by Frank et al. (2), various methods for selecting the size of a random sample from a consignment have been accepted by the U.S. courts. A summary of different procedures for sample size determination has been described by Colón et al. (3). If the consignment size is denoted by N, then some of these procedures can be described by simple formulations such as $\sqrt{N}$, $10\%N$, $4\%N$ and by methods based on the hypergeometric distribution, amongst others. These methods differ fundamentally from alternative methods described by Aitken (4) for choosing the sample size based on the incorporation of prior beliefs with certain probabilistic criteria. These criteria do not depend on the concept of frequentist ideas using confidence intervals.

In an earlier paper (4), in which consignments of discrete units such as tablets, compact disks or computer disks, were considered, the Bayesian and frequentist approaches for the estimation of sample size were compared. The units considered take one of two values, such as "illicit" or "not illicit." For example, this may be whether a pill does or does not contain an illicit drug. It is the purpose of this paper to describe a Bayesian method for the estimation of a quantity of a drug in a consignment when, for example, there may be interest in the total weight of illicit material in the consignment. This paper will compare the two approaches for the estimation of the quantity of drugs in consignments, both large and small. The distinction between large and small consignments cannot be clearly defined. For the purposes of this paper, a large consignment will be one whose size is greater than 50. The Bayesian approach permits the inclusion of uncertainty for both proportions and quantities in the analysis. This is in contrast to the frequentist approach in which only uncertainty in the estimation of the quantity is considered and a point estimate is used for the proportion.

The estimation of the quantity of a drug will be treated in two stages. First, the proportion of the units in the consignment that contain illicit drugs will be modelled. Secondly, the total weight of the illicit material in those packets that do contain anything illicit is estimated. Following previous work (4), uncertainty in the prior belief in the proportion of packets that are "illicit" is represented by a beta distribution. Beta distributions have useful mathematical properties, as well as having a form which models well the uncertainties about proportions. One can usually be chosen that fits adequately with almost any prior belief. It is assumed there is no prior information for the mean and variance of the distribution of the quantity of drugs in the packages. Details of how such prior information may be considered are given in (5) and are beyond the scope of this paper.

A recent review of statistical and legal aspects of the forensic study of illicit drugs is given by Izenman (6). This includes a discussion of various sampling procedures, various methods of choosing the sample size, a strategy for assessing homogeneity and the relationship between quantity and the possible standards of proof.

Izenman (6) gives several sampling procedures. First, for single containers, 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 multiple containers, without homogeneity, a rule is that at least one sample from each container should be conclusively tested for the presence of an illicit drug. Another procedure is that of composite sampling. In this procedure, a sample is taken from each source, the samples are then thoroughly mixed and a

[1] Department of Mathematics and Statistics, and Joseph Bell Centre for Forensic Statistics and Legal Reasoning, The King's Buildings, The University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, UK.

Received 28 Dec. 2001; and in revised form 9 April 2002; accepted 17 April 2002; published 21 Aug. 2002.

subsample is taken from the mixture. The mixture is the composite sample. Only simple random sampling is considered in this paper.

There are many choices of sample size, of which Izenman (6) gives several, such as the square-root law, the 4% rule and the 10% rule. The square-root rule apparently originated in the 1920's from a need to provide agricultural regulatory inspectors with a convenient, memorizable rule for sample size determination.

A strategy for assessing homogeneity is also described in Izenman (6). Consider a collection of containers. First, divide the containers into several batches, so that each batch has approximately the same number of containers. Next, take a random sample of the containers within each batch. From each of the sample containers in each batch, take some samples of its contents, mix these thoroughly to obtain a batch composite sample and then take at least two subsamples from each batch composite sample.

Given the sample size, and thus an estimate of the proportion of a consignment, which contains drugs and an estimate of the mean and standard deviation of the weight in the consignment, procedures have been provided for the construction of an estimate of the true quantity of drugs with a given level of confidence by Tzidony and Ravreboy (1). This interval is said to be a confidence interval to distinguish it from a probability interval. A confidence interval is defined by data (e.g., mean $\pm$ a multiple of a standard deviation) and is fixed by the data. There is no randomness in it and thus no probability can be attached to it. For a given sample size, the width of the confidence interval increases with the level of confidence. Thus, a 95% confidence interval is wider than a 90% confidence interval but shorter than a 99% confidence interval.

A probability interval is appropriate in a Bayesian context. In this context, a probability distribution is associated with a parameter (Q, say) denoting the total quantity of illicit material in the consignment. Thus it is possible to make probability statements of any desired kind. For example, these could include the probability that Q is greater than a certain value, q say, which will be of importance in sentencing hearings.

Recently, Coulson et al. (7) describe a procedure which combines a subjective prior assessment of the number of illicit tablets in a small consignment with a hypergeometric probability distribution.

Sampling Procedures

Before the quantity is estimated, a sample size has to be chosen. As well as the sample sizes based on proportions of the consignment size, such as those given above, other methods are based on the hypergeometric and binomial distributions.

The sampling problem is characterized as follows. A consignment of N items is divided into two subpopulations, one of positives (R) (say, illicit drugs) and one of negatives (N–R). A sample of size m is to be taken from the consignment and the number of positives z in the sample is noted. From these results an inference is then made about the number of items in the consignment which contain drugs. In an earlier paper Aitken (4) describes how m is chosen. The main purpose of this paper is to describe a procedure for the estimation of the quantity of drugs in the consignment. First, though, consider the hypergeometric and binomial distributions which are both used for sampling procedures.

Hypergeometric Distribution

This is the probability distribution which models the sampling without replacement from a finite population which consists of two subpopulations, one of positives (R) and one of negatives (N–R). The hypergeometric distribution considers the number of ways in which m items can be sampled from N, when the sampling is without replacement. This is

$$\binom{N}{m}$$

where $\binom{N}{m}$ is the binomial coefficient $\frac{N!}{m!(n-m)!}$ and $N!$ is the factorial coefficient $N \times (N-1) \times \ldots \times 2 \times 1$, and all the ways of sampling are equally likely. The number of ways in which z of the m may be positive and $m - z$ may be negative is

$$\binom{R}{z} \times \binom{N-R}{m-z},$$

and, again, these are all equally likely. Thus, the probability that, when sampling m items, z are positive, is

$$\frac{\binom{R}{z}\binom{N-R}{m-z}}{\binom{N}{m}}.$$

Let θ be the proportion of drug units in the population, and let θ_0 be a lower confidence bound for θ. The lower limit for θ is $R_0/N = \theta_0$ where R_0 is the maximum number R of illicit drug units in the population which satisfies the following inequality

$$\sum_{i=0}^{z} \frac{\binom{R}{i}\binom{N-R}{m-i}}{\binom{N}{m}} \leq \alpha.$$

See (1) for further details. Given m and z it is possible to determine R_0.

Binomial Distribution

If the total consignment size exceeds 50 the proportion θ of drug units in the consignment can be treated as if it were constant during the random sampling process and the binomial distribution may be used. Again, it is desired to show that θ is equal to or greater than a predetermined value θ_0. For a given sample size m and $z(\leq m)$ positives, θ_0 is the largest value of θ which satisfies the inequality

$$\sum_{i=0}^{m-z} \binom{m}{i} \theta^{m-i}(1-\theta)^i \leq \alpha.$$

Frequentist Approach

It is only possible to make a statement about the consignment as a whole with certainty if the whole consignment is analyzed. Once it is accepted that a sample has to be considered, then it is necessary to consider what level of proof is adequate. This is strictly a matter for the court to decide. According to (2), it should be sufficient to demonstrate with "good probability that most of the exhibit contains the controlled substance." Yet, summaries are given as confidence limits using a frequentist approach (an approach in which probabilities are estimated from relative frequencies determined from repetitions of experiments under hypothesised identical conditions) and not in probabilistic terms. For example, from (2), a statement of the form that "at the 95% confidence level, 90% or more of the packages in an exhibit contain the substance" is sug-

gested as being sufficient proof in cases of drug handling that 90% or more of the packages contain the substance. However, the probability with which this particular interval contains the true proportion is not known. Further comments are given in (4).

The method described by Tzidony and Ravreboy (1) considers the consignment as a population and the packages (or units) examined as a sample. The quantities (weights) of drugs in the units are assumed to be random variables which are normally distributed, with mean μ and variance σ^2, say. The mean quantity in a unit in the consignment is estimated by the mean, denoted $\bar{x}$, of the quantities found in the sample. A confidence interval is determined for μ based on the sample size *m,* the sample mean $\bar{x}$, the sample standard deviation *s* of the quantities of drugs in the units examined and an associated *t*-distribution. An estimate of the total quantity of drugs in the consignment is then determined by considering the size *N* of the consignment and the proportion θ of packages in the consignment thought to contain drugs. A confidence interval may then be constructed which may be said to contain the true quantity of drugs with a given level of confidence. The interval is a confidence interval, not a probability interval. Consider, for example a 95% confidence interval. The interpretation of this is that in 95% of occasions on which such an interval is constructed, it will contain the true value of the parameter of interest.

For example, the inequalities in expression (7) of Tzidony and Ravreboy (1) are, in the notation of the current paper,

$$\bar{x} - t_{(m-1)}(\alpha/2)\frac{s}{\sqrt{m}}\sqrt{\frac{(N-m)}{N}} \le \mu \le \bar{x} + t_{(m-1)}(\alpha/2)\frac{s}{\sqrt{m}}\sqrt{\frac{(N-m)}{N}},$$

where $t_{(m-1)}(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the *t*-distribution with $(m-1)$ degrees of freedom, $\sqrt{\frac{(N-m)}{N}}$ is the finite population correction factor and the interval is the $100(1-\alpha)\%$ confidence interval for the mean quantity in a package.

The corresponding confidence interval for *Q,* the total quantity of drugs in the consignment is obtained by multiplying all entries in the inequalities by $N\hat{\theta}$ where $\hat{\theta}$ is an estimate for θ based on the sample of size *m.* This gives as a $100(1-\alpha)\%$ confidence interval for *Q* (expression (9) of (1))

$$N\hat{\theta}\left\{\bar{x} - t_{(m-1)}(\alpha/2)\frac{s}{\sqrt{m}}\sqrt{\frac{(N-m)}{N}}\right\} \le Q \le N\hat{\theta}\left\{\bar{x} + t_{(m-1)}(\alpha/2)\frac{s}{\sqrt{m}}\sqrt{\frac{(N-m)}{N}}\right\}. \quad (1)$$

However, no account is taken of the uncertainty in the estimation of θ, only a point estimate of θ is used.

A corresponding $100(1-\alpha)\%$ lower bound for *Q* is given by the left-hand-side of the inequality.

$$N\hat{\theta}\left\{\bar{x} - t_{(m-1)}(\alpha)\frac{s}{\sqrt{m}}\sqrt{\frac{(N-m)}{N}}\right\} \le Q \quad (2)$$

where $t_{(m-1)}(\alpha)$ is the $100(1-\alpha)\%$ point of the *t*-distribution with $(m-1)$ degrees of freedom.

Bayesian Approach

Bayesian estimation can incorporate subjective information about a problem into the analysis. Objections are raised to the loss of objectivity that results from using the analyst's subjective information. However, either the data are strong enough for reasonable people to agree on their interpretation, regardless of the prior, or the analysts should be using their subjective prior information in order to make appropriate decisions related to the data. It is also possible to choose a so-called ignorance prior in which the subjective prior information is minimal.

Procedures described by Aitken (4) provide summaries for the determination of sample size in probabilistic terms. These probabilistic ideas are incorporated into the method for the estimation of quantities. Other work (5,8) describes procedures for the estimation of the quantity of drugs based on a Bayesian approach, an approach which permits the combination of so-called *subjective probabilities,* based on measures of belief, with probabilities derived from observations of random quantities assumed to follow specified probability distributions. The subjective probabilities may be derived in consultation with forensic scientists and lawyers to model background information about the consignment, such as its country of origin. More commonly, however, in the absence of such information or if there is an unwillingness to use subjective probabilities, it is possible to use what are known as vague or ignorance prior probabilities to represent a "neutral" position and this is done later. Note that, as mentioned in (4) it is not possible for the scientist's prior beliefs to have no effect on the analysis. For example, the choice of the model which is used to represent the uncertainty introduced by the sampling process is itself a subjective choice.

Lindley (9) provides the following argument in favor of subjectivity. Belief is a property of an individual and is subjective. A frequentist version of probability, what Lindley calls *chance,* is a property of a sequence and all who observe the sequence will observe its value. It is objective. Subjectivity is thought undesirable and difficult to handle. The difficulty has the following mitigation. Suppose two people have different beliefs in the truth of an event, say the proportion θ of drugs in a consignment. Their beliefs are not the same. Additional evidence *m* and *z* relevant to θ is produced. Then, as can be seen in the results for sample sizes, this additional evidence will tend to bring the two beliefs together.

The statistical methodology for informative priors is described in (5), the legal considerations in (8). The examples discussed in these two papers refer to court cases in which reasonably large quantities of drugs, of the order of several kilograms, have been involved. The methods described in this paper consider both large and small consignments in examples where units of the consignment may be divided into one or two groups only, those units which contain illicit drugs and those which do not. The two groups are assumed to be homogeneous within themselves.

The Bayesian approach considers the consignment, the population of (1), as itself a random sample from a larger super-population of units or packages, some or all of which contain illegal material. Then θ $(0 < \theta < 1)$ is the proportion of units in the super-population which contain drugs. In order to make probability statements about θ, it is necessary to represent the variability in θ with a probability distribution for θ. This variability may simply be uncertainty in one's knowledge of the exact value of θ, uncertainty which may arise because the consignment is considered as a random sample from a super-population. The Bayesian philosophy permits this uncertainty to be represented as a probability distribution. The most common distribution for θ is the so-called beta distribution, with probability density function

$$f(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}, 0 < y < 1,$$

where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha + \beta)},$$

and $\Gamma(x + 1) = x!$ for integer $x > 1$, $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{(\pi)}$. Its use in this context is described in (4).

Let n be the number of packages in the consignment which are not examined. Then N equals $m + n$. Let z ($\leq m$) be the number of units in those examined which contain drugs and let y ($\leq n$) be the number of units which contain drugs among those units which are not examined. Let $(x_1, \ldots, x_z)$ be measurements of the quantities of drugs in those units examined which contain drugs. Let $(w_1, \ldots, w_y)$ be measurements of the quantities of drugs in those units not examined which contain drugs. Let $\bar{x} = \Sigma_{i=1}^{z}\, x_i/z$ be the sample mean quantity of drugs in units containing drugs among those examined, and let s be the sample standard deviation where the sample variance $s^2 = \Sigma_{i=1}^{z}\,(x_i - \bar{x})^2/(z - 1)$. Let $\bar{w} = \Sigma_{i=1}^{y}\, w_j/y$ be the mean quantity of drugs in units containing drugs among those not examined. The total quantity q of drugs in the exhibit is then $(z\bar{x} + y\bar{w})$ and the problem is one of first estimating $\bar{w}$, given $\bar{x}$, s and z, while not knowing y and then of finding $y\bar{w}$ by finding the posterior distribution of $f(y \mid \bar{x})$. The method advocated by Tzidony and Ravreboy (1) is a so-called *estimative* approach (see Aitchison et al. (10)), in which the parameters (μ, σ^2) of the normal distribution representing the quantity of drugs in an individual unit are *estimated* by the corresponding sample mean $\bar{x}$ and sample variance s^2. The method described below is a *predictive* approach (10–12) in which the values of the unknown measurements $(w_1, \ldots, w_y)$ are *predicted* by values of known measurements $(x_1, \ldots, x_z)$. A brief, general, description for forensic scientists is given in Aitken (13).

The predictive approach *predicts* the values of $\bar{w}$ (and hence q) from $\bar{x}$ and s through the probability density function $f(\bar{w} \mid \bar{x}, s)$ where

$$f(\bar{w} \mid \bar{x}, s) = \int f(\bar{w} \mid \mu, \sigma^2)\, f(\mu, \sigma^2 \mid \bar{x}, s)\, d\mu\, d\sigma^2,$$

and $f(\mu, \sigma^2 \mid \bar{x}, s)$ is a Bayesian posterior density function for (μ, σ^2) based on a prior density function $f(\mu, \sigma^2)$ and the summary statistics $\bar{x}$ and s. Details of possible prior density functions for μ and σ^2 are given in (5). The density functions are taken to be Normal for μ and inverse chi-squared for σ^2. When informative priors for μ and σ^2 are used then the probability density function $f(\bar{w} \mid \bar{x}, s)$ can only be determined using simulation methods known as Markov Chain Monte Carlo (MCMC). When prior information for μ and σ^2 is not available, a vague prior for μ and σ^2 is used, namely $f(\mu, \sigma^2) \propto \sigma^{-2}$. The predictive density function for $f(y \mid \bar{x})$ is then a generalized t-distribution as described below.

There are two advantages of the predictive approach relative to the estimative approach. First, any prior knowledge of the variability in the parameters (μ, σ^2) of the normal distribution can be modelled explicitly. Suggestions as to how this may be done are given by Aitken et al. (5) with reference to *U.S. vs. Pirre,* (927, F.2d 694, 2nd Cir, 1991). Fifteen packages were involved and there was evidence concerning their weight and uniformity before any formal measurements were made. It was verified numerically that when determining a quantity, it is the variability among the individual packages which leads to the greatest variability in the inferences. The second advantage is that inferences about Q can be made probabilistically. Thus, it is possible to determine probability intervals for q, or more appropriately in this context, lower probability bounds for q (see Tables 1 and 2 and Figs. 1 and 2; the first for small consignments, the second for large consignments). Confidence intervals require long-term frequency properties for their validity, as explained above.

Choice of Sample Size

Consider a consignment of drugs containing N units as a random sample from some super-population of units. Let θ ($0 < \theta < 1$) be the proportion of units in the super-population which contain drugs. A prior distribution for θ may be specified by, say, a beta distribution, which has two parameters, α and β. Values for α and β may be chosen subjectively to represent the scientist's prior beliefs before inspection about the proportion of the units in the consignment (as a random sample from the super-population) which contain drugs. A large value of α relative to β would imply a belief that θ was high. Larger values of α and β would correspond to higher certainty about the value of θ. A detailed discussion is given in Aitken (4). In many cases, the scientist will not wish to quantify his prior beliefs and will wish to remain neutral. This can be done by choosing $\alpha = \beta = 1$. Also, as shown in (4), for variations in α and β, when both are small, the evidence from the sample will soon reduce the effect of the values of α and β considerably. This is intuitively reasonable: little prior information is soon subsumed by the data. A small interactive program to investigate the effects changes of α and β may have on sample size considerations is available from Dr. David Lucy.

A sample of m units from the consignment is examined and z ($\leq m$) units are found to contain drugs. The distribution of z, given m and θ, is assumed to be binomial, that is, for each unit, independently of the others in the consignment, the probability it contains drugs is taken to be equal to θ. The posterior distribution of θ is then another beta distribution with parameters $(\alpha + z)$ and $(\beta + m - z)$. This is a consequence of choosing a beta distribution as the prior distribution.

There are n units in the remainder of the consignment ($m + n = N$, the total consignment size.) Let Y ($\leq n$ and unknown) be the number of units in the remainder of the consignment which contain drugs. The total number of units in the consignment which contain drugs is then $(z + y)$ ($\leq N$). The distribution for $(Y \mid m, n, z, \alpha, \beta)$ is a Bayesian predictive distribution known as the beta-binomial distribution (14) with

$$Pr(Y = y \mid m, n, z, \alpha, \beta) = \frac{\Gamma(m + \alpha + \beta)\binom{n}{y}\,\Gamma(y + z + \alpha)\,\Gamma(m + n - z - y + \beta)}{\Gamma(z + \alpha)\,\Gamma(m - z + \beta)\,\Gamma(m + n + \alpha + \beta)}, \quad (3)$$

$$(y = 0, 1, \ldots, n).$$

The derivation of this distribution requires a beta prior and a binomial model for the data (m, z). This gives a posterior distribution for the proportion. This is then combined with a binomial model for the uninspected portion (n, y) of the consignment to give the beta-binomial distribution above. Further details are given in (4).

Estimation of Quantity of Drugs

A consignment of $m + n$ ($= N$) units is seized. A number (m) of the units are examined; the choice of m may be made following the procedures described in (4). On examination it is found that z ($\leq m$) units contain drugs and that $(m - z)$ do not. The contents of the z units which contain drugs are weighed and their weights $(x_1, \ldots, x_z)$ recorded. The remainder (n) are not examined. All of m, z and n are known.

First, consider a small consignment. Let Y ($\leq n$) denote the unknown number of units not examined which contain drugs. The estimation of quantity is able to take account of the lack of knowledge of Y. A probability function for Y may be determined using the methods described above. A weighted average of the quantities obtained for each value of Y is taken with weights the probabilities of Y obtained from an appropriate beta-binomial distribution (Eq (3) above).

Let $\mathbf{X} = (X_1, \ldots, X_z)$ and $\mathbf{W} = (W_1, \ldots, W_y)$ be the weights of the contents of the units examined and not examined, respectively, which contain drugs. It is assumed that these weights are normally distributed. Let $\overline{X} = \Sigma_{i=1}^{z} X_i/z$ and $\overline{W} = \Sigma_{i=1}^{y} W_j/y$. The total weight, Q, of the contents of the units in the consignment is then given by

$$Q = z\overline{x} + Y\overline{W}.$$

Let $\mathbf{x} = (x_1, \ldots, x_z)$ be the observed value of **X.** The distribution of $(Q \mid \mathbf{X} = \mathbf{x})$, which is a predictive distribution (10–12), is of interest. Once known, it is possible to make probabilistic statements, as distinct from confidence statements, about Q.

In the absence of prior information about the mean or variance of the distribution of the weights of drugs in the packages, a vague prior distribution is used. The probability density function of $(\overline{w} \mid z, y, \overline{x}, s^2)$ is a generalized t-distribution with $(z - 1)$ degrees of freedom (5). More explicitly, the distribution of

$$\frac{\overline{w} - \overline{x}}{s\sqrt{\frac{1}{z} + \frac{1}{y}}}$$

is a t-distribution with $(z - 1)$ degrees of freedom, denoted $t_{(z-1)}$. This distribution is not dependent on the mean and variance of the underlying normal distribution of the measurements as these have been integrated out using the information contained in **x.** It is possible to determine quantiles of this distribution and hence lower bounds for the quantity $q = z\overline{x} + y\overline{w}$, according to appropriate burdens of proof.

For given values of m, z, n, y, $\overline{x}$ and s, lower bounds for $\overline{w}$ and hence q, can be determined from the formula

$$\overline{w} = \overline{x} + s\, t_{\alpha}\sqrt{\frac{1}{z} + \frac{1}{y}}, \qquad (4)$$

where t_{α} is the $100\alpha\%$ point of the t-distribution with $(z - 1)$ degrees of freedom.

However, the value of y is not known. For a small consignment, it is a realization of a random variable which has a beta-binomial distribution as given by Eq 3. The distribution of $(\overline{w} \mid z, y, \overline{x}, s^2)$ has to be combined with Eq 3 to give a distribution of $(\overline{w} \mid s^2, \overline{x}, z)$. It is then possible to determine the distribution and corresponding probability density function of Q from the relationship $Q = z\overline{x} + y\overline{W}$ (see Appendix 1).

Consider the example from Tzidony and Ravreboy (1) in which a seized drug exhibit contained 26 street doses. A sample of six ($m = 6$) units was taken and each was analyzed and weighed. Twenty ($n = 20$) units were not examined. It was found that all six of the units examined contained drugs. The average net weight $\overline{x}$ of the powder in the six units was 0.0425 g with a standard deviation s of 0.0073 g. A 95% confidence interval for the total quantity Q in the 26 doses is 1.105 ± 0.175 g (1). Note that this interval incorporates the finite population correction factor from Eq 1 to allow for the relatively large sample size ($m = 6$) compared with the consignment size ($N = 26$). The Bayesian approach described here does not require such a correction.

Consider the approach described in Appendix 1. It is possible to determine values for Q corresponding to appropriate percentage points of the distribution. Some results are given in Table 1, for the examples given in (1), together with corresponding results with the

TABLE 2—*Estimates of quantities* q g *of drugs, in a consignment of* m + n *units, according to various possible burdens of proof, expressed as percentages* P = *100* × Pr (Q > q | m, z, n, x̄, s) *in 2600 packages when 6 packages are examined* (m = *6,* n = *2594) and* z = *6, 5, or 4 are found to contain drugs. The mean* ($\overline{x}$) *and standard deviation* (s) *of the quantities found in the packages examined which contain drugs are 0.0425* g *and 0.0073* g. *The parameters for the beta prior are* $\alpha = \beta = 1$. *Numbers in brackets are the corresponding frequentist lower bounds without using the finite population correction factor and Eq. (2).*

Percentage P	Number of Units Examined which Contain Drugs		
	6	5	4
99	54 (92)	37 (76)	24 (59)
97.5	63 (95)	44 (78)	30 (61)
95	69 (98)	51 (80)	36 (63)
90	77 (101)	58 (83)	43 (66)
70	91 (106)	74 (88)	58 (70)
60	95 (109)	79 (90)	64 (72)
50	98 (110)	84 (92)	69 (74)

TABLE 1—*Estimates of quantities* q g *of drugs, in a consignment of* m + n *units, according to various possible burdens of proof, expressed as percentages* P = *100* × Pr (Q > q | m, z, n, x̄, s) *in 26 packages when 6 packages are examined* (m = *6,* n = *20) and* z = *6, 5, or 4 are found to contain drugs. The mean* ($\overline{x}$) *and standard deviation* (s) *of the quantities found in the packages examined which contain drugs are 0.0425* g *and 0.0073* g. *The parameters for the beta prior are* $\alpha = \beta = 1$. *Numbers in brackets are the corresponding frequentist lower bounds using the finite population correction factor and Eq. (2).*

Percentage P	Number of Units Examined which Contain Drugs			Possible Burden of Proof (Illustrative)
	6	5	4	
99	0.617 (0.876)	0.435 (0.683)	0.290 (0.519)	
97.5	0.689 (0.930)	0.501 (0.744)	0.345 (0.575)	
95	0.750 (0.968)	0.559 (0.785)	0.397 (0.613)	Beyond reasonable doubt
90	0.818 (1.005)	0.628 (0.823)	0.461 (0.647)	
70	0.944 (1.067)	0.770 (0.885)	0.603 (0.704)	Clear and convincing
60	0.982 (1.087)	0.819 (0.904)	0.655 (0.721)	
50	1.015 (1.105)	0.862 (0.921)	0.704 (0.737)	Balance of probabilities

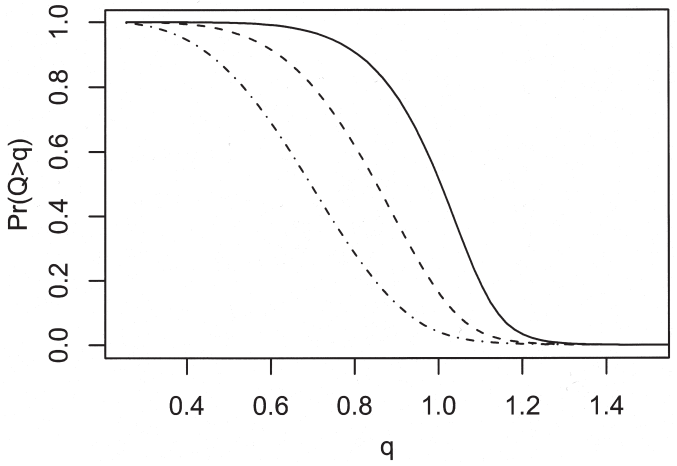

FIG. 1—*The probability that the total quantity* Q *of drugs (in grams) in a consignment of 26 units is greater than* q *when 6 units are examined and 6 (-), 5(- - -) or 4 (-·-·) units are found to contain drugs. The mean and standard deviation of the quantities found in the packages examined which contain drugs are 0.0425 g and 0.0073 g as in (1). The beta prior parameters are* $\alpha = \beta = 1$.

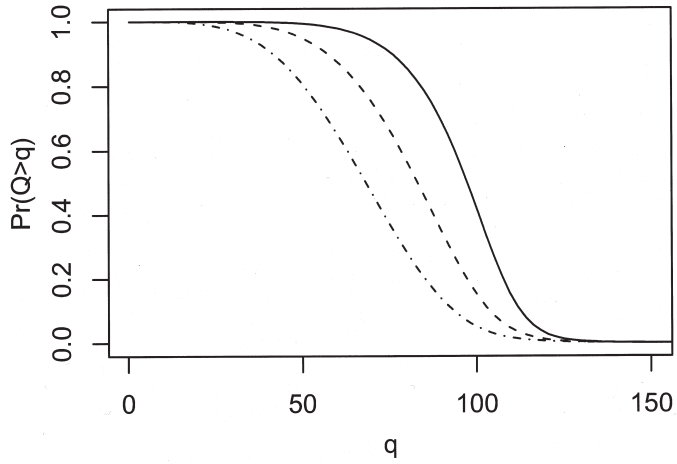

FIG. 2—*The probability that the total quantity* Q *of drugs (in grams) in a consignment of 2600 units is greater than* q *when 6 units are examined and 6 (-), 5(- - -) or 4 (-·-·) units are found to contain drugs. The mean and standard deviation of the quantities found in the packages examined which contain drugs are 0.0425 g and 0.0073 g as in (1). The beta prior parameters are* $\alpha = \beta = 1$.

method of Tzidony and Ravreboy. Predictive values only for Q are given in Fig. 1.

The lower end 0.930 g of the 95% interval determined by Tzidony and Ravreboy (1) for the quantity Q of drugs in the 26 packages may be thought of as a 97.5% lower confidence limit for Q. This can be compared with the value 0.689 *g* in the first column and second row of Table 1 which is the amount such that $Pr(Q > 0.689) = 0.975$ obtained from the predictive approach. The lower value produced by this approach arises because of the uncertainty associated with the values determined for the number of unexamined units which contain drugs. This difference is repeated throughout the table. The Bayesian approach gives smaller values for the quantities than the frequentist approach.

Further details are available in (5) and (6) where it is shown that as the burden of proof, concerning the amount of drugs in the packages, increases, the quantity for which charges may be brought decreases. For example, if proof is required *beyond reasonable doubt* and a probability of 0.99 is thought to meet this burden, then the quantity associated with this is 0.617 *g,* (assuming all six units examined contain drugs) since, from Table 1, $Pr(Q > 0.617) = 0.99$. Alternatively, if proof is required on *the balance of probabilities* and a probability of 0.50 is thought to satisfy this, then the quantity associated with this is 1.015 *g* since, again from Table 1, $Pr(Q > 1.015) = 0.50$. If less than six of the units examined are found to contain drugs, then the estimates for q decrease considerably—see the second and third columns of Table 1 and the appropriate curves in Fig. 1 for examples when only 5 or 4 of the 6 units examined are found to contain drugs.

Second, consider a large consignment. In this context, the data are used to provide a beta posterior for the proportion of illicit drugs in the whole consignment. It is assumed that the consignment size is known. The total weight, Q of the contents of the units in the consignment is given as before, by

$$Q = z\bar{x} + y\overline{W}.$$

The distribution of Q is then given by the t-density, conditional on y, with $Pr(Y = y)$ replaced by an appropriate part of a beta distribution over the interval $(0, n)$. Results for a large consignment obtained by scaling up by a factor of 100 from the results in Table 1 are shown in Table 2 and Fig. 2 with a similar pattern of results to those for small consignments. Note that in the t-density component of the expression y is treated as a discrete variable in the interval $\{0, \ldots, n\}$ and in the beta component of the expression, it is treated as a continuous variable. The treatment of y as a continuous variable for the beta integral enables the calculation of the probability that y takes a particular integer value for use with the t-density.

Acknowledgments

Thanks are due to John Dagpunar and anonymous referees who provided valuable comments on earlier drafts of this paper. This work was assisted by a Research Development Grant from the Scottish Higher Education Funding Council.

Appendix

Derivation of the distribution of the quantity of drugs in a consignment, given the results of an inspection of a sample from a small consignment

Let Q be the total quantity of drugs in the consignment. The number of units examined equals m of which z ($\leq m$) contain drugs. The mean and standard deviation of the quantity of drugs in the z units are denoted by $\bar{x}$ and s, respectively. The number of units not examined equals n, of which y (unknown) contain drugs and for which the mean quantity of drugs in these y units is $\bar{w}x$. Thus, $Q = z\bar{x} + Y\overline{W}$ in which both Y and $\overline{W}$ are random variables.

First, condition on $Y = y$. Then

$$Pr(Q < q \mid y, z, \bar{x}, s, m, n) = Pr(z\bar{x} + y\bar{w} < q \mid y, z, \bar{x}, s, m, n)$$
$$= Pr\left(\overline{W} < \frac{q - z\bar{x}}{y} \mid y, z, \bar{x}, s, m, n\right).$$

Now, given $Y = y$,

$$\frac{\overline{W} - \bar{x}}{s\sqrt{\frac{1}{z} + \frac{1}{y}}} \sim t_{z-1}.$$

Let $T = (\overline{W} - \overline{x})/\{s\sqrt{(1/z) + (1/y)}\}$. Then,

$$Pr(\overline{W} < \frac{q - z\overline{x}}{y} \mid y, z, \overline{x}, s, m, n) = Pr\left(T < \frac{q - (z + y)\,\overline{x}}{sy\sqrt{\frac{1}{z} + \frac{1}{y}}} \mid y, z, \overline{x}, s, m, n\right)$$

where $T \sim t_{z-1}$. Let

$$t_{qy} = \frac{q - (z + y)\,\overline{x}}{sy\sqrt{\frac{1}{z} + \frac{1}{y}}}.$$

Now, combine this with the result for the conditional distribution for Q, given $Y = y$, and the marginal probability function for Y, to obtain

$$Pr(Q < q \mid z, \overline{x}, s, m, n) = \sum_{y=0}^{n} Pr\left(T < \frac{q - (z + y)\,\overline{x}}{sy\sqrt{\frac{1}{z} + \frac{1}{y}}} \mid y, z, \overline{x}, s, m, n\right) Pr(Y = y) \quad (5)$$

$$= \sum_{y=0}^{n} Pr(T < t_{qy} \mid y, z, \overline{x}, s, m, n)\, Pr(Y = y).$$

The probability density function $f(q)$ of Q can be derived by differentiation of the distribution function. Let $f_{t,z-1}(.)$ denote the probability density function of the t-distribution with $(z - 1)$ degrees of freedom. Then,

$$f(q) = \sum_{y=0}^{n} f_{t,z-1}\left\{\frac{q - (z + y)\,\overline{x}}{sy\sqrt{\frac{1}{z} + \frac{1}{y}}}\right\}\left\{sy\sqrt{\frac{1}{z} + \frac{1}{y}}\right\}^{-1} Pr(Y = y).$$

Derivation of the distribution of the quantity of drugs in a consignment, given the results of an inspection of a sample from a large consignment.

For a large consignment, consider equation (5) but replace $Pr(Y = y)$ with

$$\int_{y-\frac{1}{2}}^{y+\frac{1}{2}} f(\theta)\, d\theta$$

for $y = 0, \ldots, n$,
where the density function for θ is a variation of the beta density function given earlier, with a range $(0, n)$, with

$$f(\theta \mid \alpha, \beta, n) = \frac{1}{B(\alpha, \beta)} \frac{\theta^{\alpha-1}(n - \theta)^{\beta-1}}{n^{\alpha+\beta-1}},\ 0 < \theta < n,$$

and the intervals are adjusted appropriately when $y = 0$ and n.

Then

$$Pr(Q < q \mid z, \overline{x}, s, m, n) = \sum_{y=0}^{n} Pr\left(T < \frac{q - (z + y)\,\overline{x}}{sy\sqrt{\frac{1}{z} + \frac{1}{y}}} \mid y, z, \overline{x}, s, m, n\right)$$

$$\int_{y-\frac{1}{2}}^{y+\frac{1}{2}} f(\theta \mid \alpha + z, \beta + m - z)\, d\theta.$$

As above, the probability density function $f(q)$ of Q can be derived by differentiation of the distribution function. Then,

$$f(q) = \sum_{y=0}^{n} f_{t,z-1}\left\{\frac{q - (z + y)\,\overline{x}}{sy\sqrt{\frac{1}{z} + \frac{1}{y}}}\right\}\left\{sy\sqrt{\frac{1}{z} + \frac{1}{y}}\right\}^{-1} \int_{y-\frac{1}{2}}^{y+\frac{1}{2}} f(\theta \mid \alpha + z, \beta + m - z)\, d\theta.$$

Notation

- α: a prior parameter for the beta distribution;
- β: a prior parameter for the beta distribution;
- $f_{t,z-1}\{.\}$: the probability density function of the t-distribution with $(z - 1)$ degrees of freedom;
- m: number of items inspected;
- n: number of items not inspected;
- N: consignment size $(= m + n)$;
- q: quantity to be estimated;
- Q: random variable corresponding to quantity to be estimated;
- R: number of items in the consignment which are illicit $(= y + z)$;
- s: standard deviation of measured items;
- t_α: the $100\alpha\%$ point of the t-distribution with degrees of freedom as stated;
- θ: proportion of the consignment which contains illicit items; $(w_1, \ldots, w_y)$: the weights of the contents of the items not examined which are illicit.
- $\overline{w}$: mean weight of items not inspected which are illicit; $(x_1, \ldots, x_z)$: the weights of the contents of the items examined which are illicit;
- $\overline{x}$: mean weight of inspected items which are illicit;
- y: number of items not inspected which are illicit, $y \leq n$;
- z: number of items inspected which are found to be illicit, $z \leq m$.

References

1. Tzidony D, Ravreby M. A statistical approach to drug sampling: a case study. J Forensic Sci 1992;37:1541–9.
2. Frank RS, Hinkley SW, Hoffman CG. Representative sampling of drug seizures in multiple containers. J Forensic Sci 1991;36:350–7.
3. Colón M, Rodriguez G, Diaz RO. Representative sampling of "street" drug exhibits. J Forensic Sci 1993;38:641–8.
4. Aitken CGG. Sampling—how big a sample? J Forensic Sci 1999;44: 750–60.
5. Aitken CGG, Bring J, Leonard T, Papasouliotis O. Estimation of quantities of drugs handled and the burden of proof. J Royal Statistical Soc, Series A 1997;160:333–50.
6. Izenman AJ. Statistical and legal aspects of the forensic study of illicit drugs. Statistical Sci 2001;16:35–57.
7. Coulson SA, Coxon A, Buckleton JS. How many samples from a drug seizure need to be analyzed? J Forensic Sci 2001;46:1456–61.
8. Bring J, Aitken CGG. Burden of proof and estimation of drug quantities under the Federal Sentencing Guidelines. Cardozo Law Review 1997;18:1987–1999.
9. Lindley DV. Probability. In: Aitken CGG, Stoney DA, (editors) The use of statistics in forensic science. Chichester: Ellis Horwood. 1991;27–50.
10. Aitchison J, Habbema JDF, Kay JW. A critical comparison of two methods of statistical discrimination. Applied Statistics 1977;26:15–25.
11. Aitchison J, Dunsmore I. Statistical prediction analysis. Cambridge, UK: Cambridge University Press, 1975.
12. Geisser S. Predictive inference: an introduction. London, UK: Chapman and Hall, 1993.
13. Aitken CGG. Statistics and the evaluation of evidence for forensic scientists. Chichester, UK: John Wiley & Sons Ltd, 1995.

14. Bernardo JM, Smith AFM. Bayesian Theory. Chichester, UK: John Wiley and Sons Ltd, 1994.

Additional information and reprint requests:
Colin G.G. Aitken Ph.D.
Department of Mathematics and Statistics
The King's Buildings
The University of Edinburgh
Mayfield Road
Edinburgh EH9 3JZ
E-mail: cgga@maths.ed.ac.uk.

Computer programmes for the various probability distributions referred to in the paper are available from http://www.maths.ed.ac.uk/ncgga