C. G. G. Aitken,[1] *Ph.D.*

Sampling—How Big a Sample?

REFERENCE: Aitken CGG. Sampling—how big a sample? J Forensic Sci 1999;44(4):750–760.

ABSTRACT: It is thought that, in a consignment of discrete units, a certain proportion of the units contain illegal material. A sample of the consignment is to be inspected. Various methods for the determination of the sample size are compared. The consignment will be considered as a random sample from some super-population of units, a certain proportion of which contain drugs.

For large consignments, a probability distribution, known as the beta distribution, for the proportion of the consignment which contains illegal material is obtained. This distribution is based on prior beliefs about the proportion. Under certain specific conditions the beta distribution gives the same numerical results as an approach based on the binomial distribution. The binomial distribution provides a probability for the number of units in a sample which contain illegal material, conditional on knowing the proportion of the consignment which contains illegal material. This is in contrast to the beta distribution which provides probabilities for the proportion of a consignment which contains illegal material, conditional on knowing the number of units in the sample which contain illegal material. The interpretation when the beta distribution is used is much more intuitively satisfactory. It is also much more flexible in its ability to cater for prior beliefs which may vary given the different circumstances of different crimes.

For small consignments, a distribution, known as the beta-binomial distribution, for the number of units in the consignment which are found to contain illegal material, is obtained, based on prior beliefs about the number of units in the consignment which are thought to contain illegal material. As with the beta and binomial distributions for large samples, it is shown that, in certain specific conditions, the beta-binomial and hypergeometric distributions give the same numerical results. However, the beta-binomial distribution, as with the beta distribution, has a more intuitively satisfactory interpretation and greater flexibility. The beta and the beta-binomial distributions provide methods for the determination of the minimum sample size to be taken from a consignment in order to satisfy a certain criterion. The criterion requires the specification of a proportion and a probability.

KEYWORDS: forensic science, drugs, statistics, sampling, Bayesian inference

Introduction

Consider a population or consignment, which consists of discrete units, such as individual tablets in a consignment of tablets or individual computer disks in a consignment of disks. Each unit may, or may not, contain something illegal, such as drugs or pornographic images. It is of interest to an investigating scientist to determine the proportion of the consignment which contains something illegal. This may be done exactly (assuming no mistakes are made) by examination of every unit in the consignment. Such an examination can be extremely costly. Considerable resources can be saved if information, sufficient to satisfy the needs of the investigators, may be gained from examination of a sample from the consignment. When a sample is considered, uncertainty is introduced when inference is made from the sample to the population, because the whole population is not inspected. However, this uncertainty may be quantified probabilistically. It is shown that if two numbers are specified in advance of the inspection of the consignment, then a sample size may be specified. The first of these numbers is the minimum proportion of units in the consignment which contain something illegal that the examination is to be designed to find. The second is the required probability with which the true proportion of illegal units exceeds this minimum proportion.

With reasonable assumptions, a probability distribution for the true proportion of units in the consignment is derived, based on the scientist's prior beliefs (i.e., prior to the inspection of individual units) and the outcome of the inspection of the sample. The prior beliefs may be based on presumptive tests. For example, the physical appearance of the suspected illegal material may be similar to that of other known illegal material from the scientist's experience. Alternatively, the results of an initial examination of some of the material by color spot tests may affect the prior beliefs. It is possible to choose a function to represent the strength of the scientist's prior beliefs. It may be thought inappropriate that the scientist's prior beliefs should have any effect on the decision to be made regarding the sample size. In such a case, it is possible to choose the function in such a way that the effect is very small. (It is also possible to choose the function such that the effect is very large.) 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 a subjective choice. The binomial model described here requires assumptions about independence of the probability for each unit being illegal and the choice of a constant value for this probability.

The function representing the scientist's prior beliefs is then combined with a function which accounts for the observation of the number of units in the sample which contain something illegal. The combination of these two functions provides a third function which represents the probability distribution for the true proportion of illegal units in the consignment. It is shown here that this function is of the same form as the one representing the scientist's prior beliefs. From this so-called posterior distribution (i.e., posterior to the inspection of individual units) it is possible to determine the probability with which the true proportion exceeds any specified proportion.

This approach provides a probability statement about the true proportion. This is in contrast to the inference obtainable from an approach which provides a confidence interval for the true proportion. It is no accident that the word probability is not used to describe this

[1] Department of Mathematics and Statistics, The King's Buildings. The University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom.

Received 29 May 1998; and in revised form 16 Oct. 1998; accepted 19 Oct. 1998.

interval. A confidence interval derives its validity as a method of inference on a long-run frequency interpretation of probability (1). For example, consider specifically, the 95% confidence interval for a proportion. The probability with which this interval contains the true proportion is not known. However, suppose the experiment which generated the 95% confidence interval is repeated many times (under identical conditions, a theoretical stipulation, which it is impossible to fullfil in practice) and on each of these occasions a 95% confidence interval for the true proportion is calculated. Then, it can be said that 95% of these (95%) confidence intervals will contain the true proportion. This does not provide information concerning the one 95% confidence interval which has been calculated. It is not known whether it does or does not contain the true proportion, and it is not even possible to determine the probability with which it contains the true proportion.

The method which uses the scientist's prior beliefs and enables a probability statement to be made is known as the Bayesian method, after the Rev. Thomas Bayes (2–4). The procedure by which prior beliefs and a function of the observations may be combined to provide posterior beliefs is known as Bayes' Theorem. The method which relies on the idealized long-run frequency for its validity is known as the frequentist method.

It is the purpose of this paper to compare the results obtained from the Bayesian and frequentist approaches to assessing uncertainty, to clarify the assumptions made in the two approaches, to contrast the clarity of the inferences obtainable from the Bayesian approach with the lack of clarity associated with the frequentist approach, and to illustrate the greater flexibility of the Bayesian approach with the inflexibility of the frequentist approach. The methods are illustrated with reference to sampling from consignments of drugs. However, they apply equally well to sampling in other forensic contexts, for example, glass (5) and pornographic images.

Frequentist procedures are described in (6) for choosing a sample size from a consignment. Distinction is drawn between an approach based on the binomial distribution and an approach based on the hypergeometric distribution. It is argued in (6) that the former can be used for large consignments in which the sampling of units may be taken to be sampling with replacement. For small samples, the sampling units may be taken to be sampling without replacement and the hypergeometric approach is used. The Bayesian approach also has different methods for analyzing large and small samples.

As reported in (7), various methods for selecting the size of a random sample from a consignment have been accepted by the U.S. courts. An approach based on the hypergeometric distribution is proposed in (7). A summary of different procedures used in 27 laboratories around the world in given in (8). These procedures include methods based on the square root of the consignment size, a percentage of the consignment size, and a fixed number of units regardless of the consignment size, as well as the hypergeometric distribution. The authors in (8) propose the formula.

$$m = 20 + 10\%(N - 20) \qquad (\text{for } N > 20)$$

where m is the sample size and N is the consignment size. As well as being simple to implement, this approach, as the authors rightly claim, provides the opportunity to discover heterogeneous populations before the analysis is completed. According to (7), 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 (as described above) and not in probabilistic terms. For example, from (7), a statement of the form that "at the 95% confidence level, 90% or more of the packages in an exhibit contain the substance" is suggested as being sufficient proof in cases of drug handling. The procedures to be described here provide summaries in probabilistic terms.

In general, an answer is provided to the question:

> "How big a sample should be taken for it to be said that there is a 100 p% probability that the proportion of units in the consignment which contain drugs is greater than $100\theta_0$%?"

For an example of a particular instance, an answer is provided to the question:

> "How big a sample should be taken for it to be said that there is a 95% probability that the proportion of units in the consignment which contain drugs is greater than 50%?"

Here, the value 0.95 has been substituted for p and the value 0.5 has been substituted for θ_0. This requirement may not seem very stringent but may be sufficient to satisfy certain legal requirements.

A Comparison of Two Methods of Measuring Uncertainty

Before comparing two methods of estimating the sample size necessary to make a statement about the proportion of units in a consignment which contain drugs, it is useful to consider further the two methods of measuring uncertainty on which these two methods are based, and which were outlined in the Introduction.

The Frequentist Method—This assumes that the proportion θ of the consignment which contains drugs is unknown but fixed. The data, that is the number of units in the sample which contain drugs, are variable. A so-called *confidence interval* or *level* is calculated. The name *confidence* is used since no probability can be attached to the uncertain event that the interval contains θ.

The frequentist approach derives its name from the relative frequency definition of probability. The probability that a particular event, A say, occurs is defined as the relative frequency of the number of occurrences of event A compared with the total number of occurrences of all possible events, over a long run of observations, conducted under identical conditions.

For example, consider tossing a coin n times. It is not known if the coin is fair and the outcomes of the n tosses are to be used to determine the probability of a head occurring on an individual toss. There are two possible outcomes, heads (H) and tails (T). Let $n(H)$ be the number of H and $n(T)$ be the number of T such that $n(H) + n(T) = n$. Then the probability of tossing a head on an individual toss of the coin is defined as the limit as $n \to \infty$ of the fraction $n(H)/n$.

The validity of the frequentist approach, however, relies on a belief in the long-run repetition of trials under identical conditions. This is an idealized situation, seldom, if ever, realized in practice.

The Bayesian Method: Subjective Probability—The Bayesian approach represents the uncertainty concerning knowledge of θ (the proportion of interest) with a probability distribution. The data are taken as fixed, in contrast to the frequentist method. A sample of a particular size m from a consignment has been taken and the number of units z which contain drugs noted. These data are considered fixed. There is no consideration for the long-run repetition of trials under identical conditions. Data which may have been observed but have not are not allowed to affect the analysis. In the frequentist method the probability of z given m and θ is a function represented by the binomial distribution. In the Bayesian method, the same func-

tion is expressed as a function of θ, known as the likelihood function. It has the same mathematical form as the binomial distribution but takes the data as fixed and expresses the form as a function of θ:

$$L(\theta \mid m, z) = \binom{m}{z} \theta^z (1-\theta)^{m-z} \qquad (0 < \theta < 1). \qquad (1)$$

Uncertainty about θ can be expressed as a probability distribution. Probability intervals may be determined for θ with a much clearer interpretation than with confidence intervals.

Choice of Sample Size

Large Consignments—A large consignment will be taken to be one which is sufficiently large that sampling is effectively with replacement. This can be as small as 50, though in many cases it will be of the order of many thousands.

A consignment of drugs containing *N* units will be considered as a random sample from some super-population of units which contain drugs. Let θ ($0 < \theta < 1$) be the proportion of units in the super-population which contain drugs. For consignment sizes of the order of several thousand all realistic values of θ will represent an exact number of units. For small sample sizes less than 50, θ can be considered as a nuisance parameter (5 and Appendix 1) and integrated out of the calculation leaving a probability distribution for the unknown number of units in the consignment which contain drugs as a function of known values. For intermediate calculations, θ can be treated as a continuous value in the interval ($0 < \theta < 1$), without any detriment to the inference. As before, let *m* be the number of units sampled and let *z* be the number which are found to contain drugs.

Frequentist Approaches to the Estimation of θ—The sample proportion $p = z/m$ is an unbiased estimate of θ. The variance of *p* is given by (9)

$$\frac{\theta(1-\theta)}{m}\left(\frac{N-m}{N-1}\right).$$

The factor $(N - m)/(N - 1)$ is known as the *finite population correction* (fpc). Provided that the sampling fraction m/N is low, the size of the population has no direct effect on the precision of the estimate of θ. For example, if θ is the same in the two populations, a sample of 500 from a population of 200,000 gives almost as precise an estimate of the population proportion as a sample of 500 from a population of 10,000. The estimated standard deviation of θ in the second case is 0.98 times the estimated standard deviation in the first case.

Consider the following example. To simplify matters, the fpc is ignored and the sample proportion *p* is assumed to be normally distributed. Assume that θ is thought to be about 75%. It is stipulated that a sample size *m* is to be taken to estimate θ to within 25%, i.e., in the interval (0.50, 1.00) with 95% confidence. (This may be thought a very wide interval but it is consistent with the form of question posed at the end of the Introduction that a sample size be determined such that it can be said that, if all of the sample are found to contain drugs, then there is a 95% probability that θ is greater than 50%.) The criterion for the sample size is that there should be a confidence of 0.95 that the sample proportion *p* lies in the interval 0.75 ± 0.25. From known results of the normal distribution, this implies that two standard deviations equal 0.25. The standard deviation of *p*, ignoring the fpc, is

$$\sqrt{\frac{\theta(1-\theta)}{m}}$$

(see also equation (5) of (5)). Setting two standard deviations equal to 0.25, and solving for *m*, gives the following expression for *m*:

$$m = \frac{4\theta(1-\theta)}{0.25^2}.$$

When $\theta = 0.75$, $m = 12$. Thus a sample of size 12 is sufficient to estimate θ to be greater than 0.5 with *confidence* 0.95. Similar calculations are reported in (6), where one standard deviation is set equal to 0.1 and the sample size is chosen to maintain this value of the standard deviation for various values of *N* and estimated values of θ. Later, it will be shown using Bayesian techniques, that, if all the inspected samples are found to contain drugs, the required sample size to enable one to say that $\theta > 0.5$ with *probability* 0.95, is 4.

An alternative approach, based on the binomial distribution, is discussed in (5). Consider a specific value θ_0 of θ. The forensic scientist wishes to show that $\theta > \theta_0$. Another probability, α, is selected as the probability that the null hypothesis $\theta > \theta_0$ is rejected based on an inspection of a sample of units from the consignment. It is also assumed that the inspection of the sample reveals that all the units in the sample contain drugs. An example when not all the units which are examined contain drugs is given in (6).

This analysis may be considered as a hypothesis test. The null hypothesis is that $\theta > \theta_0$; the alternative is that $\theta < \theta_0$. As stated above, it is assumed that $m = z$. Standard statistical theory shows that the null hypothesis is rejected in favor of the alternative if $\theta_0^m \leq \alpha$. The probability α is the probability that the hypothesis $\theta > \theta_0$ is rejected. The complementary probability $(1 - \alpha)$ is the probability that the hypothesis $\theta > \theta_0$ is not rejected. Failure to reject the null hypothesis is not the same as saying it is true. There will be many occasions in which data are such that the null hypothesis is not rejected but for which the data provide stronger support for an alternative hypothesis. The probability that the hypothesis $\theta > \theta_0$ is not rejected is not the same as the probability that $\theta > \theta_0$. For this reason the word *confidence* is used instead of probability when referring to the first situation. It is said that one has $100(1 - \alpha)\%$ confidence that $\theta > \theta_0$.

Table 1, a subset of Table 1 of (6) shows the values of $(1 - \alpha)$ for various values of θ_0 and *m*.

The meaning of the results in Table 1 can be illustrated by consideration of $\theta_0 = 0.7$. Assume $\theta_0 = 0.7$. Five units of the consignment are examined and all are found to contain drugs. The probability of this happening, when $\theta_0 = 0.7$, is 0.7^5 which equals 0.17. Thus, there is 83% confidence that θ, the true proportion of drugs in the consignment, is greater than 0.7. Similarly, there is 67% confidence that θ is greater than 0.8 and 41% confidence that θ is greater than 0.9. However, these statements are not probability statements about the value of θ. The statement concerning the first result, written more fully, is as follows.

> Suppose $\theta < 0.7$. The probability that, when 5 units are examined, all are found to contain drugs is less than 0.17.

TABLE 1—*Values of* $(1 - \alpha)$ *for various values of* θ_0 *and m, from (6).*

	θ_0		
m	0.7	0.8	0.9
5	0.83	0.67	0.41
10	0.97	0.89	0.65
15	0.99	0.96	0.79

What is taken to be known in the probability calculation is that $\theta < 0.7$. The probability statement is concerned with the probability that all the examined units contain drugs, **if** $\theta < 0.7$. The data are variable, θ is fixed.

This is the transpose of what is required. In practice, it is known that a sample has been examined and all of the units in the sample have been found to contain drugs. It is then the probability that $\theta < 0.7$, say, which is of interest.

In the first case, illustrated in Table 1, probability statements of the form

$$Pr(z = m \mid m \text{ units examined}, \theta = 0.7)$$

are considered. In the transpose, probability statements of the form

$$Pr(\theta < 0.7 \mid m \text{ units examined and } z = m)$$

are considered. In both cases, z is the number of units examined which are found to contain drugs. It is one of the main purposes of this paper to demonstrate that this second approach provides a more intuitively satisfactory method of determining the sample size.

A Bayesian Method for the Estimation of θ—In order to make probability statements about θ, it is necessary to have a probability distribution for θ, to represent the variability in 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. However, the Bayesian philosophy permits one to represent this uncertainty as a probability distribution. The most common distribution for θ is the so-called beta distribution (10). Its use in another forensic context, that of sampling glass fragments, is described in (5).

A continuous random variable θ has a beta distribution with parameters (α, β, $\alpha > 0$, $\beta > 0$), denoted Beta(α, β), if its probability density function $f(\theta \mid \alpha, \beta)$ is

$$f(\theta \mid \alpha, \beta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}/B(\alpha, \beta), \qquad 0 < \theta < 1, \quad (2)$$

where

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

and

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$$

is the gamma function. Integer and half-integer values of the gamma function are found from the recursive relation $\Gamma(x+1) = x\,\Gamma(x)$ and the values $\Gamma(1) = 1$ and $\Gamma(^1/_2) = \sqrt{\pi} \simeq 1.7725$.

Note that the beta distribution models characteristics which only takes values in the range (0,1), which is particularly appropriate for proportions. Graphs of the beta distributions with parameters (3,2), (3,1) and (10,1) are shown in Fig. 1 (a), (b), and (c). The graph of Beta (3,1) is proportional to θ^2. This reflects belief that the most likely outcome is that all units in the consignment contain drugs with a belief that reduces as a quadratic as θ decreases from 1 to 0. The graph of Beta (10,1) is proportional to θ^9. The graph of Beta (3,2) has a mode at $\theta = {}^2/_3$, reflecting a belief that it is quite likely that there are units in the consignment which do not contain drugs.

The beta distribution is technically convenient in the context of sampling from a discrete consignment because it is a so-called conjugate prior distribution for the binomial distribution. It combines with the binomial distribution to provide a posterior distribution which is also a beta distribution. See Appendix 1 for further details. Thus, if m units are examined and z are found to contain drugs then the probability density function which combines this information with the prior distribution is given by

$$f(\theta \mid z, m, \alpha, \beta) = \theta^{z+\alpha-1}(1-\theta)^{m-z+\beta-1}/B(z+\alpha, m-z+\beta), \qquad 0 < \theta < 1,$$

denoted Be($z+\alpha$, $m-z+\beta$). In the particular case where $z = m$, the density function is given by

$$f(\theta \mid m, m, \alpha, \beta) = \theta^{m+\alpha-1}(1-\theta)^{\beta-1}/B(m+\alpha, \beta), \qquad 0 < \theta < 1.$$

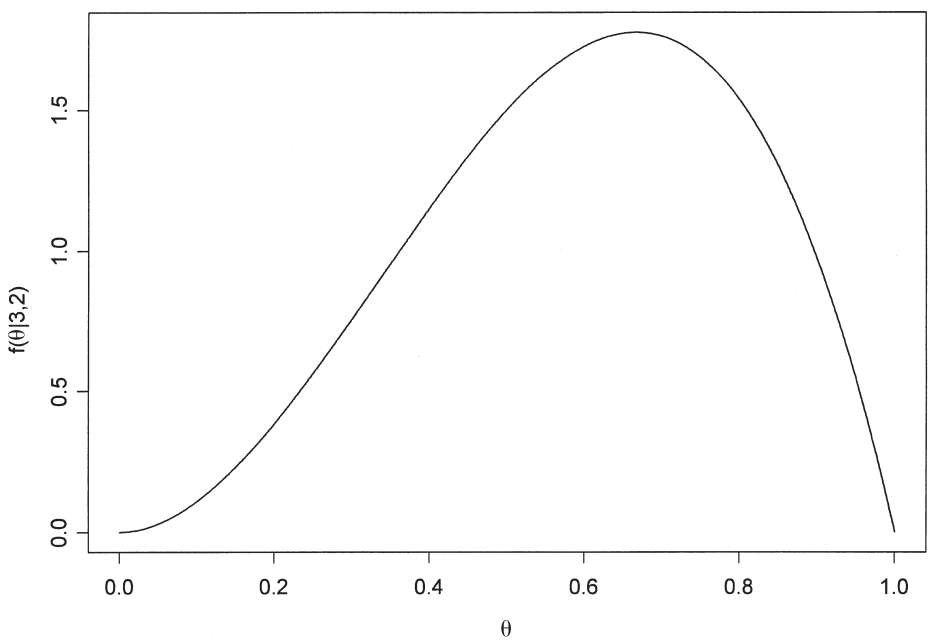
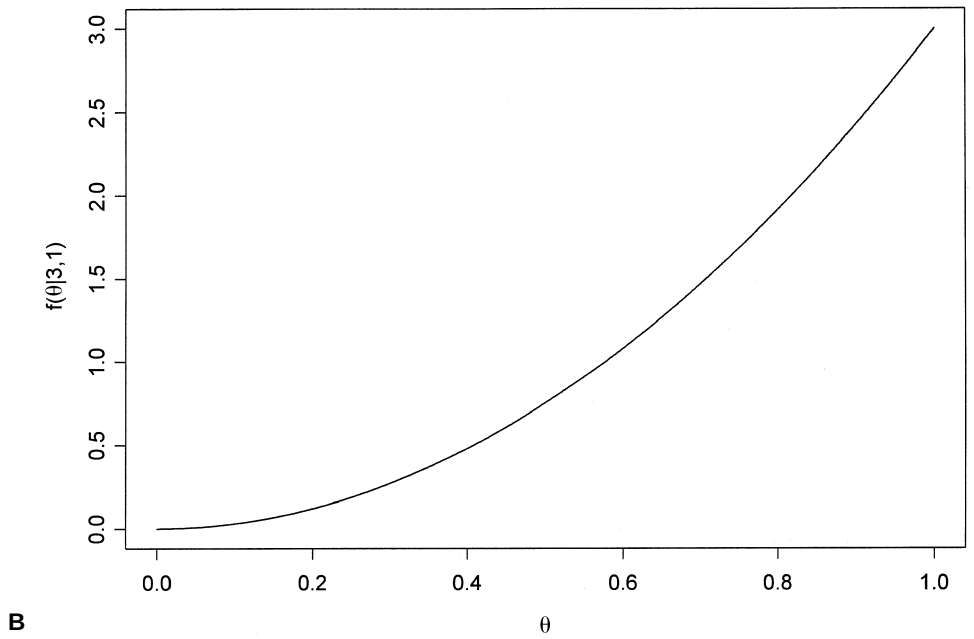
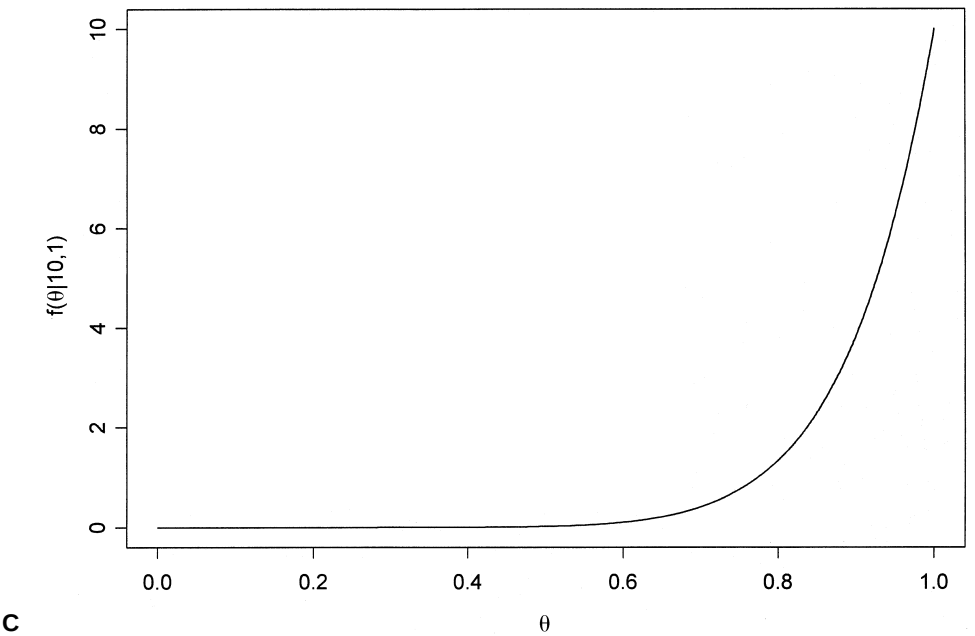


FIG. 1—*Probability density functions from (2) for the beta distribution with parameters* (a) $\alpha = 3$, $\beta = 2$, (b) $\alpha = 3$, $\beta = 1$, (c) $\alpha = 10$, $\beta = 1$.



B



C

FIG. 1—(*Continued*)

There is an interesting comparison with the frequentist formulation of the problem, when a limiting case is considered. Let $\alpha \to 0$ and let $\beta = 1$. This limiting case gives rise to an improper prior distribution, so-called because it is not a probability density function. However, the analysis is justified in that the posterior distribution is a probability density function,

$$f(\theta \mid m, m, \alpha, \beta) \to \theta^{m-1}/B(m, 1) = m\theta^{m-1}.$$

Thus,

$$Pr(\theta < \theta_0 \mid m, m, \alpha, \beta) \to \int_0^{\theta_0} m\theta^{m-1}\, d\theta = \theta_0^m,$$

the same expression as used in the frequentist case but with a much more coherent interpretation, since it is now a probability statement about θ. For example, if $m = 5$,

$$Pr(\theta < 0.7 \mid 5, 5, \alpha, \beta) \to 0.7^5 = 0.17,$$

the same numerical value as before.

Another commonly used prior distribution is one in which $\alpha = \beta = 1$. Then

$$f(\theta \mid 1, 1) = 1 \qquad 0 < \theta < 1.$$

This distribution is the so-called uniform distribution and is often used to represent maximum uncertainty about θ. Another representation of uncertainty is the case where $\alpha = \beta = {}^1/_2$ where greater belief is placed at the ends of the range, in favor of all items or no

items containing drugs, than in the middle. Then

$$f(\theta \mid 0.5, 0.5) = \pi^{-1}\,\theta^{-1/2}(1-\theta)^{-1/2}.$$

In practice, a criterion has to be specified in order that the sample size may be determined. Consider the criterion from the Introduction that the scientist wishes to be 95% certain that 50% or more of the consignment contains drugs when all units sampled contain drugs. Then the criterion may be written mathematically as

$$Pr(\theta > 0.5 \mid m, m, \alpha, \beta) = 0.95$$

or

$$\int_{0.5}^{1} \theta^{m+\alpha-1}(1-\theta)^{\beta-1} d\theta / B(m+\alpha, \beta) = 0.95. \qquad (3)$$

The general question in which p and θ_0 are specified at the end of the Introduction may be answered by finding the value of m which solves the equation

$$\int_{\theta_0}^{1} \theta^{m+\alpha-1}(1-\theta)^{\beta-1}\, d\theta / B(m+\alpha, \beta) = p. \qquad (4)$$

Such integrals are easy to evaluate using standard statistical packages (e.g., *SPLUS* (11)) given values for m, α and β. It is then a simple matter to substitute specified values for θ_0 and p and given values for α and β and then select m by trial and error to solve (4).

TABLE 2—*The probability that the proportion of drugs in a large consignment is greater than 50% for various sample sizes* m *and prior parameters* α *and* β.

		m			
α	β	2	3	4	5
1	1		0.94	0.97	
0.5	0.5	0.92	0.97	0.985	0.993
0.065	0.935		0.90	0.95	0.97

For example: Consider the following pairs of values for α and β: (1,1), (0.5,0.5), and (0.065,0.935) (the last pair suggested by Professor T. Leonard, personal communication). For the first two pairs there is a prior probability of 0.5 that $\theta > 0.5$; for the third pair there is a prior probability of 0.05 that $\theta > 0.5$. This third choice was made since 0.05 is the complement of 0.95. Table 2 shows the results for (3) for various values of m.

This gives the remarkable result that, for large consignments, of whatever size, the scientist need only examine 4 units, in the first instance. If all are found to contain drugs, there is a 95% probability that 50% of the consignment contains drugs. Compare this with the result derived from a frequentist approach using a normal approximation to the binomial distribution which gave a value of 12 for the sample size. These sample sizes are not large. However, there is not very much information gained about the exact value of θ. It is only determined that there is probability of 0.95 that $\theta > 0.5$. This is a wide interval (from 0.5 to 1) within which the true proportion may lie.

Figures 2 and 3 illustrate how varying prior beliefs have little influence on the conclusions once some data have been observed. Figure 2 shows the prior probability that $\theta > \theta_0$ for $0 < \theta_0 < 1$, for the values of (α, β) given in Table 2, decreasing from a value of 1 when $\theta_0 = 0$ to a value of 0 when $\theta_0 = 1$. There are considerable differences in the curves. Figure 3 shows the corresponding posterior probabilities for $\theta > \theta_0$ given four units have been observed and all have been found to contain drugs, with the values for $\theta_0 = 0.5$ emphasized. There is very little difference in these curves. There may be concerns that it is very difficult for a scientist to formalize his prior beliefs. However, if α and β are small, large differences in the probabilities associated with the prior beliefs will not lead to large differences in the conclusions.

The methodology can be extended to allow for units which do not contain drugs. For example, if one of the original four units inspected is found not to contain drugs then three more should be inspected. If they all contain drugs, then it can be shown that the probability that $\theta > 0.5$, given that six out of seven contain drugs, is 0.96.

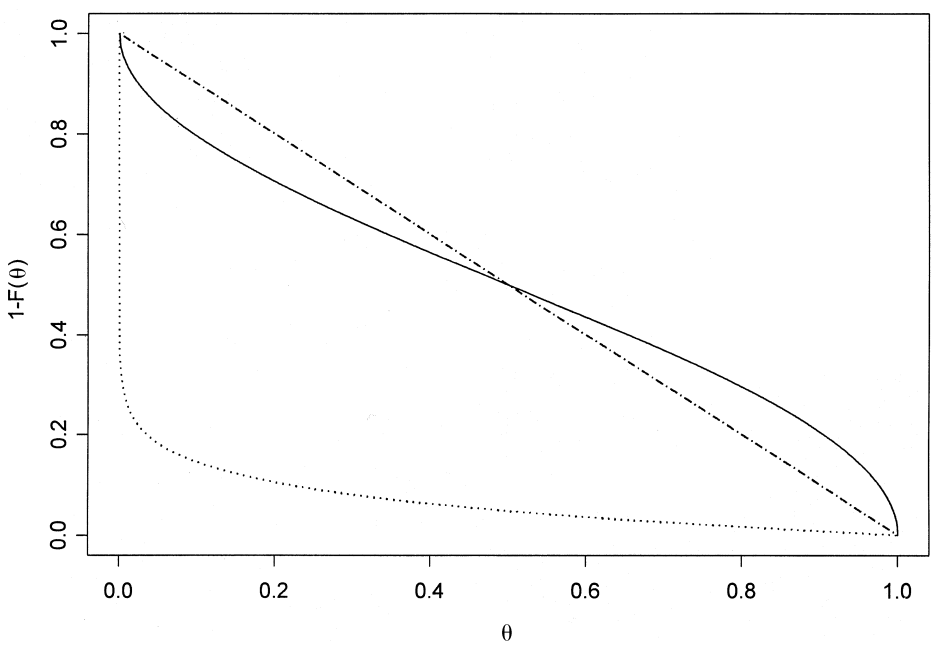


FIG. 2—*The prior 1-F* (θ_0) *probability that the proportion* θ *of units in a consignment which contain drugs is greater than* θ_0, *for various choices of* α *and* β: $\alpha = \beta = 1$(– · – ·), $\alpha = \beta = 0.5$(– – –), $\alpha = 0.065$, $\beta = 0.935$(···).

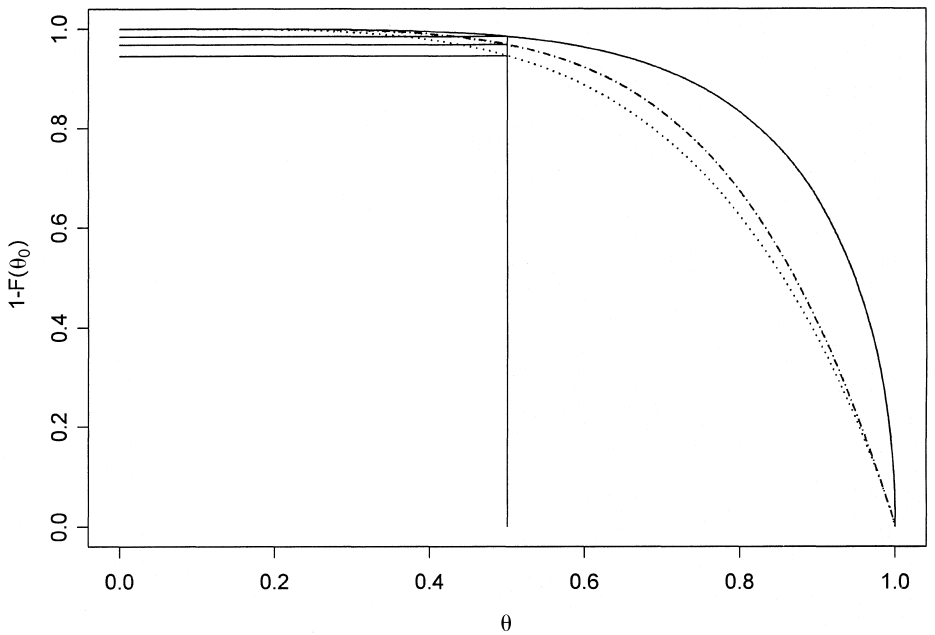


FIG. 3—*The posterior probability 1-F* (θ_0) *that the proportion* θ *of units in a consignment which contain drugs is greater than* θ_0*, after inspection of m = 4 units has shown them all to contain drugs, for various choices of* α *and* β*:* $\alpha = \beta = 1$*(– · – ·),* $\alpha = \beta = 0.5$*(– – –),* $\alpha = 0.065$*,* $\beta = 0.935$*(···). The solid lines show the probabilities that* $\theta > 0.5$ *for the various choices of* α *and* β *(from Table 2, with m = 4, the probabilities are 0.97, 0.985 and 0.95, respectively, for* (α, β) *= (1, 1), (0.5, 0.5) and (0.065, 0.935)).*

TABLE 3—*The sample size required to be 100p% certain that the proportion of units in the consignment which contain drugs are greater than* θ_0*, when all the units inspected are found to contain drugs. The prior parameters* $\alpha = \beta = 1$*.*

	p		
θ_0	0.90	0.95	0.99
0.5	3	4	6
0.6	4	5	9
0.7	6	8	12
0.8	10	13	20
0.9	21	28	43
0.95	44	58	89
0.99	229	298	458

TABLE 4—*The probability* p *that the proportion of a large consignment which contains drugs is greater than* θ_0 *when a sample of size* m *is inspected and the entire sample is found to contain drugs* ($\alpha = \beta = 1$*.)*

	m					
θ_0	3	4	5	50	100	150
0.5	0.94	0.97	0.98			
0.6	0.87	0.92	0.95			
0.7	0.76	0.83	0.88			
0.8	0.59	0.67	0.74			
0.9	0.34	0.41	0.47	0.995		
0.95	0.19	0.23	0.26	0.927		
0.99	0.04	0.05	0.06	0.401	0.638	0.781

The dependency of the sample size on the values of p and θ_0 is illustrated in Table 3. The prior parameters α and β are set equal to 1. Consider $p = 0.90$, 0.95 and 0.99 and consider values of $\theta = 0.5$, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99. The sample size m required to be $100p\%$ certain that θ is greater than the specified value is then given by the value of m which satisfies the equation

$$Pr(\theta > \theta_0 \mid m, m, 1, 1) = 1 - \theta_0^{m+1} = p,$$

a special case of (4). The value of m is thus given by the smallest integer greater than

$$[\log(1 - p)/\log(\theta_0)] - 1.$$

Obviously, when considering the results in Table 3, the consignment size has to be taken into account in order that the sample size may be thought small with respect to the size of the consignment. Thus, for the last row in particular to be useful, the size of the consignment from which the sample is to be taken will have to be of the order of several tens of thousands.

An alternative representation, given in Table 4, considers the value of p which is obtained for various values of θ_0 and for given sample sizes m, when the entire sample of size m is found to contain drugs. The first row of Table 2 is a special case of Table 4.

There may be situations in which different choices of α and β may be wanted. The outcome for small samples of some different choices is shown in Table 8. It may be the scientist has some substantial prior beliefs about the proportion of the consignment which may contain drugs. These beliefs may arise from previous experiences of similar consignments (from what may be considered to be the same super-population), for example. In such cases, use can be made of various properties of the beta distribution to assist the scientist in choosing values for α and β. The mean of the beta distribution is $\alpha/(\alpha + \beta)$ and the variance is $\alpha\beta/(\alpha + \beta)^2(\alpha + \beta + 1)$. Thus, a prior belief of the proportion of the consignment which contains drugs would set that proportion equal to the mean of the distribution and a belief about how precise that belief was, would provide a value for the variance. Alternatively, if it was felt that β

could be set equal to 1, so that the shape of the probability density function is similar to those in Fig. 1(b) and 1(c) (i.e., monotonic increasing with respect to θ), and that there was a prior belief about a lower bound for the proportion, say that

$$Pr(\text{Proportion} > \theta_0 \mid \alpha, \beta) = p$$

then use could be made of the result that

$$\alpha = \log(1 - p)/\log(\theta_0).$$

Small Consignments—Suppose now that the consignment size N is small. A sample of m units from the consignment is examined and $z(\leq m)$ units are found to contain drugs. Denote the number, $N - m$, of units not examined by n, so that $m + n = N$.

A Frequentist Approach—Consider a frequentist approach based on the hypergeometric distribution (6). Let $R = Z + Y$ be the total number of units in the consignment which contain drugs, where Z is the number of units in the sample of size m and Y is the number of units in the remainder which contain drugs. Then the distribution of Z is hypergeometric with

$$Pr(Z = z) = \frac{\binom{R}{z}\binom{N-R}{m-z}}{\binom{N}{m}}, \qquad z = 0, 1, \ldots, \min(R, m).$$

To satisfy a given confidence level $(1 - \alpha)$, the maximum value of R is needed, (6), such that

$$\sum_{x = \min(o,\, m-R)}^{\max(m-z,\, N-R)} \frac{\binom{R}{m-x}\binom{N-R}{x}}{\binom{N}{m}} \leq \alpha.$$

When $m = z$ this reduces to

$$\frac{R!(N - m)!}{N!(R - m)!} \leq \alpha.$$

Table 5, part of Table 3 in (6), shows the confidence levels when $m = z$ and $\theta = 0.7$ for $N = 10$, 20 and 30 (and hence $R = 7$, 14, 21) and for $m = 5$, 10, 15.

Consider the value 0.92 when $N = 10$ and $m = 5$. This is the probability that, if $\theta = 0.7$, 5 units in a consignment of size 10, when examined, will be found to all contain drugs. Again this assumes θ to be known to be 0.7 and gives a value for the probability that $m = z$. This is translated in frequentist terms to read that "one is 92% confident that, when 5 units in a population of 10 are examined and all are found to contain drugs, $\theta > 0.7$."

The Beta-Binomial Distribution—The interpretation in the previous section is not so clear as that obtained from the Bayesian method which uses a so-called beta-binomial distribution (10). The beta-binomial distribution provides a probability statement about the number of units in the consignment which contain drugs.

TABLE 5—*Confidence levels when the true proportion of drugs in the consignment is 0.7, for consignments of size* N *and samples of size* m.

	N		
m	10	20	30
5	0.92	0.86	0.847
10	—	0.98	0.982
15	—	≃ 1.0	0.998

As before, let θ $(0 < \theta < 1)$ be the proportion of units in the super-population which contain drugs. The probability distribution of z, given m and θ, may be taken to be binomial. 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 another beta distribution with parameters $(\alpha + z)$ and $(\beta + m - z)$ (see Appendix 1).

Since the consignment size is small, a better representation of the variability of the number of unexamined units in the consignment which contain drugs is obtained by considering a probability distribution for this number, Y, explicitly. There are n units in the remainder of the consignment $(m + n = N)$ which have not been examined. Then Y (unknown and $\leq n$) is the number of units in this remainder which contain drugs. Given θ, the distribution of $(Y \mid n, \theta)$, like that of $(Z \mid m, \theta)$, is binomial. However, θ has a beta distribution and the distribution of $(Y \mid n, \theta)$ and the distribution of $(\theta \mid m, z, \alpha, \beta)$ can be combined to give what is known as a Bayesian predictive distribution for $(Y \mid m, n, z, \alpha, \beta)$, known as a beta-binomial distribution (10).

$$Pr(Y = y \mid m, n, z, \alpha, \beta) = \frac{\Gamma(m + \alpha + \beta)\binom{n}{y}\Gamma(y + z + \alpha)\Gamma(m + n - z - y + \beta)}{\Gamma(z + \alpha)\Gamma(m - z + \beta)\Gamma(m + n + \alpha + \beta)},$$
$$(y = 0, 1, \ldots, n), \quad (5)$$

(see Appendix 1).

From this distribution, inferences can be made about Y, such as probability intervals or lower bounds for Y. Note the flexibility given by the ability to vary α and β to incorporate prior beliefs about the proportion of units in the consignment which contain drugs. Also, for integer values of α and β, (5) reduces to a function of factorials.

Comparison of Beta-Binomial and Hypergeometric Approaches—For large samples, it has been shown that in a limiting case, the beta distribution and the binomial distribution give the same numerical answers, though with different interpretations. A corresponding comparison can be made with the beta-binomial and hypergeometric distributions.

Suppose m units are examined and z are found to contain drugs; n units are not examined. Let y be the number of unexamined units which contain drugs and let $r = z + y$. Then y is unknown. It can be shown (see Appendix 2) that

$$Pr(R = r \mid m, n, z) = \frac{(m + 1)\binom{m}{z}\binom{n}{y}}{(m + n + 1)\binom{m+n}{z+y}} = Pr(Y = y \mid m, n, z, 1, 1), \quad (6)$$

the beta-binomial (5) distribution with $\alpha = \beta = 1$, where r can take values $z, \ldots, n + z$, and y can take values $0, \ldots, n$. This result depends on a theorem known as Vandermonde's theorem (12,13). This is not a new result. Todhunter (14) credits Condorcet in 1785 and Prevost and Lhuilier in 1799 with recognizing it. Todhunter (14) comments that "the coincidence of the results obtained on the two different hypotheses is remarkable." This result has also been used more recently in forensic science in the context of glass sampling (5).

Use of the Beta-Binomial Distribution—As an example, consider a consignment of size $N = 10$, where five units are inspected and all five are found to contain drugs $(m = z = 5)$. For the pro-

portion of units in the consignment which contain drugs to be at least 0.7 ($\theta \geq 0.7$), it is necessary for the number of units Y in the five units not inspected to be at least 2($Y \geq 2$). The beta-binomial probability, with a uniform prior $\alpha = \beta = 1$, is given by

$$Pr(Y \geq 2 \mid 5, 5, 5, 1, 1,) = \sum_{y=2}^{5} \frac{6\binom{5}{5}\binom{5}{y}}{11\binom{10}{5+y}} = 0.985,$$

the summation of terms in (6), with $m = n = z = 5$ and for y taking values from 2 to 5.

The hypergeometric distribution has the interpretation that if $m = z = 5$, one is 92% confident that $\theta > 0.7$. The beta-binomial approach enables one to assign a probability of 0.985 to the event that $\theta > 0.7$.

As with large consignments, 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.

When considering large consignments, the criterion was stipulated that a sample size was to be chosen such that if all the units contained drugs then there was to be a probability of 0.95 that the proportion of units in the whole consignment which contains drugs was greater than 0.5. This criterion was used to investigate sample sizes required for small consignments in which $N \leq 50$. The results are shown in Table 6, where $\alpha = \beta = 1$.

Thus, for consignments no greater than 50 in size, samples of size 3 need to be examined. For samples greater than 50, theory relating to large consignments may be used and samples of size 4 may be used. For convenience, samples of size 4 can be used for all sample sizes.

Results can also be obtained for proportions of N different from 50%. For example, in Table 7, $R \geq 0.8N$ is chosen. The probability that $R \geq 0.8N$ is given for sample sizes 1, 2, 3 and 4 in which all units contain drugs ($\alpha = \beta = 1$).

The results in Tables 6 and 7 are for the special circumstances in which $\alpha = \beta = 1$ and all units inspected contain drugs. More general results can be obtained. The problem is then to choose m such that, given n, α, and β, (and possible values for z, consequential on the choice of m and the outcome of the inspection), a value for y can be determined to satisfy some probabilistic criterion, e.g., the value y_0 such that $Pr(Y \geq y_0 \mid m, n, z, \alpha, \beta) = p$. Some results are given in Table 8 for $p = 0.9$, where the consignment size N is taken to be 30. Note from the last two rows that if one or two of the six inspected units do not contain drugs then the number of units in the remainder of the consignment which can be said, with probability 0.9, to contain drugs drops from 17 to 12 to 9. Note also that even if 16 units (out of 30) are inspected and all are found to contain drugs, then it can only be said, with probability 0.9, that 12 of the remaining 14 contain drugs (and this is so even with $\alpha = 4$, $\beta = 1$).

Summary

The following summary of the main results of the paper is phrased in the context of inspecting a consignment of drugs. However, the ideas expressed in the paper are just as applicable to other forensic contexts, such as the inspection of computer disks for pornographic images. In such a situation sampling may be beneficial as it exposes the investigators to as little stress as possible.

For small consignments, the beta-binomial distribution provides a probability distribution for Y and hence for the total number of units R (and hence the proportion of units) in the consignment which contain drugs. For large consignments, the beta distribution is used. These probability distributions can then be used to make inferences for the number of units in a consignment which contain drugs. There are no problems of interpretation as the uncertainty

TABLE 6—*For consignments of size* N *in which* R*(≤* N *and unknown) units contain drugs, the probability that* R ≥ N*/2 is given for various sample sizes* m*, in which all units in the sample contain drugs. (α = β = 1.)*

	m			
N	1	2	3	4
10	0.818	0.939	0.984	0.998
20	0.786	0.910	0.965	0.988
30	0.774	0.899	0.957	0.982
40	0.768	0.893	0.952	0.979
50	0.744	0.890	0.949	0.972

TABLE 7—*For consignments of size* N *in which* R *(≤* N *and unknown) units contain drugs, the probability that* R ≥ *0.8*N *is given for various sample sizes* m *in which all units contain drugs. (α = β = 1.)*

		m			
N	0.8*N*	1	2	3	4
10	8	0.491	0.661	0.788	0.879
20	16	0.429	0.579	0.696	0.785
30	24	0.406	0.550	0.662	0.750
40	32	0.395	0.535	0.645	0.731
50	40	0.388	0.526	0.634	0.720

TABLE 8—*Determination of the sample size required from a consignment of 30 units, to satisfy certain criteria. Parameters* α *and* β *are representative of prior beliefs about the proportion of units which contain drugs. The number of units inspected equals* m*, the number of those which contain drugs is* z*. The number of units not inspected is* n *(equals 30 −* m*).* y_0 *is the largest number of those units not inspected for which it can be said that 'the probability is 0.9 or greater that* y_0 *or more units contain drugs'.* Pr(Y = n) *is the probability that all the units not inspected contain drugs.*

α	β	*m*	*z*	*n*	y_0	$Pr(Y = n)$
1	1	4	4	26	16	0.16
2	1	4	4	26	17	0.19
3	1	4	4	26	18	0.21
4	1	4	4	26	19	0.24
5	1	4	4	26	20	0.26
6	1	4	4	26	20	0.28
4	2	4	4	26	16	0.06
6	2	4	4	26	17	0.08
10	2	4	4	26	19	0.13
0.5	0.5	4	4	26	19	0.38
1	1	6	6	24	17	0.23
1	1	8	8	22	16	0.29
1	1	10	10	20	16	0.35
1	1	12	12	18	15	0.42
1	1	14	14	16	13	0.48
1	1	16	16	14	12	0.55
4	1	16	16	14	12	0.59
10	1	4	4	26	22	0.35
1	1	6	5	24	12	0.05
1	1	6	4	24	9	0.01

concerning the number of units in the consignment which contain drugs may be expressed by a probability distribution. The cost, however, lies in the choice of the parameters α and β. This choice is made subjectively. If a small change in their values leads to a large change in the outcomes then considerable care has to be exercised in that choice so that it may be fully justified. However, if this were the case, this would be indicative that little information had been gained from inspection of the consignment. Care would be needed in the interpretation, regardless of the statistical input to the investigation.

For large consignments, once choices of α, β, θ_0 and p have been made, it is a simple matter to determine the value of m which provides a solution for (4).

For small consignments the inferences which can be made are illustrated in Tables 6 and 7 (for $\alpha = \beta = 1$) and in Table 8 for other values of α and β. The general formula is given by (5).

Finally, consideration has to be given to the effect of the discovery that some units in the inspected sample do not contain drugs. Such a discovery can have quite an effect on $Pr(Y = n)$ and on y_0 as illustrated in Table 8. Further units can be inspected in a sequential process. An example for large consignments has been given.

Appendix 1

Derivation of the Beta-Binomial Distribution—A sample of size m is taken from a consignment which contains $m + n$ units ($m + n = N$). Let θ be the proportion of the total number of units in the consignment which contain drugs and let Z be the number of units in the sample of size m which contain drugs. Then

$$Pr(Z = z \mid m, \theta) = \binom{m}{z} \theta^z (1 - \theta)^{m-z}, \qquad z = 0, 1, \ldots, m,$$

and

$$\binom{m}{z} = \frac{m!}{z!(m - z)!}$$

is the binomial coefficient, where the ! notation denotes factorials (e.g., for integer x, $x! = x(x - 1)(x - 2)\ldots1$) and the distribution for Z is a binomial distribution.

The binomial distribution, written as a function of θ, is the likelihood function (1) which can then be combined with a prior beta distribution (2) for θ to give a posterior beta distribution for θ. This follows from Bayes' Theorem which states that

$$f(\theta \mid z, m, \alpha, \beta) \propto Pr(Z = z \mid \theta, m) \times f(\theta \mid \alpha, \beta).$$

The posterior beta distribution for $(\theta \mid z, m, \alpha, \beta)$ can be shown to be

$$f(\theta \mid z, m, \alpha, \beta) = \theta^{z+\alpha-1}(1 - \theta)^{m-z+\beta-1}/B(z + \alpha, m - z + \beta), \qquad 0 < \theta < 1.$$

However, for small consignments, the most interesting quantity is the number Y of units in the remaining n unexamined units which contain drugs. The distribution of $(Y \mid n, \theta)$ is binomial. When this distribution is combined with the posterior beta distribution for θ the resulting distribution is known as a beta-binomial distribution (10).

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

($y = 0,1, \ldots, n$), expression (5).

Appendix 2

Similarity of Results from a Beta-Binomial Distribution with $\alpha = \beta = 1$ and a Hypergeometric Distribution.

First, define

$$\binom{-a}{k} \text{ to be equal to } \frac{(-a)(-a - 1)\ldots(-a - k + 1)}{k!}.$$

This in turn is equal to

$$\frac{(-1)^{a+k-1}(a + k - 1)!}{(-1)^{a-1}(a - 1)!k!} = (-1)^k \binom{a+k-1}{k},$$

(12).

Then, by comparing the coefficients of t^k in the two sides of the equation

$$(1 + t)^{-a}(1 + t)^{-b} = (1 + t)^{-a-b}$$

it can be shown that

$$\sum_{j=0}^{k} \binom{a+k-j-1}{k-j}\binom{b+j-1}{j} = \binom{a+b+k-1}{k},$$

(13). With a suitable change of notation this result can be written as

$$\sum_{k=z}^{n+z} \binom{k}{z}\binom{m+n-k}{m-z} = \binom{m+n+1}{n}.$$

Consider the beta-binomial distribution (5) with $\alpha = \beta = 1$. Then it can be shown that

$$Pr(Y = y \mid m, n, z, 1, 1) = \frac{(m + 1)\binom{m}{z}\binom{n}{y}}{(m + n + 1)\binom{m+n}{z+y}}, \qquad (7)$$

for $y = 0,1,2, \ldots, n$.

Let R be the total number of units which contain drugs in a consignment of size N. Thus R takes a value in $\{0,1,2, \ldots, N\}$. A uniform prior distribution for R assigns equal probability $1/(N + 1)$ to each of these $(N + 1)$ integers.

$$Pr(R = r \mid N) = 1/(N + 1).$$

The distribution of Z, the number of inspected units which contain drugs, given m, n and R, is hypergeometric. For ease of notation, let $N = m + n$ and $R = Z + Y$. The distribution of Y, given m, n and z and given the uniform prior for R can be written as

$$\begin{aligned} Pr(Y = y \mid m, n, z) &= Pr(Y + z = r \mid m, n, z) \\ &= \frac{Pr(Z = z \mid m, n, r)Pr(R = r \mid N)}{Pr(Z = z \mid m, n)} \\ &= \frac{Pr(Z = z \mid m, n, r)Pr(R = r \mid N)}{\sum_{k=z}^{n+z} Pr(Z = z \mid m, n, k)Pr(R = k \mid N)} \\ &= \frac{\binom{z+y}{z}\binom{m+n-z-y}{m-z}}{\sum_{k=z}^{n+z}\binom{k}{z}\binom{m+n-k}{m-z}} \\ &= \frac{\binom{z+y}{z}\binom{m+n-z-y}{m-z}}{\binom{m+n+1}{n}} \\ &= \frac{(m + 1)\binom{m}{z}\binom{n}{y}}{(m + n + 1)\binom{m+n}{z+y}}. \end{aligned}$$

for $y = 0,1,2, \ldots, n$, which equals the beta-binomial probability (7).

Acknowledgment

I wish to thank Dr. Freda Kemp for helpful comments concerning the beta-binomial and hypergeometric distributions.

References

1. Barnett V. Comparative statistical inference. 2nd edition. Chichester, John Wiley and Sons Ltd., 1982;37.
2. Bayes T. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London for 1763, 1764;53:370–418. Reprinted with Barnard (1958) in Pearson and Kendall 1970;131–53.
3. Barnard GA, Bayes T. A biographical note (together with a reprinting of Bayes, 1764), Biometrika 1958;45:293–315. Reprinted in Pearson and Kendall 1970;131–53.
4. Pearson EG, Kendall MG, editors. Studies in the history of statistics and probability. London: Charles Griffin, 1970.
5. Curran JM, Triggs CM, Buckleton J. Sampling in forensic comparison problems. Science and Justice 1998;38:101–7.
6. Tzidony D, Ravreby M. A statistical approach to drug sampling: a case study. J Forensic Sci 1992;37:1541–9.
7. Frank RS, Hinkley SW, Hoffman CG. Representative sampling of drug seizures in multiple containers. J Forensic Sci 1991;36:350–7.
8. Colón M, Rodriguez G, Diaz RO. Representative sampling of 'street' drug exhibits. J Forensic Sci 1993;38:641–8.
9. Cochran WG. Sampling techniques, 3rd edition. Chichester: John Wiley and Sons Ltd., 1977.
10. Bernardo JM, Smith AFM. Bayesian Theory. Chichester, John Wiley and Sons Ltd., 1994;117.
11. Venables WM, Ripley BD. Modern applied statistics with S-Plus. 2nd edition. New York, Springer Verlag, 1997.
12. Johnson NL, Kotz S, Kemp AW. Univariate discrete distributions. 2nd edition. Chichester, John Wiley and Sons Ltd., 1992;3.
13. Feller WF. An introduction to probability theory and its applications. 3rd edition, New York, John Wiley and Sons Ltd., 1968;1:65.
14. Todhunter I. A history of the mathematical theory of probability. Cambridge and London, Macmillan and Co., 1865. Reprinted 1965;454–7. New York, Chelsea.

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

ERRATUM

Erratum/Correction of Aitken CGG, Sampling—How Big a Sample? J Forensic Sci 1999 Jul;44(4):750–60.

On page 750, in the second column, second paragraph.

An alternative approach, based on the binomial distribution, is discussed in (6). Consider a specific value . . . . .

should read:

An alternative approach, based on the binomial distribution, is discussed in (5). Consider a specific value . . . . .

The Journal regrets this error. Note: Any and all future citations of the above-referenced paper should read: Aitken CGG. Sampling—How Big a Sample? [published erratum appears in J Forensic Sci 2000 May;45(3)] Forensic Sci 1999 Jul;44(4):750–60.