J Forensic Sci, January 2009, Vol. 54, No. 1
doi: 10.1111/j.1556-4029.2008.00920.x
Available online at: www.blackwell-synergy.com

Dimitris Mavridis,[1] *Ph.D. and Colin G.G. Aitken,*[1] *Ph.D.*

Sample Size Determination for Categorical Responses

ABSTRACT: Procedures are reviewed and recommendations made for the choice of the size of a sample to estimate the characteristics (sometimes known as parameters) of a population consisting of discrete items which may belong to one and only one of a number of categories with examples drawn from forensic science. Four sampling procedures are described for binary responses, where the number of possible categories is only two, e.g., licit or illicit pills. One is based on priors informed from historical data. The other three are sequential. The first of these is a sequential probability ratio test with a stopping rule derived by controlling the probabilities of type 1 and type 2 errors. The second is a sequential variation of a procedure based on the predictive distribution of the data yet to be inspected and the distribution of the data that have been inspected, with a stopping rule determined by a prespecified threshold on the probability of a wrong decision. The third is a two-sided sequential criterion which stops sampling when one of two competitive hypotheses has a probability of being accepted which is larger than another prespecified threshold. The fifth procedure extends the ideas developed for binary responses to multinomial responses where the number of possible categories (e.g., types of drug or types of glass) may be more than two. The procedure is sequential and recommends stopping when the joint probability interval or ellipsoid for the estimates of the proportions is less than a given threshold in size. For trinomial data this last procedure is illustrated with a ternary diagram with an ellipse formed around the sample proportions. There is a straightforward generalization of this approach to multinomial populations with more than three categories. A conclusion provides recommendations for sampling procedures in various contexts.

KEYWORDS: forensic science, sample size, evidence evaluation, likelihood ratio, ternary diagram, multinomial data, misleading evidence, power priors

Sample size determination (SSD) is a crucial aspect of any experimental design and there have been a number of papers addressing this subject both from a frequentist and a Bayesian approach. A review of the subject up to the mid-1990's can be found in Ref. (1) and references therein. Most examples in the literature come from medical studies and from the quality assessment of products. Another field where SSD may play a crucial role in the saving of resources is in forensic analysis. For instance, there may be a consignment of discrete units with certain proportions containing illegal materials of different types. Such units may be pills (which may be drugs, possibly of more than one type), CDs or pornographic computer files. The traditional approach to SSD from a frequentist perspective is to control some aspects of the sampling distributions of the statistics that are used for drawing inference and to define null and alternative hypotheses for the value of the characteristic of interest (e.g., proportion of pills of a certain type). The sample size is then determined by controlling the probabilities of type 1 and type 2 errors, respectively, the probabilities of rejecting the null hypothesis (e.g., that the proportion of pills is less than a certain value) when it is true and of not rejecting the null hypothesis when it is false.

Emphasis is given here on the use of Bayesian methodology in which inferences are made directly about the characteristic of interest which is categorical. The characteristic, conventionally denoted θ, is considered to be random and to have an associated probability distribution in some relevant population from which all relevant information about θ may be obtained. Such information can include the mean, the variance and distributional results such that the probability that θ is greater than a certain value, for example, may be determined. An extension to consider quantities is described in Refs (2,3).

It is common in forensic analysis to encounter a consignment of discrete units, some of which may contain illegal material. Examples of such units are pills, some of which may be illicit, CDs, some of which may be pirated, or computer files, some of which may be pornographic. For illicit drugs in pills there may be two or more mutually exclusive categories for classification (e.g., powder cocaine, crack cocaine, heroin, LSD, and marijuana). Consider a sample of known size, n say, taken from the consignment. When there are only two mutually exclusive categories, such as licit and illicit, a common distribution associated with the number of pills in one of the categories, conditional on the total number of pills in the sample, is the *binomial* distribution. When there are more than two mutually exclusive categories, the analogous distribution for the number of pills in each of the categories is known as a *multinomial* distribution. Izenman points out that inaccuracies may occur when the whole seizure is being analyzed due to time and manpower constraints (4). Furthermore, he argues that certain chemical testing destroys the evidence and that evidence may need to be shown to the jury or given to the defense to make their own testing. Also forensic scientists may be exposed to potential health hazards through airborne dust or physical contact. Thus, for various reasons, as little analysis as possible is desirable so a sample is analyzed rather than every member of the consignment. Criteria are required in order to give meaning to the phrase "as little...as possible."

Many simple approaches to the determination of a sample size have been adopted. These approaches include choosing the sample size to be the size of the square root of the size of the whole consignment. Other rules are to analyze a number equal to half the square root of the size of the whole consignment, or equal to a certain proportion, such as 10%, of the size of the whole consignment. These methods, although simple to remember, have little or

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

Received 24 Mar. 2007; and in revised form 31 May 2008; accepted 8 Aug. 2008.

© 2008 American Academy of Forensic Sciences

no statistical justification and may lead to the inspection of quantities of pills considerably in excess of those required for an inference to be made that is sufficient for legal purposes. Methods based on the beta and binomial distributions for samples in which all sampled items are illicit are described in Ref. (5) and extended here to sequential sampling. Samples contain items of more than one category. For binary models these categories could be licit and illicit in drug cases and four models are described in this context.

A fifth model is described for samples with more than two categories. The method described here assumes that the number of categories is known. The purpose of the sampling is to estimate the proportions of each category in the consignment. Further work is required in order to develop a sampling protocol for situations in which the number of categories is unknown. The example described here considers three categories for pills in a drug case where the categories are licit, ecstasy, and LSD. Other examples include:

- a mixture of glass fragments of a known number of categories;
- an autosomal locus with a known number of different alleles;
- soil which is a mixture of several different soil types;
- a pollen composition which is a mixture of several different types.

In each example, the purpose of the sampling is to estimate the proportion of each type or category.

Binomial Sampling

Consider circumstances in which a sample of size n items are taken at random and without replacement from a large population of size m. Items may be assigned to one and only one of two categories, conventionally known as "success" and "failure." As mentioned above, examples include illicit or licit drugs, pirated or legal CDs, pornographic or nonpornographic computer files.

Consider the case of illicit or licit drugs. In some jurisdictions, the numbers of pills seized is a contributory factor in the determination of the defendant's sentence and hence accurate estimation of these numbers, with knowledge of the associated uncertainty in the estimation, is important. These numbers may be obtained from estimates of the proportions of drugs in each category by multiplication of these estimated proportions by the consignment size. The procedures described here consider estimation of proportions as this is statistically the best way to proceed. It is straightforward to transform the results into numbers in a consignment.

Denote the proportion of items in the population that are categorized as successes (or, more briefly, known as "successes") as θ. In the context of illicit and licit drugs, the proportion of pills that are illicit is of interest so a "success" would be an illicit tablet. This information about the population is then translated to record that for an item drawn at random from the population (pill from the consignment), the probability it is a success (illicit) is θ. The population is deemed to be sufficiently large relative to the sample size that the probabilities for successive drawings without replacement from the sample to be successes may be treated as constant and equal to θ; i.e., sampling is taken to be equivalent to sampling with replacement. For small consignments, analyses using the beta–binomial distribution are appropriate and described in the section on "Predictive sample size determination." Let X denote the phrase "the number of drawings from a sample of size n that are successes" (the number of pills in the sample that are illicit) and let x be the symbol denoting the number of successes (illicit pills) in a sample of size n. Thus, the phrase "the probability the number of drawings from a sample of size n that are successes equals x" may be written symbolically as $\Pr(X = x)$. When there are two, and only two categories for a population into which an item may be placed, with probabilities θ and $(1 - \theta)$, respectively, then the number of successes, X, in a sample of size n, has a binomial distribution

$$\Pr(X = x|n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{(n-x)}, \quad 0 < \theta < 1,\ x = 0, 1, \ldots, n$$

The vertical bar | denotes conditioning in that symbols to the right of the bar are taken to be known. Here these are n, the sample size and θ, the probability of a success. The expression to the left of the bar is taken to be unknown and the expression whose probability it is desired to determine. Thus, the probability statement concerns the probability the number of successes equals x, conditional on (or "given") n, the sample size and θ, the probability of a success. Sometimes, as in the discussion of power priors, this function is expressed as a function of θ and it is then known as a *likelihood*

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

The sample proportion $\hat{\theta} = x/n$ provides a good estimate of θ. The variance of $\hat{\theta}$ including a so-called "finite population correction" $(m - n)/(m - 1)$ is

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

(6). As an example of the use of the finite population correction consider the example where the sample is the whole population so that $n = m$. Then the sample proportion is the population proportion and there is no uncertainty; the variance is zero which is the result given from the expression of the variance using the finite population correction. If the sampling fraction n/m is low the finite population correction $(m - n)/(m - 1) = 1 - (n - 1)/(m - 1)$ can be ignored. Assume that the sample proportion is asymptotically Normally distributed. Then

$$\hat{\theta} \sim N\left(\theta, \frac{\theta(1-\theta)}{n}\right) \tag{2}$$

One criterion for the choice of sample size in such a context is that there should be $100(1 - \alpha)\%$ confidence that the sample proportion lies within an interval of desired length $2d$ of the true proportion θ. Then

$$z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}} \leq d$$

and hence $n \geq z^2_{\alpha/2}\theta(1 - \theta)/d^2$ where $z_{\alpha/2}$ is the $100(1 - \alpha/2)\%$ point of the standard Normal distribution. For example, when $\alpha = 0.05$, $z_{\alpha/2} = 1.96$, the 97.5% point of the standard Normal distribution. As θ is not known in advance there are two courses of action. One is to use a prior subjective estimate for θ. The other is to use the value of θ for which $\theta(1 - \theta)$ is a maximum which is when $\theta = 0.5$. This latter choice leads to the rule

$$n \geq \frac{z^2_{\alpha/2}}{4d^2} \tag{3}$$

which is conservative in that it gives the largest sample size necessary to satisfy the criterion.

Thus, for $\alpha = 0.05$ and $d = 0.01$, the sample size n should be greater than or equal to $1.96^2/(4 \times 0.0001) = 3.84/0.0004 =$

9600; i.e., to obtain an estimate of the true proportion in a category to within 0.01 of the true proportion, with 95% confidence, a sample of size 9600 items is needed. This is a large sample but also a stringent criterion. To obtain an estimate of the true proportion in a category to within 0.1 of the true proportion, with 95% confidence, a sample of size 96 items is needed. The sample size has to be increased by a factor of 100 to narrow the width of the interval by a factor of 10. As an example, for consignments of CDs, with 95% confidence, an estimate could be given to within 0.1 of the true proportion of pirated CDs in a consignment if 96 were examined. In practice, it is suggested 100 CDs be examined.

In a Bayesian paradigm, uncertainty in parameter estimation is modeled with probability distributions for the parameters of interest. The beta distribution is commonly chosen to represent uncertainty about the parameter θ and details are given in the Appendix. This distribution is a so-called *conjugate prior distribution* for the binomial distribution in that the posterior distribution is also a beta distribution, but with different parameters.

Let uncertainty about θ, the probability of success for an item drawn in a sample of size n from a population, be represented with a beta distribution beta(v_1,v_2). The number of successes, X, in the sample has a binomial distribution. The combination of the beta prior and binomial distribution gives a posterior distribution beta($v_1 + x$, $n - x + v_2$) for θ which is also a beta distribution, as a result of the property of conjugacy. The probability density function is

$$f(\theta|v_1+x, v_2+n-x) = \frac{\Gamma(n+v_1+v_2)}{\Gamma(v_1+x)\Gamma(v_2+n-x)}\theta^{v_1+x-1}(1-\theta)^{n-x+v_2-1}; \quad v_1>0,\ v_2>0,\ 0<\theta<1,\ x=0,1,\ldots,n \tag{4}$$

The Bayesian approach provides answers to the questions of interest of the forensic scientist in that it provides probabilities for the uncertainties about the probabilities of success (proportions of illicit drugs, proportions of pirated CDs, or pornographic files).

This is in contrast to the frequentist approach in which confidence limits are provided. Confidence limits are limits which apply in the long run in that if identical conditions apply many times then these limits will contain the true answer a certain proportion of the time; no statement is made about the particular occasion under inspection. Thus, as stated above, for consignments of CDs, the 95% confidence limits are such that in 95% of cases in which CDs are examined, the proportion of pirated CDs in a sample of size 96 will be within 0.1 of the true proportion of pirated CDs in the whole consignment.

Likelihood Principle

A method that has been widely used for evaluating statistical evidence for one hypothesis versus another is the likelihood ratio. The term "likelihood ratio" is used because reference is made to "how likely the data x are if one hypothesis (denoted H_2 say) is true relative to how likely the data are if another hypothesis (denoted H_1 say) is true." Hacking (7) defined the *likelihood law* as

If one hypothesis, H_1, implies that a random variable X takes the value x with probability $f(x|H_1)$, while another hypothesis, H_2, implies that the probability is $f(x|H_2)$, then the observation $X = x$ is evidence supporting H_2 over H_1 if $f(x|H_2) > f(x|H_1)$, and the likelihood ratio, LR,

$$\frac{f(x|H_2)}{f(x|H_1)}$$

measures the strength of that evidence.

This definition provides a common approach to the evaluation of evidence in forensic science when H_2 is taken as the prosecution proposition, H_1 as the defense proposition, and x as the evidence that is being evaluated (8). Note that in evidence evaluation the term *proposition* is preferred to the term *hypothesis*, because of the statistical frequentist connotations of the latter term. The term "proposition" will be used from now on as far as is appropriate.

The LR may also be given as the ratio of the posterior odds in favor of H_2 to the prior odds in favor of H_2:

$$\mathrm{LR} = \frac{f(x|H_2)}{f(x|H_1)} = \frac{f(H_2|x)/f(H_1|x)}{f(H_2)/f(H_1)} \tag{5}$$

where $f(H_1)$ and $f(H_2)$ are the prior probabilities, the probabilities that H_1 and H_2 are true, respectively, prior to the conduct of the experiment, their ratio is the prior odds in favor of H_2, $f(H_1|x)$ and $f(H_2|x)$ are the posterior probabilities for propositions H_1 and H_2, and their ratio is the posterior odds in favor of H_2. The expression $f(H_2|x)$, for example, may be read as the probability H_2 is true, given x successes out of a sample of size n. The *LR* is the factor which converts prior odds into posterior odds. An interpretation of the likelihood ratio is to say that the evidence is so many times more likely if H_2 is true than if H_1 is true.

The LR is non-negative and can take values greater than 1. An LR equal to one indicates that the evidence is equally probable under either of the two competing propositions. Values of LR greater than one indicate that the evidence supports H_2 over H_1 and values smaller than one favor H_1 over H_2. For ease of interpretation and for the better understanding of the strength of the evidence, the possible values of the LR can be divided into regions to indicate the differing strengths of the evidence. Therefore, it may be taken that LRs close to one represent weak evidence and that LRs greater than some threshold t ($t > 1$) or less than t^{-1} represent moderate or strong evidence in favor of H_2 or H_1, respectively, according to some numerical criterion for t. Thresholds ($t = 8$ and $t = 32$) with conventional descriptions "weak," "moderate," and "strong" have been suggested by Royall (9) such that

- weak evidence
  - for H_2 over H_1: $1 \le \mathrm{LR} < 8$,
  - for H_1 over H_2: $1/8 < \mathrm{LR} \le 1$;
- moderate evidence
  - for H_2 over H_1: $8 \le \mathrm{LR} < 32$,
  - for H_1 over H_2: $1/32 < \mathrm{LR} \le 1/8$;
- strong evidence
  - for H_2 over H_1: $\mathrm{LR} \ge 32$,
  - for H_1 over H_2: $\mathrm{LR} \le 1/32$.

There is a nonzero probability that the likelihood ratio may yield strong evidence supporting H_2 over H_1 when, in fact, H_1 is correct, or vice versa. The wrong proposition is then accepted. In such a case the evidence is known as misleading evidence. A probabilistic limit on this situation is described.

Probability of Accepting the Wrong Proposition

Consider two propositions of interest H_1 and H_2 concerning a binomial characteristic, θ say, such that H_1:$\theta \le \theta_l$ and H_2:$\theta \ge \theta_u$. This characteristic could be the proportion of illicit drugs in a

consignment (and by multiplication, the total number in the consignment) with interest being in the value for sentencing purposes. A binomial experiment (i.e., one with a fixed number of items with two possible categories into which items can be assigned, with a constant assignation probability for each trial and such that each trial is independent of all other trials) is conducted and x successes are observed out of n trials. The LR is given by $f(x|H_2)/f(x|H_1)$ where f is the probability function of the binomial distribution. The probability of H_2 being accepted is

$$f(H_2|x) = \frac{f(x|H_2)f(H_2)}{f(x|H_1)f(H_1) + f(x|H_2)f(H_2)} \quad (6)$$

which can be rewritten as

$$f(H_2|x) = \frac{\mathrm{LR}f(H_2)}{f(H_1) + \mathrm{LR}f(H_2)} \quad (7)$$

by dividing the numerator and denominator of the right-hand side by $f(x|H_1)$. By further dividing both the numerator and denominator of the right-hand side by $f(H_1)$,

$$f(H_2|x) = \frac{\mathrm{LR}\frac{f(H_2)}{f(H_1)}}{1 + \mathrm{LR}\frac{f(H_2)}{f(H_1)}} \quad (8)$$

There are two and only two propositions. If H_2 is not true then H_1 is true. Thus, $f(H_2|x) + f(H_1|x) = 1$ and hence

$$f(H_1|x) = \frac{1}{1 + \mathrm{LR}\frac{f(H_2)}{f(H_1)}} \quad (9)$$

It is evident from Eqs (8) and (9) that the probability of the truth of either of the two propositions given the data (x successes out of n items sampled) is a function of the likelihood ratio and the prior odds $f(H_2)/f(H_1)$ in favor of H_2.

If both competing propositions are equiprobable *a priori* $\left(f(H_1) = f(H_2) = \frac{1}{2}\right)$, then

$$f(H_2|x) = \frac{\mathrm{LR}}{\mathrm{LR} + 1}$$

and

$$f(H_1|x) = \frac{1}{\mathrm{LR} + 1}$$

It can be deduced that for any given constant $k > 0$, $\Pr(\mathrm{LR} \geq k|H_1 \text{ is true}) \leq 1/k$; i.e., the probability of evidence that supports H_2 with an $\mathrm{LR} \geq k$ when H_1 is true is $\leq 1/k$. Evidence that supports H_2 when H_1 is true is misleading. Consider S to be the set of values of x that produce a value of the LR in favor of H_2 versus H_1 of at least k. For $x \in S$, $f(x|H_2)/f(x|H_1) \geq k$ and hence $f(x|H_1) \leq f(x|H_2)/k$, where $\in$ is read as "is a member of." Then $\Pr(S) = \sum_{x \in S} f(x|H_1) \leq \sum_{x \in S} f(x|H_2)/k \leq 1/k$. Equation (7) enables the formation of a scale of evidence in favor of one proposition or the other. For instance, sampling could be stopped when the probability that one proposition is correct, given the sampled data, is above some predefined threshold.

Sequential Sampling

Sequential analysis was developed during the Second World War (10) mainly because war production and development required results as quickly as possible. In sequential sampling a consignment (of pills, for example) is inspected, usually one item at a time (but sometimes in small batches). After inspection of each item (or small batch) a decision is made as to whether to continue sampling or to terminate the process. Sampling is terminated when the cumulative sample contains enough information to make a decision based on some prespecified probabilistic criterion. Analysis happens as the data are collected in contrast with sampling plans where statistical analysis is conducted after a sample of a size fixed in advance has been collected; see Eq. (3) for an example of a sample size fixed in advance. Sequential sampling is best used when the emphasis is on decision making and there are well-defined propositions about which decisions can be made. The methods used to make decisions from sequential sampling are called stopping rules. The accumulated data are analyzed at each step to see if one of the stopping rules has been attained and hence sampling may stop, otherwise sampling is continued.

Large values (>1) of the likelihood ratio constitute statistical evidence in favor of one proposition whereas small values (<1) are supportive of the other proposition. The likelihood ratio may be computed sequentially as data are inspected. In such a case, sampling is stopped when enough data have been collected to support one of the competing propositions in the sense that the LR is greater than a threshold t, (e.g., $t = 32$) or smaller than t^{-1} (e.g., 1/32). A well-known test that distinguishes between two competing propositions by using the likelihood ratio and controlling the probabilities of type 1 and type 2 errors is the sequential probability ratio test (SPRT) (10).

Sequential Probability Ratio Test

Suppose that there are two competing propositions H_1 and H_2 for the value of the parameter θ, where θ is the proportion of items in the population falling into a certain category. For example, H_1 could be that $\theta = \theta_1$ and H_2 could be that $\theta = \theta_2$, respectively.

As an example consider a seizure of 5000 pills. The exact size of the seizure is not important for determination of proportions except that it must be sufficiently large that sampling of the pills may be considered to be with replacement; i.e., the proportions of licit and illicit pills remain effectively unchanged with the removal of a few pills from the consignment. The exact size of the seizure is relevant when estimates of the absolute numbers of licit and illicit pills are required, for example when sentencing. Assume there are three levels of criminality associated with the seizure, other than the one of innocence in which no pills are illicit. These levels depend on the proportion θ of illicit pills in the seizure and are defined by $0 < \theta \leq 0.2$, $0.2 < \theta < 0.6$ and $\theta \geq 0.6$. Therefore, the propositions being tested are $H_1{:}\theta \leq 0.2$ versus $H_2{:}\theta \geq 0.6$. The error probabilities are set as $\alpha = 0.01$ (probability of a type 1 error, rejecting $H_1[\theta \leq 0.2]$ when H_1 is true) and $\beta = 0.1$ (probability of a type 2 error, not rejecting H_1 when it is false). It is considered more serious to convict an innocent person (i.e., increase the likelihood of such a verdict by deciding the proportion of illicit pills is larger than it actually is) than to fail to convict a guilty person (i.e., decrease the likelihood of a conviction by deciding the proportion of illicit pills is less than it actually is). In this context this would suggest it is more serious to decide $\theta \geq 0.6$ when in fact $\theta \leq 0.2$ than vice versa. Hence the probability of the former error is set at a value a factor of 10 lower than the probability of the latter error. The exact values of α and β chosen are a matter of subjective judgment based on consideration of the consequences of incorrect decisions. A numerical solution is given after the mathematical principles are explained. Notice also that inequalities are given here for θ while the theory is developed for exact values for θ. The results developed for the exact values for θ (0.2 and 0.6 in this example) are conservative for the inequalities in the sense that

the sample sizes derived using the exact values for θ will give error probabilities no greater than those specified for the inequalities.

Each item sampled is inspected immediately after collection. Following such an inspection a decision has to be made as to whether one of the propositions should be treated as being true (accepted) or sampling should be continued. Sampling is stopped when enough information has been accumulated to accept one of the competing propositions. Another possibility is that limitations on resources are such that it is not possible to continue sampling. Sampling is then stopped with the conclusion that there is insufficient evidence to choose between the two propositions. Observations are assumed to be independent and are denoted by x_i ($i = 1,\ldots,n$) where i is the number of the sampling unit. For the consignment of pills, an observation is the licit or illicit nature of the inspected pill and x_i is set equal to 1 if the pill is illicit and to 0 if the pill is not illicit. Similar notation may be used for pirated CDs or pornographic computer files or any other similar contexts. The probability of observing a sequence $\mathbf{x} = (x_1,x_2,\ldots,x_n)$ (of zeros and ones) assuming H_1 to be true is

$$f(\mathbf{x}|H_1) = f(x_1|H_1)\cdots f(x_n|H_1) \tag{10}$$

and, assuming H_2 to be true, is

$$f(\mathbf{x}|H_2) = f(x_1|H_2)\cdots f(x_n|H_2) \tag{11}$$

where in both situations independence is assumed for the results from each sampled unit.

The likelihood ratio $f(\mathbf{x}|H_2)/f(\mathbf{x}|H_1)$ is computed after the inspection of each additional x_i. Sampling is stopped either when this ratio is very small and less than 1 (with an acceptance of H_1) or when the ratio is very large and greater than 1 (with an acceptance of H_2). Two constants, A and B, for the likelihood ratio are upper and lower limits as follows:

$$B \le \frac{f(\mathbf{x}|\theta_2)}{f(\mathbf{x}|\theta_1)} \le A \tag{12}$$

The constants A and B are determined in such a way that the probability H_1 is rejected (i.e., H_2 is accepted as being true) when H_1 is actually true is at most α and the probability that H_2 is rejected (i.e., H_1 is accepted as being true) when H_2 is actually true is at most β. The SPRT controls the probability of observing evidence that is misleading in the sense that it leads to acceptance of a certain proposition when the alternative proposition is true.

The probabilities of making a decision with respect to a pair of propositions under the set of circumstances that each of the competing two propositions in turn is correct is given in Table 1.

The process is terminated with the acceptance of H_2, if $f(\mathbf{x}|\theta_2)/f(\mathbf{x}|\theta_1) > A$. This inequality can be written as $f(\mathbf{x}|\theta_2) > A\, f(\mathbf{x}|\theta_1)$ which is equivalent to $1 - \beta > A\alpha$ when H_2 is correct and hence $A < (1 - \beta)/\alpha$. Similarly, the process is terminated, with the acceptance of H_1, if $f(\mathbf{x}|\theta_2)/f(\mathbf{x}|\theta_1) < B$ and $B > \beta/(1 - \alpha)$. As setting A and B further from one decreases the probabilities of errors, $(1 - \beta)/\alpha$ is a lower limit for A and $\beta/(1 - \alpha)$ is an upper limit for B.

Consider an example using the binomial distribution with success probability θ. Denote individual members of a sample of size n by x_i with $x_i = 1$ for a success (illicit pill, pirated CD, pornographic computer file) and $x_i = 0$ for a failure (licit pill, legal CD, nonpornographic computer file) and let $x = \sum_{i=1}^{n} x_i$ denote the total number of successes in the sample. Then the total number of successes has the binomial distribution

$$\Pr(X = x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{(n-x)}, \quad 0 < \theta < 1, x = 0,1,\ldots,n$$

Equation (12) can be analyzed further as

$$\begin{aligned} B \le \frac{f(x|\theta_2)}{f(x|\theta_1)} &\le A, \\ \frac{\beta}{1-\alpha} \le \frac{\binom{n}{x}(1-\theta_2)^{n-x}\theta_2^x}{\binom{n}{x}(1-\theta_1)^{n-x}\theta_1^x} &\le \frac{1-\beta}{\alpha}, \\ \log\frac{\beta}{1-\alpha} \le n\log\frac{1-\theta_2}{1-\theta_1} + x\log\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)} &\le \log\frac{1-\beta}{\alpha} \end{aligned} \tag{13}$$

where log denotes Napierian logarithms, i.e., logarithms to base "e." With the help of Eq. (13), the SPRT of the proposition $H_1{:}\theta = \theta_1$ versus the proposition $H_2{:}\theta = \theta_2$ with probabilities of type 1 and type 2 errors α and β, respectively, can be summarized as:

- accept H_1, if $x \le k_1 + \lambda n$;
- accept H_2, if $x \ge k_2 + \lambda n$;
- continue sampling if $k_1 + \lambda n < x < k_2 + \lambda n$ where x is the number of successes and

$$k_1 = \log\frac{\frac{\beta}{1-\alpha}}{\log\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}} \tag{14}$$

$$k_2 = \frac{\log\frac{1-\beta}{\alpha}}{\log\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}} \tag{15}$$

$$\lambda = \frac{\log\frac{1-\theta_1}{1-\theta_2}}{\log\frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)}} \tag{16}$$

This sequential test, denoted as $Q_1(\alpha, \beta)$ in Table 2, can be seen as testing the proposition $H_1{:}\theta \le \theta_1$ because if the acceptance region is attained it means strictly that $\theta \le \theta_1$ and not just that $\theta = \theta_1$. Similarly, the rejection region for H_1 corresponds to $\theta \ge \theta_2$.

For $\alpha = \beta$, the limits A and B are $t = (1 - \alpha)/\alpha$ and $t^{-1} = \alpha/(1 - \alpha)$, respectively. The SPRT is the sequential estimation of the LR until a value larger than $(1 - \alpha)/\alpha$ or smaller than $\alpha/(1 - \alpha)$ is observed.

TABLE 1—*Probabilities of accepting a certain hypothesis.*

	LR>A (Accept H_2)	LR<B (Accept H_1)
H_1 is correct	α	$1-\alpha$
H_2 is correct	$1-\beta$	β

An Application of the SPRT

Consider the seizure of 5000 pills and propositions $H_1{:}\theta \le 0.2$ and $H_2{:}\theta \ge 0.6$; with $\alpha = 0.01$, $\beta = 0.1$. As stated above, the inequalities for θ may be replaced with equalities when developing the test protocol. Insertion of the values $\theta_1 = 0.2$, $\theta_2 = 0.6$, $\alpha = 0.01$, and $\beta = 0.1$ into Eqs (14)–(16) gives values for k_1, k_2, and λ of −1.3, 2.8, and 0.4, respectively. Figure 1 illustrates the procedure. The two parallel lines represent the lower and upper thresholds ($k_1 + \lambda n$, $k_2 + \lambda n$) where

$$k_1 + \lambda n = -1.3 + 0.4n$$

$$k_2 + \lambda n = 2.8 + 0.4n$$

Suppose the first five pills inspected in a sample were found to be illicit. The line $2.8 + 0.4n$ is crossed and it can be decided to act as if H_2 is true ($\theta \ge 0.6$).

TABLE 2—*Simulation results for sequential sampling from a consignment with* m *members from a population with proportion* θ *of successes.*

θ	Criterion	Propositions Tested	Mean	% False	Min	Median	Max
0.2	$Q_1(0.01,0.01)$		130.40	0	30	125	371
	$Q_1(0.01,0.1)$		127.52	0	30	120	323
	$Q_2(0.9)$		37.63	0.7	1	8	455
	$Q_2(0.99)$	$\theta \le 0.1$ vs. $\theta \ge 0.15$	183.91	0	2	143.5	753
	$Q_3(0.1)$		300.40	0	257	273	639
	$Q_3(0.3)$		210.13	0	47	185	611
0.5	$Q_1(0.01,0.01)$		27.26	0	12	27	51
	$Q_1(0.01,0.1)$		27.00	0	13	26	65
	$Q_2(0.9)$		2.75	0	1	2	28
	$Q_2(0.99)$	$\theta \le 0.1$ vs. $\theta \ge 0.15$	7.62	0	2	5	52
	$Q_3(0.1)$		349.71	0	345	350	350
	$Q_3(0.3)$		53.42	0	9	54	54
0.5	$Q_1(0.01,0.01)$		70.32	0	23	67	167
	$Q_1(0.01,0.1)$		35.76	0	8	32	171
	$Q_2(0.9)$		26.87	1.7	2	8	300
	$Q_2(0.99)$	$\theta \le 0.6$ vs. $\theta \ge 0.7$	100.59	0.1	5	78	461
	$Q_3(0.1)$		350.02	0	345	350	350
	$Q_3(0.3)$		112.10	0.2	9	89	361
0.8	$Q_1(0.01,0.01)$		69.44	0	33	65	179
	$Q_1(0.01,0.1)$		70.25	0	30	67	176
	$Q_2(0.9)$		27.28	5.5	2	15	241
	$Q_2(0.99)$	$\theta \le 0.6$ vs. $\theta \ge 0.7$	97.85	0.3	5	86	394
	$Q_3(0.1)$		255.19	0	30	257	356
	$Q_3(0.3)$		66.05	0	9	49	332
0.12	$Q_1(0.01,0.01)$		62.50	0	5	54	261
	$Q_1(0.01,0.1)$		59.60	4	5	54	213
	$Q_2(0.9)$		107.48	0	1	10	922
	$Q_2(0.99)$	$\theta \le 0.03$ vs. $\theta \ge 0.1$	508.70	26.8	1	513	1000
	$Q_3(0.1)$		494.70	0	216	501.5	885
	$Q_3(0.3)$		464.28	0	42	500.5	898

The proportion θ is estimated by the number of successes divided by the sample size n and n is increased incrementally in steps of 1. Sampling in a particular simulation is stopped if the appropriate criterion is met, or after 1000 trials if no decision has been made under the relative criterion about the proportion of successes in the population from which the consignment has been selected. The process is repeated 1000 times. The criteria are $Q_1(\alpha,\beta)$: sequential test; stop sampling if $x \le k_1 + \lambda n$ or $x \ge k_2 + \lambda n$ where k_1,k_2,λ are given by Eqs (14)–(16). $Q_2(p)$: Normal approximation to the beta–binomial posterior distribution and sampling is stopped when one of the competing propositions is accepted with probability p from inequality Eqs (29) or (30). $Q_3(l)$: predictive sample size where l is the width of the interval given by expression (27); sampling stops when the width of the interval is less than l.

This example may be used to illustrate the result that $\Pr(\mathrm{LR} \ge k|H_1) \le 1/k$. For 5 "successes" out of 5 pills, the $\mathrm{LR} = \theta_2^5/\theta_1^5 = (0.6/0.2)^5 = 3^5 = 243$. Set $k = 243$. Then

$$\Pr(\mathrm{LR} \ge 243|H_1) = \Pr(5 \text{ successes out of } 5 \text{ pills } |\theta = 0.2) = 0.2^5 = 1/3125 < 1/243$$

Bayesian Approaches to Sample Size Determination for Binary Responses

The work of Royall (9) was extended by De Santis (11) to a Bayesian setting. De Santis used the LR and determined an appropriate sample size to be one for which there was a large probability of observing strong, correct evidence while there was a small probability of observing weak, misleading evidence. The probability of observing strong evidence is associated with the other two probabilities of observing weak and moderate evidence. As before (9), thresholds need to be set in order to determine what constitutes strong and weak evidence. The probability of accepting a certain proposition after data have been observed (Eq. [7]) may provide such thresholds or Royall's benchmarks

$$\left(8, 32, \frac{1}{8}, \frac{1}{32}\right)$$

might be used (9).

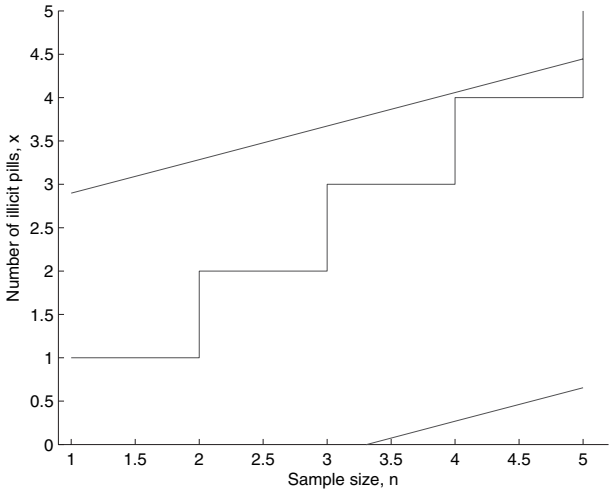

FIG. 1—*Monitoring the SPRT. Solid lines represent the lower and upper thresholds of the procedure.*

Other Bayesian approaches determine the sample size as that for which a function, such as the variance (12) of the posterior distribution of the characteristic of interest, θ, satisfies some prespecified criterion, e.g., the variance is less than a certain value. This would correspond to a requirement to estimate the characteristic to within a certain precision. Let $T(\theta|x_n)$ denote a function of the posterior distribution of θ whose performance is to be controlled. This is to be carried out by the design of an experiment that will provide a sample of size n, and x_n denotes the number of members of the sample with the characteristic, where the subscript in this context denotes the sample size. Other examples of such functions are the average posterior interquartile range, the width of the highest posterior density (HPD) interval (13) (a procedure which considers the posterior density of θ, $f(\theta|x)$, and finds the shortest interval for which the probability that θ lies in that interval is a predetermined probability, say 0.95) and the posterior probability of a certain proposition (14). Most Bayesian SSD techniques select the minimal n for chosen values of ϵ (>0) and α (>0) (significance level) that satisfy either of the two following statements

$$E[T(\theta|x_n)] \le \epsilon \tag{17}$$

or

$$\Pr[T(\theta|x_n) \notin R] \le \alpha \tag{18}$$

$$\text{equivalently } \Pr[T(\theta|x_n) \in R] \ge 1 - \alpha \tag{19}$$

for an appropriate interval R, where $\notin$ indicates "is not a member of."

Average Posterior Variance

In the examples that follow, the criterion that is used is the mean posterior variance where $T(\theta|x_n) = \mathrm{var}(\theta|x_n)$. A reason for using

that function is its simplicity both in intuitive terms as it is an analogue of Cochran's (1977) method (Eq. [2]) of determining the sample size as well as in mathematical terms as a beta conjugate prior can be used for θ leading to a beta posterior with a mathematical expression for the variance (Eq. [49]). The mean posterior variance criterion finds the minimum n for which

$$E[\text{var}(\theta|x_n)] \leq \epsilon \tag{20}$$

where ϵ is some prespecified limit.

Predictive Sample Size Determination

When the consignment size m is known it is possible to determine an appropriate sample size by estimation of the distribution of the number y of items that are illicit in the $m-n$ units not inspected (15). This is in contrast to the estimation of θ, a proportion. The reason for the contrast may be explained in the context of a super-population. The consignment may itself be considered as a sample from a larger population, known as a super-population (such as the overall output of a drug factory), within which θ denotes the proportion of items that are illicit. This proportion may be estimated from a sample from the consignment under inspection from the super-population. The super-population may be conceptually infinite, for example as the total output of the drug factory may be unknown other than it is extremely large.

The approach that estimates y directly is an alternative to consideration of properties of θ, the proportion of illicit items in a consignment and, by extension, the super-population. A beta(v_1,v_2) prior for θ is considered which yields an updated posterior distribution for θ of beta($v_1 + x$, $n - x + v_2$). The so-called predictive distribution of Y is then given by

$$\Pr(Y = y|x) = \int_0^1 \Pr(Y = y|\theta) f(\theta|x) \mathrm{d}\theta \tag{21}$$

where $f(\theta|x)$ is the posterior distribution of θ (beta($v_1 + x$, $n - x + v_2$)) and

$$\Pr(Y = y|\theta) = \binom{m-n}{y} \theta^y (1-\theta)^{m-n-y} \tag{22}$$

It can be shown that

$$\Pr(Y = y|x) = \binom{m-n}{y} \frac{B(v_1 + x + y, m - x - y + v_2)}{B(v_1 + x, n - x + v_2)}, \quad y = 0, \ldots, m-n$$

a beta–binomial distribution with parameters ($v_1 + x + y$, $m - x - y + v_2$) (5). It is necessary to work with the cumulative distribution function in order to determine probabilities that Y is greater than a certain value and hence the total size of the illicit part of the consignment is greater than a certain value. If m is large this will involve the summation of many values. It is computationally intensive but feasible with computer software packages such as MATLAB. If a suitable computer package is not available an alternative option is to use the beta distribution (5). Alternatively, the normal approximation to the beta–binomial distribution may be used (5) where the mean μ is given by

$$\mu = \frac{(m-n)(x+v_1)}{n + v_1 + v_2} \tag{23}$$

and the variance σ^2 is given by

$$\sigma^2 = \frac{(m-n)(x+v_1)(n-x+v_2)(m+v_1+v_2)}{(n+v_1+v_2)^2(n+v_1+v_2+1)} \tag{24}$$

The sample size may then be determined as the smallest n, for given m and θ, such that

$$\Pr(Y \leq cm - x_n) \geq 1 - \alpha \tag{25}$$

where $\Pr(Y \leq y_0) = \sum_{y=0}^{y_0} f(y|x_n), x_n$ is the number of illicit items, and c, $0 \leq c \leq 1$, is a prespecified threshold. Similarly, there may be interest in satisfying a criterion of the form

$$\Pr(Y \geq cm - x_n) \geq 1 - \alpha \tag{26}$$

Note that $(x_n + y)/m$ is the proportion of illicit pills in the consignment and hence $y \leq cm - x_n$ is equivalent to the proportion $(x_n + y)/m \leq c$. Hence, the above inequalities Eqs (25) and (26) denote probabilistic bounds on the sample sizes. In summary, after n and x_n have been observed and for given m, a Normal approximation may be used to determine the total number of illicit pills in the consignment, with mean and variance given by Eqs (23) and (24), respectively. Therefore, in an extreme scenario and for significance level α, either y_α or $y_{1-\alpha}$ illicit pills are found in the remaining $m - n$ trials, depending on which of the propositions Eqs (25) and (26) "Y less than a certain value" or "Y greater than a certain value" are to be tested, with $y_\alpha = \mu + z_\alpha\sigma$, $\Phi(z_\alpha) = \alpha$, $\Phi(z_{1-\alpha}) = 1 - \alpha$, where Φ denotes the cumulative distribution of the standard normal distribution and $0 \leq \alpha \leq 0.5$. Therefore the $100(1 - \alpha)\%$ interval for the proportion, θ, of illicit pills in the population is the interval

$$\left(\frac{x + \mu + z_\alpha\sigma}{m}, \frac{x + \mu + z_{1-\alpha}\sigma}{m}\right)$$

where the subscript n has been dropped from the x for ease of notation. This method can be used both for making an inference from the sample to the population and for a sequential sampling scheme where sampling is stopped when the probability interval

$$\left(\frac{x+\mu+z_{1-\alpha}\sigma}{m}, \frac{x+\mu+z_\alpha\sigma}{m}\right) \tag{27}$$

has a width less than a certain value l or when the estimated probability, $(x + y)/m$, of an illicit pill in the population falls into a prespecified interval. In simulation results reported in Table 2 the first method is used and denoted by $Q_3(l)$. There may be cases where there is interest only in rejecting one of the two competing propositions without any need for an accurate estimation of θ. A requirement to control the width of the probability interval may result in a big sample size especially if θ is close to 0.5. Alternatively, sampling may be stopped when along with the upper and lower bounds satisfying either H_1 or H_2, a specific number of sampling units has been inspected, e.g., 10 or 20.

Items are tested sequentially and stopping rules are defined. For example, a rule may be to stop if the proportion of illicit drugs in the consignment is estimated to exceed 60%. Alternatively, a rule may be to stop if the proportion in the consignment is estimated to be below 20%. For intermediate values as well as for values that lie both in the acceptance and rejection regions [i.e., $(x + \mu + z_\alpha\sigma)/m$ lies in the lower region and $(x + \mu + z_{1-\alpha}\sigma)/m$ in the upper region] sampling is continued until only one of the two criteria is satisfied, or an upper limit, e.g., 1000, is reached when it is decided to behave as if $0.2 < \theta < 0.6$.

Suppose a sample of six units ($n = 6$) from a seizure of $m = 5000$ pills is taken and there are six successes (i.e., the number x of illicit pills equals the sample size six). The beta prior is taken to be beta(1,1). The mean μ (Eq. [23]) of the posterior

beta–binomial distribution is 4369.75, a proportion 87.4% of 5000. The variance σ^2 (Eq. [24]) is 550.978^2. The significance level α is taken to be 0.01 so that $z_\alpha = -2.3263$ and $z_{1-\alpha} = 2.3263$. Then

$$y_\alpha = \mu + z_\alpha\sigma = 3088.01 \quad \text{and} \quad \frac{x + y_\alpha}{m} = 0.62 > 0.6$$

i.e., the probability that the true proportion of illicit pills is greater than 0.62 is 0.99. Sampling is stopped with the decision to act as if the seizure is contaminated to a degree larger than 60%.

Criterion for Sample Size Calculations for Proportions with Binary Responses

Return now to consideration of a population proportion rather than a number of items in a consignment. A criterion where the scientist wants to be $100p\%$ certain that at least $100l\%$ of a consignment contains drugs when all n units in the sample contain illicit drugs is provided by (5). As an example, when $p = 0.95$ and $l = 0.5$, the criterion can be written mathematically as

$$\Pr(\theta > 0.5|v_1 + n, v_2) = \frac{\int_{0.5}^{1} \theta^{n+v_1-1}(1-\theta)^{v_2-1}\,\mathrm{d}\theta}{B(n+v_1, v_2)} \geq 0.95 \quad (28)$$

A context different from that of drugs is that of the inspection of a hand for gunshot residue. A person is suspected of firing a gun. A sample of particles is taken from his hands and wrists. Sampling of particles can stop when the first particle of gunshot residue is found. The problem is to determine a number for the particles that should be sampled before stopping if no particle has been found. This number can be determined by using a criterion that the scientist wishes to be $100p\%$ certain that the probability there is no gunshot residue present is at least $100l\%$. In this context possible values for p and l are 0.95 and 0.99, say. Consider $v_1 = v_2 = 1$ (the uniform prior mentioned in the Appendix and the skeptical prior of the following section) in Eq. (28). Denote the probability that no gunshot residue is present by θ. Then the criterion may be written mathematically as

$$\Pr(\theta > 0.99|1 + n, 1) = \frac{\int_{0.99}^{1} \theta^n\,\mathrm{d}\theta}{B(n+1,1)} = (n+1)\int_{0.99}^{1} \theta^n \mathrm{d}\theta \geq 0.95$$

(Note that $(1-\theta)^{v_2-1} = (1-\theta)^0 = 1$ when $v_2 = 1$.) The sample size n is chosen as the smallest integer that satisfies this inequality. This value is determined as follows.

$$\begin{aligned}(n+1)\int_{0.99}^{1} \theta^n \mathrm{d}\theta \geq 0.95 &\Rightarrow [\theta^{n+1}]_{0.99}^{1} \geq 0.95\\ &\Rightarrow 1 - 0.99^{n+1} \geq 0.95\\ &\Rightarrow 0.99^{n+1} \leq 0.05\\ &\Rightarrow (n+1) \geq \log(0.05)/\log(0.99)\\ &\Rightarrow n \geq 297.07\end{aligned}$$

Thus if the scientist wishes to be 95% certain that the probability there is no gunshot residue present is at least 99% then just under 300 particles have to be examined. This is a very strict criterion and leads to a large sample size which may not be possible to achieve in practice. An alternative approach is to consider the inference that may be made if a fixed sample size is chosen and no particles of gunshot residue are found in that sample. For example, if the sample size n is chosen to be 10, then it can be shown that it is 50% certain that the probability no gunshot residue is present is greater than 0.94 and approximately 70% certain that the probability no gunshot residue is present is greater than 0.90.

It is suggested in Ref. (16) that a community of priors representing skeptical, enthusiastic, and weak prior beliefs should be considered in every experiment and that all three beliefs should lead to the same conclusion in order to make inference about the target population. Parameter values v_1 and v_2 need to be found for the beta distribution that will represent the three different beliefs. Such parameters can be $v_1 = 1$ and $v_2 = 1$ for the skeptical belief, $v_1 = 10$ and $v_2 = 1$ for the enthusiastic belief, and $v_1 = 1$ and $v_2 = 10$ for the weak belief. The sample sizes following this method and for $p = 0.95$ and $l = 0.5$ are 4, 1, and 18, respectively. The last two figures show that an enthusiastic prior belief requires little extra evidence to satisfy the criterion and a weak prior belief requires much extra evidence. The results also illustrate how previous knowledge can lead to variations in the sample size and hence the cost of analysis. However, such a criterion should be tested sequentially because if the first item sampled is "negative" sampling has to continue beyond these values. Application of a sequential sampling scheme enables the prior beliefs to be updated as samples are investigated.

A Bayesian Two-Sided Sequential Criterion

Suppose that there are two competing propositions $H_1{:}\theta \leq \theta_\ell$ and $H_2{:}\theta \geq \theta_u$, $(\theta_u > \theta_\ell)$. These two propositions are tested sequentially. First, some stopping rules are defined. It is decided to act as if H_1 is true if there is at least a $p_1\%$ probability that $\theta < \theta_\ell$ and to act as if H_2 is true if there is a $p_2\%$ probability that $\theta > \theta_u$. Sampling is continued until either of these rules is satisfied. Assume a beta prior, beta(v_1,v_2), for θ. The posterior distribution after the inspection of the ith sampling unit is also a beta distribution with parameters $v_1 + x_i$ and $v_2 + i - x_i$ where x_i is the number of "successes" up to the ith inspected sampling unit, and θ is the probability of a success.

Therefore, sampling is stopped, either when

$$\begin{aligned}&\Pr(\theta < \theta_\ell|v_1 + x_i, v_2 + i - x_i)\\ &\quad = \frac{\int_0^{\theta_\ell} \theta^{v_1+x_i-1}(1-\theta)^{v_2+i-x_i-1}\,\mathrm{d}\theta}{B(v_1+x_i, v_2+i-x_i)} \geq p_1 \qquad (29)\end{aligned}$$

or when

$$\begin{aligned}&\Pr(\theta > \theta_u|v_1 + x_i, v_2 + i - x_i)\\ &\quad = \frac{\int_{\theta_u}^{1} \theta^{v_1+x_i-1}(1-\theta)^{v_2+i-x_i-1}\,\mathrm{d}\theta}{B(v_1+x_i, v_2+i-x_i)} \geq p_2 \qquad (30)\end{aligned}$$

For $\theta_u = 0.5$ and $p_2 = 0.95$, Eq. (30) is equivalent to the criterion suggested in Eq. (28) when only "successes" are observed ($x_i = i$).

The Use of Historical Data for Determination of the Sample Size with Power Priors

Prior information from historical data may lead to a substantial saving of time and financial resources. The so-called *power priors* for the incorporation of information from previous studies were used in Ref. (17) to form a suitable prior for a current study.

Consider the previous example with 5000 illicit pills. There are no historical data from the suspect but there are historical data associated with the conditions under which the seizure is captured. For instance, there may be information about the location where

the seizure has been found (e.g., hidden in a boat in transit) or there may be historical data from circumstantial evidence associated with the suspect (e.g., previous convictions, illegal possession of weapons, fake transport or other papers, possession of large amounts of money). Suppose now that the seizure was caught under circumstances similar to those of a previous seizure of illicit pills in which 25 pills were analyzed and all of them were found to be illicit. These 25 pills may be used as historical data. The way in which the data may be used is explained in general and then this particular example is developed.

Previous studies should be similar to the current one in that the same likelihood should be able to be used for inference about the characteristics of interest. Suppose that the data (sample size and the number of illicit drugs in the sample) from a previous similar study are denoted by D_0. The power prior $f^{\mathrm{p}}/(\theta|D_0)$ that will be used in the current study is

$$f^{\mathrm{p}}(\theta|D_0) \propto L(\theta|D_0)^{\zeta_0} f(\theta) \tag{31}$$

where θ is the parameter of interest, superscript "p" denotes power prior, L denotes the likelihood and ζ_0 is a coefficient, between 0 and 1, weighting the effect of historical data on the current study and $f(\theta)$ is a prior before consideration of the historical data. As $\zeta_0 \to 1$ the standard posterior of $(\theta|D_0)$ is obtained whereas as $\zeta_0 \to 0$ the prior that would have been used in the absence of historical data is obtained. Intermediate values of ζ_0 are associated with different weights for historical data, the closer ζ_0 is to 1, the stronger the belief in the validity and relevance to the case in hand of the historical data.

Suppose a binomial experiment is conducted. A beta prior, beta(v_1,v_2), is considered. There is also some information from a similar experiment conducted in the past. The likelihood obtained from that previous experiment is

$$L(\theta|D_0) = \binom{n_0}{x_0} \theta^{x_0}(1-\theta)^{n_0-x_0}$$

with n denoting the sample size, x the number of successes and index 0 denoting reference to a previous study; see Eq. (1) with $(n_0,x_0) = D_0$. The power prior is a beta with parameters $(\zeta_0 x_0 + v_1, \zeta_0(n_0 - x_0) + v_2)$,

$$\begin{aligned} f^{\mathrm{p}}(\theta|D_0) &\propto L(\theta|D_0)^{\zeta_0} f(\theta) \\ &= \left(\binom{n_0}{x_0} \theta^{x_0}(1-\theta)^{n_0-x_0} \right)^{\zeta_0} \theta^{v_1-1}(1-\theta)^{v_2-1} \\ &\propto \theta^{\zeta_0 x_0 + v_1 - 1}(1-\theta)^{(n_0-x_0)\zeta_0 + v_2 - 1} \end{aligned} \tag{32}$$

Suppose that instead of one, there are multiple (G) prior independent sets of results from historical data, denoted by $\mathbf{D}_0 = (D_{01},\ldots,D_{0G})'$. Each previous case is given a weight $\zeta_g\zeta_0$ ($g = 1,\ldots,G$) where ζ_0 is the overall weight that is assigned to previous data and $\zeta_g(>0)$ is the specific weight assigned to case g and $\sum_{g=1}^{G} \zeta_g = 1$. The power prior in such a situation is defined as

$$f^{\mathrm{p}}(\theta|\mathbf{D}_0) \propto (L(\theta|D_{01})^{\zeta_1} \cdots L(\theta|D_{0G})^{\zeta_G})^{\zeta_0} f(\theta) \tag{33}$$

Power priors have been combined with results from simulations, conducted under experimental conditions, to determine appropriate sample sizes (18). For various sample sizes, values were generated from the power prior distribution Eqs (32)–(33) of the parameter of interest and the information in the posterior distribution was summarized by some statistic such as the posterior variance. Then the value of that statistic for various sample sizes was plotted against the corresponding sample sizes. A minimal sample size was chosen so that a certain criterion was met. Let $T(\theta|x_n)$ denote the statistic from a sample of size n from the posterior distribution of θ, given data x_n, whose performance is to be controlled by appropriate sampling. Examples of such statistics are, as before, the posterior variance (12), the mean posterior interquartile range and the width of the HPD interval (13). De Santis' method (18) for estimation of the precision of a statistic T in a power prior where the statistic cannot be determined analytically consists of the following steps:

- Draw a number, let it be b, of θ^*s $(\theta_1^*,\ldots,\theta_b^*)$ from the power prior distribution $f^{\mathrm{p}}(\theta^*|D_0)$ with given values n_0, x_0, ζ_0, v_1, and v_2. (The symbol * denotes a simulated value.) A typical value of b may be 1000.
- Draw, for each θ^*, a simulated sample $\mathbf{x}_n^*$ of size n from the sampling distribution $f(\mathbf{x}|\theta^*)$, to give a likelihood $L(\theta|\mathbf{x}_n^*)$.
- Compute $f(\theta|\mathbf{x}_n^*, D_0) \propto L(\theta|\mathbf{x}_n^*) f^{\mathrm{p}}(\theta^*|D_0)$ for each of the b generated samples.
- Compute $T(\theta|\mathbf{x}_n^*)$ for each of the b generated samples.
- Approximate $\Pr(T(\theta|\mathbf{x}_n)) \in A$ with the proportion of the b generated samples $T(\theta|\mathbf{x}_n^*)$ that belong to the set A. Similarly, $E[T(\theta|\mathbf{x}_n)]$ is estimated by the sample arithmetic mean of the b generated values $T(\theta|\mathbf{x}_\mathbf{n}^*)$ and $\mathrm{var}(T(\theta|\mathbf{x}_n^*)$ by the sample variance.

This method, with the same b, is applied repeatedly to larger values of n until the required criterion is met, for example that the posterior standard error of the statistic T is less than a certain value. This gives the sample size to be used in future cases for which the corresponding power prior is relevant.

Suppose the posterior variance for a sample of size n because of its relative simplicity is chosen as the statistic T of the posterior distribution. Assume a beta prior and a binomial sample. The variance of the beta posterior (or power prior) described here is given by $(v_1^* v_2^*)/((v_1^* + v_2^*)^2(v_1^* + v_2^* + 1))$ where $v_1^* = \zeta_0 x_0 + v_1 + x^*$ and $v_2^* = \zeta_0(n_0 - x_0) + v_2 + n - x^*$. This is divided by the sample size n and then the square root is taken to obtain the posterior standard error. The simulation process is not required here as the posterior standard error can be determined analytically.

The sample size could then be chosen as the minimum sample size for which the posterior standard error is lower than some prespecified value, 0.01 or 0.05, for example. In practice by drawing plots of the posterior standard error against the sample size the behavior of the procedure can be monitored by observing decreases in the posterior standard error as the sample size increases and determining the appropriate sample size as that one after which the posterior standard error decreases only slightly. A crucial aspect of the power prior approach is the choice of the weight ζ_0 given to the previous study. Optimal sample sizes are decreasing functions of ζ_0 as the less weight that is given to a previous, similar, study the more uncertainty there is about the current study. The major advantage of this method is its simplicity and the fact that it enables numerous scenarios to be considered without any constraint of time or finance as everything is based on the previous study and on simulated results.

Consider the consignment of 5000 illicit pills with $n_0 = x_0 = 25$. A beta(1,1) prior is taken. Figure 2 shows the mean posterior standard error, using De Santis' method (18), for various weights ζ_0 given to the previous study, as the sample size increases. For large values of n the weight given to the previous study is of little importance. A sample size of 60 seems adequate regardless of the weight attached to the previous study.

If a large degree of trust (e.g., $\zeta_0 = 0.8$) is permitted for the historical data, a sample size of 10 yields a posterior standard error

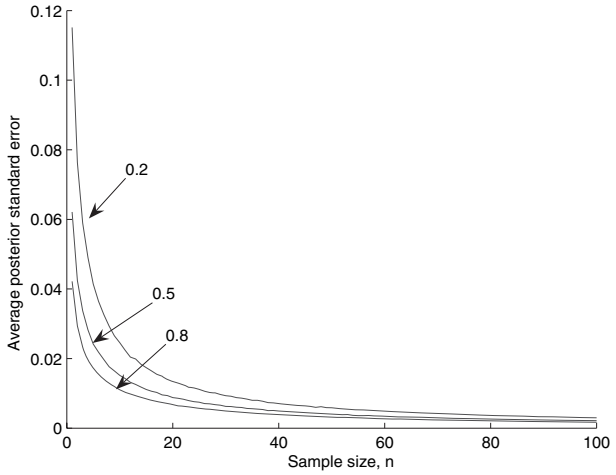

FIG. 2—*Graphs of the sample size versus the average posterior standard error for various weights assigned to historical data [$\zeta_0 = (0.2,0.5,0.8)$] for inspection of a consignment of 5000 pills, each of which is either licit or illicit. Historical data are available of a sample of size* $n_0 = 25$ *pills in which all pills were illicit* ($x_0 = 25$).

smaller than 0.01 and that would lead to a considerable saving of time and financial resources. This can be verified numerically. First, consider $\zeta_0 = 0.8$, $n_0 = x_0 = 25$, $v_1 = v_2 = 1$, $n = x^* = 10$. Then $v_1^* = 31$ and $v_2^* = 1$. The posterior variance $= 31/(32^2 \times 33) \simeq 0.030288^2$ and the posterior standard error equals $0.030288/\sqrt{n} = 0.0096 = 0.01$, to two decimal places, when $n = 10$. Second, consider $\zeta_0 = 0$. Values of $n = x^* = 10$ give a posterior standard error of 0.024 and values of $n = x^* = 20$ give a posterior standard error of 0.01. Thus, without the historical data, the sample size required to provide the same standard error is bigger by a factor of two.

If there are no historical data available, a similar approach may be followed by investigating a fraction of the data as if it were historical data. This fraction should be given full weight ($\zeta_0 = 1$) as it is part of the data. By applying the power prior approach, after inspecting an initial fraction of the data, the extra samples that should be taken may be determined.

Comparison by Simulations of Methods for Sample Size Determination for Binomial Populations

So far, three sequential methods have been suggested for SSD for binomial populations. The results of comparisons of their performances using simulations are presented in Table 2. The first method is the SPRT which, in the simulations presented in Table 2, is denoted by $Q_1(\alpha, \beta)$ where α and β are the probabilities of type 1 and type 2 error, respectively. The other two methods are Bayesian and they employ the conjugacy property of the beta distribution with respect to binomial sampling to obtain a closed-form posterior (beta–binomial). In all simulations conducted a beta(1,1) prior was assumed for the probability of "success" θ. One method uses only the posterior distribution of the population parameter and sampling is stopped when one of the competing propositions is accepted with a certain probability (either inequality Eq. [29] or inequality Eq. [30] is satisfied). This two-sided sequential criterion is denoted by $Q_2(p)$ in Table 2 with p denoting the probability that one of the competing propositions is accepted. The last method combines information both from the inspected units, up to a specific point, and from the predictive distribution of the units not inspected. The criterion $Q_3(l)$ is used and sampling is stopped when the predictive probability interval, expression (27), is less than l in width.

The analysis of random samples is used for evaluating the different criteria. The results presented in Table 2 are obtained from 1000 random samples. All three methods [$Q_1(\alpha, \beta)$, $Q_2(p)$, $Q_3(l)$] are compared on the same simulated samples.

The Q_1 criterion (SPRT) yields the most stable results even if very small type 1 and type 2 error probabilities (α and β) are used. Also, the mean sample size is always close to the median (slightly larger), although the distance of the maximum sample to median sample size is much larger than the corresponding distance of the median sample size to the minimum one. Very large sample sizes were observed in less than 0.5% of the samples.

The Q_2 criterion (two-sided sequential criterion) is very much dependent on the probability p defined *a priori* for selecting one of the two competing propositions. It gives very small sample sizes when the true population proportion θ is far away from both θ_1 and θ_2 or when θ_1 is not very close to θ_2. The population proportion is not known in advance and when it is close to either θ_1 or θ_2 the sample size required is increased considerably. Also the distribution of the sample size is skewed to the right and there is a considerable probability of obtaining a large sample size. There are also cases where the entirety of the samples (maximum equals 1000) is investigated without reaching any conclusion.

The Q_3 criterion is dependent on the maximum width of the interval that is specified. For values of θ close to 0.5, with a population size equal to 1000 and for a maximum permitted width of the interval equal to 0.1, the sample size required is around 350 with little variation. An increase in the maximum permitted width of the interval leads, not surprisingly, to a reduction in the required sample size.

If there is no restriction on the width of the interval, the method may lead to sample sizes of just one unit with a high probability of accepting the wrong proposition. Alternatively, the restriction placed on the width of the probability interval for θ can be removed and a restriction instead placed on the number of sampling units, e.g., to be at least equal to 20. This method leads to small sample sizes with zero probability of accepting the wrong proposition in all cases presented in Table 2 with the exception of the last case ($\theta = 0.12$ and $\theta \leq 0.03$ vs. $\theta \geq 0.1$) which gives a probability very close to 0.6 (0.571). From various simulations, under different scenarios, the conclusion is that when θ is very close to either of the propositions being tested there is a high probability of obtaining misleading results.

Multinomial Sampling

Previous examples have considered a binary response. There are situations in forensic science where a consignment may have more than two categories of items in it. For example, in a drug case a consignment of pills may have three categories such as licit, ecstasy, and LSD and several other examples have been given in the introduction. In all of these examples the problem is to determine the size of the sample needed in order to estimate the proportions of each category.

There are extensions to the ideas presented here for which further work is needed. In the examples in the previous paragraph the number of categories is assumed known. If this is not the case then a different approach is needed to estimate the number of categories as well as the proportion of each. Also, even if the number of categories is known, another problem is to determine how big a sample is needed in order to ensure there is at least one item in the sample from each type. This problem has already been considered above in the example of gunshot residue where the sample size was determined in order to have a certain probability of detecting

the presence of a particle of gunshot residue if a certain proportion of the total number of particles were gunshot residue. The problem can be extended further. Consider a collection of glass fragments. The problem is to determine the size of the sample that is needed to ensure there is at least one item in the sample from each category of glass.

Study of the problem of estimation of proportions when there are more than two categories requires a generalization of the situation in which there is a binary response and for which a binomial distribution is appropriate. The binomial distribution models the variation for the number of outcomes of a particular type in a sequence of independent trials where there are only two possible, mutually exclusive outcomes and the probability of a particular outcome is constant and fixed from trial to trial. The generalization models the variation for the number of outcomes of a particular type in a sequence of independent trials where there are several possible mutually exclusive outcomes and the probability of a particular outcome is constant and fixed from trial to trial. The distribution which generalizes the binomial distribution is the multinomial distribution. The probability of a success in a binomial context is denoted θ with the corresponding probability of a failure denoted $(1 - \theta)$. Consider, now, k categories, where $k \geq 2$ and $k = 2$ corresponds to the binomial context. The parameters of the multinomial distribution may be denoted $\theta = (\theta_1,\ldots,\theta_k)'$, where $\sum_{j=1}^{k} \theta_j = 1$ and $\theta_1,\ldots,\theta_k > 0$. The distribution is itself denoted $\mathrm{Mn}(\theta_1,\ldots,\theta_k)$ where Mn is short for "multinomial." The probability function for $\mathbf{x}$ is given by

$$\Pr(\mathbf{x}|n, \theta_1, \ldots, \theta_k) = \frac{n! \prod_{j=1}^{k} \theta_j^{x_i}}{\prod_{j=1}^{k} x_j!} \tag{34}$$

where

$$\sum_{j=1}^{k} \theta_j = 1, \quad \sum_{j=1}^{k} x_j = n \tag{35}$$

It is desired to obtain the set of k intervals S_j, $j = 1,\ldots,k$, of the shortest length such that

$$\Pr\left\{\bigcap_{j=1}^{k}(\theta_j \in S_j)\right\} \geq 1 - \alpha \tag{36}$$

The sample proportions are denoted by the vector $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)'$ where $\hat{\theta}_j = x_j/n$. An example for the intervals S_j may be those defined by the absolute difference of the estimates $\hat{\theta}_j$ and the corresponding parameters θ_j, and a criterion that the absolute difference be less than d_j, $j = 1,\ldots,k$. It is required that the probability will be at least $1-\alpha$ that all of the estimated proportions $\hat{\theta}_j$ will simultaneously be within d_j of the true population proportions θ_j, that is,

$$\Pr\left\{\bigcap_{j=1}^{k}|\hat{\theta}_j - \theta_j| \leq d_j\right\} \geq 1 - \alpha \tag{37}$$

It is assumed that the population is large enough for finite population correction factors to be ignored and that sample sizes are large enough for the normal approximation to be used. The sample proportions $\hat{\theta}$ converge asymptotically to a so-called degenerate multivariate normal distribution

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, V) \tag{38}$$

where $V = (1/n)\ (\mathrm{diag}(\theta) - \theta\theta')$ is the $k \times k$ covariance matrix and diag denotes a diagonal matrix (i.e., a matrix in which the diagonal terms are the components $(\theta_1,\ldots,\theta_k)$ and the off-diagonal terms are zero. The covariance matrix V has the elements $(1/n)\theta_j(1 - \theta_j)$ on its main diagonal and off-diagonal elements $-(1/n)\theta_{j_1}\theta_{j_2}$ for $j_1 \neq j_2$ and $j_1, j_2 = 1,\ldots,k$. This is a singular covariance matrix (i.e., one whose inverse does not exist) of dimension $k - 1$ due to the restriction $\sum_{j=1}^{k} \theta_j = 1$, hence the term "degenerate."

A method for constructing simultaneous confidence intervals for multinomial proportions is presented in Ref. (19). The method assumes that $n\theta_j$ is large enough (at least 5) for the square of the Pearson's residual

$$\sum_{j=1}^{k} \frac{(x_j - n\theta_j)^2}{n\theta_j}$$

to be chi-squared distributed with $k - 1$ degrees of freedom. This result was improved through the construction of less conservative confidence intervals (19). This improved method was based on the normal approximation for a binomial proportion and used Bonferroni's inequality to put a bound on the probability that all of the intervals would be simultaneously correct. Neither Goodman (20) nor Quesenberry and Hurst (19) addressed the problem of the sample size. Goodman's (20) equation for a $(1 - \alpha)\%$ confidence interval for $\hat{\theta}_i$ is

$$\left(\hat{\theta}_j - z_{\frac{\alpha}{2k}}\sqrt{\frac{\theta_j(1-\theta_j)}{n}}, \hat{\theta}_j + z_{\frac{\alpha}{2k}}\sqrt{\frac{\theta_j(1-\theta_j)}{n}}\right) \tag{39}$$

Angers (21) noted that because the distribution of $\hat{\theta}_j$ converges asymptotically to a degenerate multivariate normal distribution with a singular variance covariance matrix of rank $k - 1$ the correct $(1 - \alpha)\%$ confidence interval for $\hat{\theta}_j$ is

$$\left(\hat{\theta}_j - z_{\frac{\alpha}{2(k-1)}}\sqrt{\frac{\theta_j(1-\theta_j)}{n}}, \hat{\theta}_j + z_{\frac{\alpha}{2(k-1)}}\sqrt{\frac{\theta_j(1-\theta_j)}{n}}\right) \tag{40}$$

The probability α_j, that the sample estimate $\hat{\theta}_j$ is further than d_j from θ_j is given by the normal approximation to a binomial proportion

$$\alpha_j = \Pr(|\hat{\theta}_j - \theta_j| > d_j) \simeq 2(k-1)\left(1 - \Phi\left(\frac{d_j\sqrt{n}}{\sqrt{\theta_j(1-\theta_j)}}\right)\right) \tag{41}$$

The smaller the α_js the larger the sample size required in order to attain them. A decrease in α_j implies a decrease in

$$\left(1 - \Phi\left(\frac{d_j\sqrt{n}}{\sqrt{\theta_j(1-\theta_j)}}\right)\right)$$

which arises from an increase in n, for unchanged d_j and θ. In the multinomial setting, two simple methods for deriving the appropriate sample size are as follows:

1 Assume the parameter vector $\theta = (\theta_1,\ldots,\theta_k)'$ is known. Select a sample size n, observe $x_1,\ldots,x_k$, calculate $\hat{\theta}_j$, $j = 1,\ldots,k$ and compute $\sum_{j=1}^{k} \alpha_j$ for given d_j, $j = 1,\ldots,k$. If $\sum_{j=1}^{k} \alpha_j < \alpha$, repeat with a smaller value of n. Otherwise, repeat with a larger value for n until the smallest n is found such that $\sum_{j=1}^{k} \alpha_j \leq \alpha$. The algorithm is usually initialized with sample size equal to one and the sample size is incremented gradually by one unit at a time until $\sum_{j=1}^{k} \alpha_j \leq \alpha$. This method is employed in Ref. (22) where the vector θ was considered known and the sample

size for obtaining confidence intervals of specified lengths was to be determined. An obvious disadvantage of this method is that the vector of proportions attributed to each category is never known in advance.

2 The second method is to carry out the first procedure with all possible parameter values to determine the parameter vector which gives the largest sample size and use this sample size. This method was applied by Thompson (23) and the form of the worst case scenario was established. More specifically, Thompson proved that the worst case scenario for a multinomial distribution occurs when $k - h$ ($0 \le h \le k$) proportions are equal to zero and the remaining h have probability h^{-1}. If, for a case with k categories, zero proportions for $k - h$ of the categories are observed the dimensionality of the data is reduced by $k - h$. Zero observations for a category are taken to imply the true probability for the category is zero. For example if there is one category with zero frequency this means that the appropriate sample size is estimated considering a vector of equal probabilities assigned to $k - 1$ categories. A disadvantage of this method is that it might lead to unnecessarily large samples especially when there are categories with high or low proportions and, therefore, small variances.

Table 3 gives the appropriate sample sizes needed for a trinomial experiment so that the probability that one or more of the k estimates of proportions θ_j, $j = 1,\ldots,k$ is outside an interval of length $2d = 2 \times 0.05 = 0.1$ centered on the true, unknown, proportion θ, will be less than or equal to α. For larger numbers of categories, the sample sizes are the same as with $k = 4$ as the worst case scenario gives zero frequencies to many cells leaving only four with nonzero and equal probabilities. This is a restriction of the worst case scenario defined by Thompson.

The effect on sample size of assuming the worst case can be illustrated with other choices of the set of proportions. Consider $k = 3$ and $\alpha = 0.05$. First, assume the true set of proportions is $\theta = (0.2,0.3,0.5)'$ rather than the worst case scenario. In that case the sample size reduces from 624 (Table 3) to 593. If the set $\theta = (0.05,0.05,0.9)'$ is considered the necessary sample size is reduced to 182. Assumption of the worst case scenario leads to a much larger sample size than necessary, with a consequent unnecessary expenditure of resources. An alternative Bayesian procedure is shown to reduce the sample size considerably.

Bayesian techniques for SSD

Given a multinomial likelihood and a Dirichlet prior (Eq. [49]) in the Appendix, the posterior distribution of the parameter vector θ is also a Dirichlet distribution

TABLE 3—*Appropriate sample sizes* n *using Thompson's method ($d = 0.05$) which assumes the worst assignment of proportions, $\theta_1 = \ldots = \theta_k$ for various significance probabilities* α *and numbers* k *of categories such that the probability one or more of the estimates of the individual category proportions are outside the intervals of length 2d centered on the true, unknown, proportions* $\theta_1,\ldots,\theta_k$.

α	n	k
0.1	510	3
0.05	624	3
0.025	748	3
0.02	788	3
0.01	915	3
0.1	574	4
0.05	684	4
0.01	946	4

This posterior distribution can be approximated by the singular multivariate normal distribution with mean vector μ and variance–covariance matrix V. Then, it can be shown that (1)

$$T^2 = \left(n + \sum_{j=1}^{k} v_j + 1\right)(\theta - \mu)' V^{-}(\theta - \mu) \sim \chi^2_{k-1} \qquad (42)$$

where χ^2_{k-1} denotes the chi-squared distribution with $(k - 1)$ degrees of freedom, V^{-} is the generalized inverse of V (see the Appendix for an explanation of a generalized inverse) and $(v_1,\ldots,v_k)$ are the prior parameters of the Dirichlet distribution. The sample size may be estimated using the requirement $\Pr[T^2 \le d^2] = 1 - \alpha$ that leads to the rule

$$n + \sum_{j=1}^{k} v_j + 1 \ge \chi^2_{k-1,\alpha}/d^2 \qquad (43)$$

where $d^2 = (\theta - \mu)' V^{-}(\theta - \mu)$ is an ellipsoid in $(k - 1)$-dimensional space, centered on μ and $\chi^2_{k-1,\alpha}$ denotes the percentage point of χ^2_{k-1} such that for a random variable X^2 with a χ^2_{k-1} distribution $\Pr(X^2 > \chi^2_{k-1,\alpha}) = \alpha$. Thus, from Eq. (43) a value for n may be obtained (1). Note, although, that the choice of d and the choice of parameters for the prior distribution both affect the choice of sample size. Note that for large values of the prior parameters, and hence large $\sum_{j=1}^{k} v_j$, one obtains smaller sample sizes that in extreme cases may even become negative. A negative result implies that our prior beliefs are so strong that there is no need to collect more data. The choice of n such that $n \ge \chi^2_{k-1,\alpha}/d^2$ is an appropriate conservative choice.

Power Priors for Multinomial Experiments

The use of historical data for determining the sample size using power priors for beta and binomial distributions can be extended to multinomial data by using a Dirichlet prior for the parameters of interest.

Suppose a multinomial likelihood from a previous experiment leads to a Dirichlet power prior. More specifically, a Dirichlet prior with parameters $\mathbf{v} = (v_1,\ldots,v_k)$ is combined with data $(x_{01},\ldots,x_{0k})$ from an experiment subsequent to the choice of prior but before the planned experiment. These data are given a weight ζ_0. The Dirichlet power prior is then

$$f(\theta_1,\ldots,\theta_k|\mathbf{v}) \propto \left(\prod_{j=1}^{k} \theta_j^{x_{0j}}\right)^{\zeta_0} \prod_{j=1}^{k} \theta_j^{v_j - 1} = \prod_{j=1}^{k} \theta_j^{\zeta_0 x_{0j} + v_j - 1}$$

The power prior, combined with current data $(x_1,\ldots,x_k)$ yields a Dirichlet posterior with parameters $(\zeta_0 x_{0j} + x_j + v_j;\ i = 1,\ldots,k)$. The current data are not known in advance and a simulation-based approach, as that presented earlier in the context of the beta distribution, is used to determine the sample size as that for which a function of the posterior distribution satisfies a prespecified threshold. Such a threshold may be the trace of the posterior covariance matrix (sum of parameter variances) or the sum of the posterior standard deviations. The parameters of the Dirichlet posterior distribution are $v_j^* = \zeta_0 x_{0j} + x_j + v_j$. Let $v_0^* = \sum_{j=1}^{k} v_j^*$. The posterior covariance matrix has variances

$$\mathrm{var}(x_j|\mathbf{v}^*) = \frac{v_j^*(v_0^* - v_j^*)}{v_0^2(v_0 + 1)} = \xi_{jj}$$

and covariances

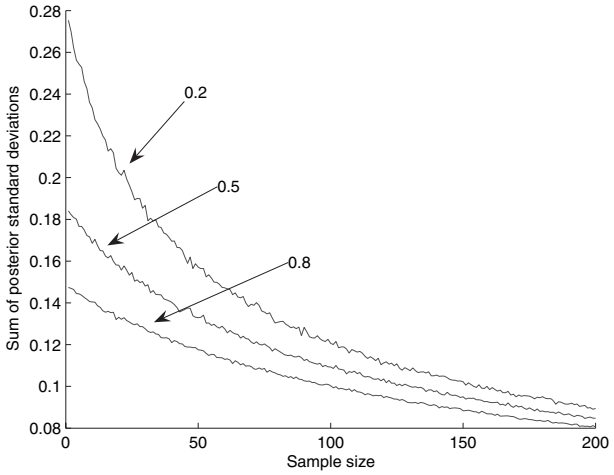

FIG. 3—*Graphs of the sample size versus the sum of the posterior standard deviations for various weights assigned to historical data* $[\zeta_0 = (0.2,0.5,0.8)]$ *based on historical data of a sample of size* $n_0 = 100$ *pills from a trinomial population with 20 pills licit, 30 pills LSD, and 50 pills ecstasy* $[x_0 = (20,30,50)]$.

$$C(x_{j_1}, x_{j_2}|\mathbf{v}^*) = \frac{-v^*_{j_1} v^*_{j_2}}{v_0^2(v_0+1)} = \xi_{j_1 j_2} = \xi_{j_2 j_1}$$

where $\mathbf{v}^* = (v^*_1, \ldots, v^*_k)'$.

Suppose that there are historical data of $n_0 = 100$ observations from a trinomial population with 20 observations falling in the first category, 30 observations falling in the second category, and 50 observations falling in the third category $[x_0 = (20, 30, 50)]$. The criterion for the choice of sample size is to choose as that size of sample, the value for which the sum of the posterior standard deviations for the three categories is less than some prespecified value. Figure 3 plots the sample size for the current data versus the sum of the posterior standard deviations for various weights assigned to historical data. It can be seen that large sample sizes are needed, even for $\zeta_0 = 0.8$, to provide estimates of the trinomial proportions in which the sum of the posterior standard deviations is less than 0.1.

Ternary Diagrams

Ternary diagrams are very popular in the geochemical sciences (24) and have been extensively used to represent the relative percentages of three components as points in an equilateral triangle. A necessary requirement for the construction of such a diagram is that the three components should sum to a fixed amount, for example, to 1 if proportions are being used or to 100 if percentages are being used. This requirement places the restriction that once two of the components are known the other is obtained by subtracting the sum of the two known components from the fixed amount. Hence, only two out of the three components are freely selected reducing the dimension of the problem from three to two. Therefore, whilst ternary diagrams provide a visual representation of apparently three-dimensional data, they are able to be plotted in two dimensions which eases interpretation. A ternary plot may be represented as an equilateral triangle as shown in Fig. 4.

Consider a trinomial experiment. Each subject gives only one response but the cumulative proportion of each category may be seen as a form of compositional data as the sum of these is a constant (100%). This is illustrated in Table 4 where the cumulative percentages, up to any number of subjects, may be seen as a composition.

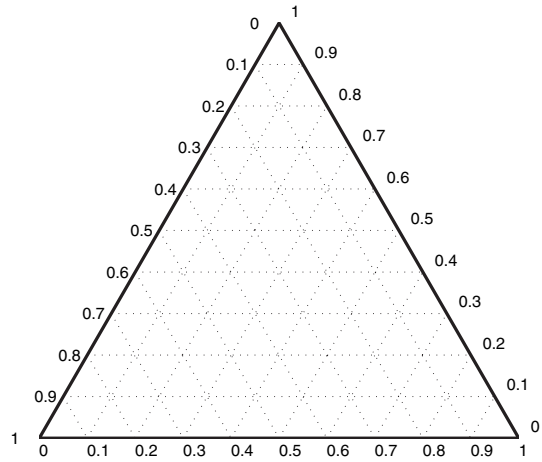

FIG. 4—*Ternary diagram.*

TABLE 4—*Sequential representation of data with three components.*

Subject j	x_1	x_2	x_3	$\frac{\sum x_{j1}}{n}\%$	$\frac{\sum x_{j2}}{n}\%$	$\frac{\sum x_{j3}}{n}\%$
1	0	1	0	0	100	0
2	1	0	0	50	50	0
3	1	0	0	67	33	0
4	0	0	1	50	25	25
5	0	1	0	40	40	20
⋮	⋮	⋮	⋮	⋮	⋮	⋮

A brief discussion of the construction of such a diagram is given. Figure 5 shows the increments along the first axis. If the composition of the first element is 0.1 then that element would lie on the line that corresponds to 0.1 in Fig. 5. All elements on that line refer to two different combinations of the other two variables which should sum to 0.9.

The representation of the point (0.2, 0.3, 0.5) on a ternary diagram is shown in Fig. 6. Approximation of the multinomial probability by a multivariate normal (Eq. [38]) enables the construction of ellipsoidal contours of any precision. Figure 7 shows a 95% probability ellipse contour. The area E under the ellipse is given as the product

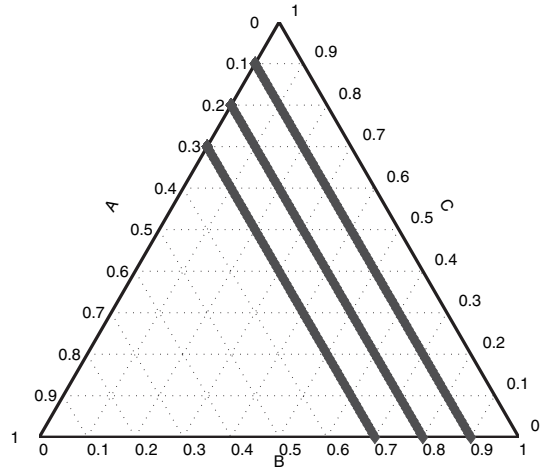

FIG. 5—*Ternary diagram showing increments along the* A*-axis parallel to the* C*-axis; the line* A = 0 *is the line labeled C in the diagram. The* B- *and* C*-axes may be illustrated similarly.*

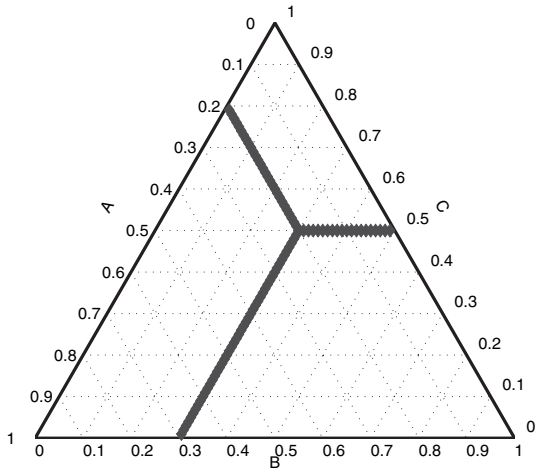

FIG. 6—*Representation of point (0.2, 0.3, 0.5) in a ternary diagram.*

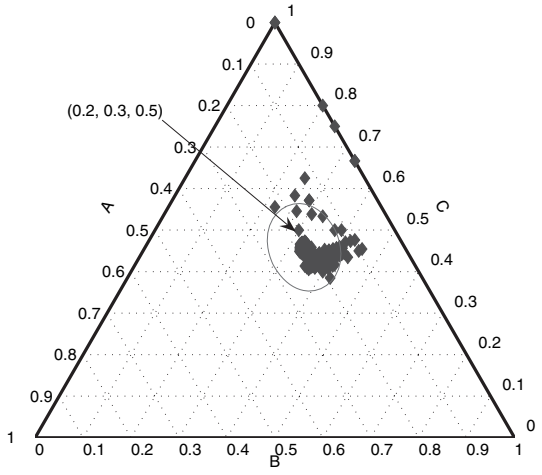

FIG. 7—*Ternary diagram with confidence region. The arrow indicates the point (0.2, 0.3, 0.5). The first 200 points in the sequential composition are depicted.*

$$E = \pi\lambda_1\lambda_2 \tag{44}$$

where λ_1 and λ_2 are half the lengths of the major and minor axes, respectively. For example, for $\lambda_1 = \lambda_2 = \lambda$ say, the ellipse is a circle and the area is $\pi\lambda^2$. Note that each point in Fig. 7 does not represent a sampling unit but the composition of the data up to, and including, the inspection of that point. The first 200 units are depicted in Fig. 7. Units have converged in the ternary plot and the nucleus of the points has converged around the true proportions ($A = 0.2$, $B = 0.3$, $C = 0.5$).

A Sequential Method Stopping when the Joint Probability Interval or Ellipsoid for the Parameter Estimates is Less than a Given Threshold

In the method described in Ref. (23), the sample size is defined *a priori* under the proposition that the worst possible scenario will happen. Alternatively, a sequential sampling scheme using Thompson's method can be adopted allowing estimation of the value of the parameter vector incrementally and thus avoiding use of the worst possible parameter vector which may be far from being true. The parameter vector is estimated as sampling units are examined and the criterion $\sum_{j=1}^{k} \alpha_j \leq \alpha$ is attained much sooner, with a consequent saving of resources.

Furthermore, ellipsoids around the data points can be formed sequentially and sampling can be stopped when the volume of the ellipsoid is below a prespecified threshold so long as it is within the sample space. The procedure can be illustrated in the trinomial case where data points can be represented graphically in two dimensions using ternary diagrams but is applicable to any number of dimensions. After each unit has been sampled an ellipsoid of prespecified probability volume (e.g., 95%) is formed around the latest data point. Sampling can be stopped when the area of the ellipse is below a certain proportion of the area of the equilateral triangle (e.g. 1% and 5%). The ellipses are defined by the set of θ such that

$$(\theta - \hat{\mu})'\hat{V}^{-}(\theta - \hat{\mu}) \leq \chi^2_{k-1,\alpha}$$

where $\hat{\mu}$ and $\hat{V}^{-}$ are given by the posterior mean and covariance of the Dirichlet distribution in the Appendix.

Suppose that there is a seizure of 10,000 pills suspected to contain illicit drugs. Suppose that out of the $N = 10{,}000$ pills, 10% are ecstasy pills, 20% are nonillicit drugs, and 70% are LSD pills, although in practice these proportions are not known. The population is assumed to be homogeneous in the sense that a simple inspection of the pills did not reveal any visual differences (e.g., shape and color). Thompson's method leads to a sample size of 624 for $d = 0.05$ and significance level $\alpha = 0.05$. The worst possible scenario in that case is one for which two of the cells have probability 0.5 while the other has zero probability. This scenario is far from being true for the specific data. If Thompson's method is applied sequentially, vector $\theta = (\theta_1,\theta_2,\theta_3)'$ is estimated sequentially and comes closer to the vector of true estimates. This leads to a sample size of 437. For application of the third method, sampling is stopped when the area of the ellipse formed around the data is below 5% of the area of a ternary diagram, $0.05\sqrt{3}/4$. This way, the sample size from one simulation is 88. The ternary diagram and the corresponding ellipse after the inspection of the 88 sampling units are shown in Fig. 8. The ellipse shows a 95% region for the composition of the data when sampling is stopped. The estimated proportions are very close to the true ones.

Sequential methods do not necessarily give an invariant sample size as they are dependent on the data being inspected. Repetitions may give different sample sizes. The sequential sampling scheme with a stopping rule defined by the area of the 95% ellipse around the data points was repeated with 100 simulations from the consignment. The average sample size needed was found to be 83.89 with values ranging from 38 to 112 and a standard deviation of 17.

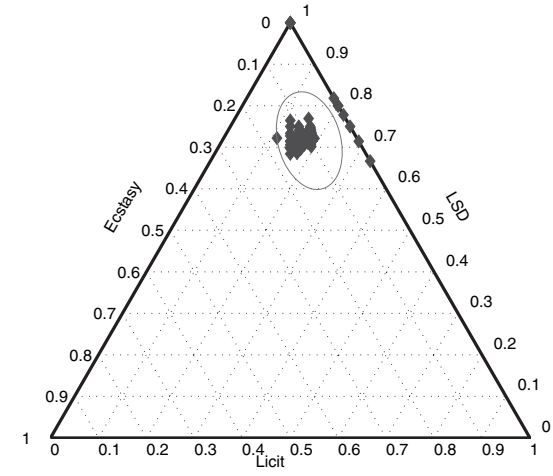

FIG. 8—*Ternary diagram and a 95% ellipse around data points.*

TABLE 5—*Mean sample sizes, n̄, based on the sizes of 100 simulations from multinomial distributions with parameters θ = (0.2,0.3,0.5) and θ = (0.05,0.05,0.90).*

Criterion	$\bar{n}, \theta = (0.2, 0.3, 0.5)'$	$\bar{n}, \theta = (0.05, 0.05, 0.90)'$
$C(0.01)$	498.1	153.19
$C(0.02)$	254.37	77.58
$C(0.05)$	55.42	42.14
$I(0.05, 0.05)$	593.51	181.67

$C(t)$ is the criterion to stop sampling when the area under the 95% ellipse formed by the data and using the posterior Dirichlet distribution is below the threshold t for $t = 0.01,0.02,0.05$. $I(\alpha, d)$ is the criterion to stop sampling when the sum of the individual significance probabilities α_j ($j = 1,\ldots,k$) is $\leq\alpha$ and the probability is at least $(1 - \alpha)$ that the estimated proportions, $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$, will simultaneously be within specified distances d (i.e., $d_1 = d_2 = d_3 = d$) of the true population proportions.

In all cases a much smaller sample size than both Thompson's fixed sample size method and the sequential Thompson's method was obtained. Prior information will reduce the variability and lead to smaller ellipses and consequent smaller sample sizes.

In practice, the guidance to a practicing forensic scientist is to sample sequentially, calculate the probability ellipse at regular intervals and stop when the necessary criterion is attained. The simulations described here demonstrate that if one wishes to have a 95% probability that the true proportions lie within an ellipse of area no greater than 5% of the total sample space then the expected sample size is 84 (83.89 to two decimal places).

Simulations

The ability of many methods to determine the appropriate sample size for drawing valid inference is illustrated using a simulation study and a parametric bootstrap method. One hundred data sets are generated from multinomial distributions and some of the above approaches are compared. The criterion to stop sampling when the area under the 95% ellipse formed by the data is below a threshold t is denoted by $C(t)$ in Table 5. The modified sequential Thompson's criterion to stop sampling when $\sum_{j=1}^{k} \alpha_j \leq \alpha$ so that the probability is at least $(1 - \alpha)$ that the estimated proportions will simultaneously be within specified distances d of the population proportions is denoted $I(\alpha, d)$ in Table 5.

Column $\bar{n}$ gives the mean sample size for two different multinomial populations, namely Mn(0.2, 0.3, 0.5) and Mn(0.05, 0.05, 0.9). Smaller sample sizes are obtained by controlling the area under the ellipse formed from the data points than by controlling the widths of the multiple confidence intervals. These two methods cannot be compared directly as they do not necessarily have the same coverage probability. However, the last row [$I(0.05, 0.05)$] of Table 5 can be compared to Thompson's method (under the worst possible scenario) which gives a sample size of 624, larger than the one obtained using a sequential approach. Furthermore, the more unequal the multinomial proportions, the smaller the sample sizes needed to distinguish between them.

Conclusions

The paper considers the use of some sampling schemes for testing propositions about binomial and multinomial variables. Examples are given from the forensic sciences. Traditional SSD techniques (6, 23) are designed to ensure that, in the absence of prior information for the parameters of interest, the sample size should be adequate for any possible parameter values. Therefore, the sample size is determined as that for which a prespecified threshold is defined under the worst possible scenario. This may lead to unnecessarily large data sets which require large amounts of time and financial resources.

Five sampling procedures of which four are for binomial variables and one is multinomial have been described. One is based on power priors, informed from historical data. The other four are sequential. One is an SPRT with a stopping rule derived from the probabilities of type 1 and type 2 errors. One is a sequential variation of a procedure based on the predictive distribution of the data yet to be inspected and the distribution of the data that have been inspected, with a stopping rule determined by a threshold. One is based on estimating the posterior probabilities of the competing propositions sequentially and sampling is stopped when these probabilities exceed a prespecified threshold. The fifth is for use with more than two categories. Sampling is stopped when the joint probability interval or ellipsoid for the parameter estimates is less than a given threshold. For trinomial data this last procedure is illustrated in ternary diagrams with ellipses formed around the sample points. There is a straightforward generalization to multinomial populations with more categories. Sequential methods do not yield standard sample sizes every time they are applied to data sets of similar contexts. Thus it is not possible to specify in advance of inspection the required size of sample. In all cases, however, they yield smaller sample sizes than those taken under the worst possible scenario.

Consider binomial data. The SPRT (Eq. [12]) gives more reliable and stable results, using fewer samples, than the predictive approach Eqs (25) and (26) and the two-sided sequential criterion Eqs (29) and (30). The two-sided sequential criterion yields smaller sample sizes and there is a very small probability of accepting the wrong proposition when the true value of θ does not lie very close to the propositions being tested (θ_1 and θ_2). Conversely, larger sample sizes occur when θ_1 is close to θ_2 and θ is not far from θ_1 or θ_2. The predictive method (14) yields very large samples when restrictions are imposed on the width of the probability interval of θ. The value of θ is not known *a priori*. Instead of applying just one SSD method, all three methods (SPRT, predictive method, two-sided sequential criterion) may be applied simultaneously. If either the predictive method or the two-sided sequential criterion provide a conclusion (accept either H_1 or H_2) check if $\hat{\theta}$ is far from either of the propositions and if this is the case sampling can be stopped because this suggests the original propositions are incorrect, otherwise continue sampling until the SPRT provides a conclusion.

Prior elicitation is one of the key aspects of Bayesian analysis and the power prior approach takes advantage of previous similar studies. If there is a large seizure of pills suspected to contain illicit substances, prior information may be acquired by circumstantial evidence. Alternatively, a fraction of the data may be investigated as if they were historical data. This fraction should be given full weight ($\zeta_0 = 1$) as it is part of the data. By applying the power prior approach, after inspecting an initial fraction of the data, the number of extra samples that should be investigated may be simply determined.

The method illustrated for trinomial data is easily extended to multinomial populations with more categories. Obviously, in such a case, the sampling units cannot be represented graphically using ternary diagrams but the method can also be applied estimating the volume (now that the dimension is more than two) of an ellipsoid sequentially until a threshold is satisfied.

Acknowledgments

This research was supported by ESRC grant Res-000-23-0729. The authors also acknowledge very helpful advice from Marjan Sjerps and Annabel Bolck and constructive reviews by anonymous referees.

References

1. Adcock CJ. Sample size determination: a review. The Statistician 1997;46(2):261–83.
2. Aitken CGG, Bring J, Leonard T, Papasouliotis O. Estimation of quantities of drugs handled and the burden of proof. J R Stat Soc Ser A, 1997;160(2);333–50.
3. Aitken CGG, Lucy D. Estimation of the quantity of a drug in a consignment from measurements on a sample. J Forensic Sci 2002;47; 968–75.
4. Izenman AJ. Statistical and legal aspects of the forensic study of illicit drugs. Stat Sci 1997;16(1):35–57.
5. Aitken CGG. Sampling–how big a sample? J Forensic Sci 1999; 44(4):750–60.
6. Cochran WG. Sampling techniques, 3rd edn. Chichester: Wiley, 1977.
7. Hacking I. Logic of statistical inference. New York: Cambridge University Press, 1965.
8. Aitken CGG, Taroni, F. Statistics and the evaluation of evidence. Chichester: John Wiley and Sons Ltd, 2004.
9. Royall R. On the probability of observing misleading statistical evidence. J Am Stat Assoc 2001;95(451):760–8.
10. Wald A. Sequential tests of statistical hypotheses. Ann Math Stat 1945;16(2):117–86.
11. De Santis F. Statistical evidence and sample size determination for Bayesian hypothesis testing. J Stat Plan Inference 2004;124(1):121–44.
12. Pham-Gia T, Turkkan N. Sample size determination: a review. The Statistician 1992;46(2):392–7.
13. Joseph L, Wolfson DB, Berger RD. Sample size calculations for binomial proportions via highest posterior density intervals. The Statistician 1995; 44(2):143–54.
14. Weiss R. Bayesian sample size calculations for hypothesis testing. The Statistician 1997;46(2):185–91.
15. Woodward P, Branson, J. A graphical approach for incorporating prior knowledge when determining the sample size for the assessment of batched products. The Statistician 2001;50(4):417–26.
16. Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials (with discussion). J R Stat Soc Ser A 1994;157(3):357–416.
17. Ibrahim JG, Chen MH. Power prior distributions for regression models. Stat Sci 2000;15(1):46–60.
18. De Santis F. Using historical data for Bayesian sample size determination. J R Stat Soc Ser A 2007;170(1):95–113.
19. Quesenberry CP, Hurst DC. Large sample simultaneous confidence intervals for multinomial proportions. Technometrics 1964;6(2):191–5.
20. Goodman LA. On simultaneous confidence intervals for multinomial proportions. Technometrics 1965;7(2):247–54.
21. Angers C. Note on quick simultaneous confidence intervals for multinomial proportions. Am Stat 1989;43(2):91.
22. Angers C. A graphical method to evaluate sample sizes for the multinomial distribution. Technometrics 1974;16(3):16.
23. Thompson SK. Sample size for estimating multinomial proportions. Am Stat 1987;41(1):42–6.
24. Aitchison J. The statistical analysis of compositional data. New York: Chapman & Hall, 1986.

Additional information and reprint requests:
Colin G.G. Aitken, Ph.D.
School of Mathematics
The King's Buildings
The University of Edinburgh
Mayfield Road
Edinburgh EH9 3JZ
U.K.
E-mail: c.g.g.aitken@ed.ac.uk

Appendix

Notation

The following notation is used throughout the paper. Distributional formulae are given at appropriate points in the text or in the last part of the Appendix.

- θ: the true proportion of a category in the relevant population (or consignment) (e.g., the proportion of illicit pills in a consignment);
- n: the size of the sample taken from the population;
- m: the population size (perhaps unknown, perhaps very large) from which the sample is to be taken;
- x: number of members of a sample of size n belonging to the category of interest when θ is the probability an individual member belongs to the category of interest (e.g., the number of illicit pills in a sample of size n from a consignment of pills of size m);
- y: number of (unknown) members in the $(m - n)$ members of the consignment which have not been examined (e.g., the number of illicit pills in the $m - n$ pills not examined in a consignment of pills of size m from which a sample of size n has been examined);
- $\hat{\theta} = x/n$, the sample proportion of successes in n trials, this ratio is used as an estimate of θ;
- t: threshold for the strength of evidence as measured by the likelihood ratio;
- k: the number of categories in a multinomial experiment (e.g., for a trinomial experiment k would be set equal to 3);
- $\mathbf{x} = (x_1,\ldots,x_k)'$ with a fixed sum, $\sum_{i=1}^{k} x_i = n$, the observed cell frequencies in a sample size n from a multinomial distribution. A bold $\mathbf{x}$ denotes a vector of two or more numbers;
- The $'$ symbol denotes that the set $(x_1,\ldots,x_k)$ is transposed to be a column of counts rather than a row. This is a mathematical convention for representation of a vector.
- The $\hat{}$ symbol denotes that the characteristic over which it is placed is an estimate of the uncovered characteristic. The symbol is read as 'hat.' Thus $\hat{\theta}_i$ (read as 'theta-i-hat') is an estimate of θ_i, and $\hat{\mathrm{var}}(\theta_i)$ is an estimate of the variance of θ_i (see the section on the Dirichlet distribution later in the Appendix);
- θ_j: for a multinomial experiment, the true proportion of category j in the population ($j = 1,\ldots,k$, $\sum_{j=1}^{k} \theta_j = 1$) (for the trinomial example, k would equal 3, and $\theta_1,\ldots,\theta_3$ would be the proportions [unknown] of the three categories in the consignment); the column vector $(\theta_1,\ldots,\theta_k)'$ is denoted θ;
- $\hat{\theta}_j = x_j/n$: the sample proportion of category j in n trials, an estimate of θ_j;
- H_1: $\theta \le \theta_l$ (l for "lower") vs. H_2:$\theta \ge \theta_u$, (u for "upper") the two competing hypotheses being tested throughout the paper; for drugs cases, the length of sentence may be determined partially on the estimated values for the proportions, θ_1 and $\theta_2 = 1 - \theta_1$, of illicit and licit categories and where θ_l and θ_u are lower and upper bounds on a sentencing guideline interval;
- beta(v_1,v_2): beta distribution with parameters v_1 and v_2. The true value of the proportion θ may be unknown but there may be prior information about its value; this information is modeled probabilistically by a distribution known as a beta distribution; see later in the Appendix for more details;
- Dir(v): Dirichlet distribution with parameter vector $v = (v_1,\ldots,v_k)'$, $v_j > 0$, $j = 1,\ldots,k$. The Dirichlet distribution is a generalization of the beta distribution. The beta distribution models uncertainty about the proportions, θ_1 and $\theta_2 = 1 - \theta_1$, in two categories. The Dirichlet distribution models uncertainty about the proportions, $\theta_1,\ldots,\theta_k$ in each of k categories with $\theta_k = 1 - (\theta_1 + \cdots + \theta_{k-1})$; $\theta_j > 0$, $j = 1,\ldots,k$; see later in the Appendix for more details;
- d_j: the maximum permitted distance $|\theta_j - \hat{\theta}_j|$ between θ_j and $\hat{\theta}_j$, the proportion of category j and its estimate in the population, to satisfy a threshold. For example, a criterion for the choice of sample size n may be to choose n such that d_j is less than some prespecified value;
- α: the significance level of a test, the probability of rejecting a null hypothesis, H_1, when it is true. This error is also known as a type 1 error and can be written as $\alpha = \Pr(H_1 \text{ is rejected } | H_1 \text{ is true})$;

- β: the probability of not rejecting the null hypothesis H_1 when the alternative H_2 is true. This error is known as a type 2 error and can be written as $\beta = \Pr(H_1 \text{ is not rejected } | H_2 \text{ is true})$;
- α_j: the significance level of category j for comparison of observed to expected frequencies for that category, $j = 1,\ldots,k$, $k \geq 2$;
- ζ_0 is the overall weight that is assigned to previous data in a power prior;
- $\zeta_g (> 0)$ is the proportion of the overall weight ζ_0 assigned to case g in the power prior with $\sum_{g=1}^{G} \zeta_g = 1$.

Probability Distributions

Beta Distribution

The beta distribution for a random variable θ is a two-parameter (α, β) continuous probability distribution, denoted beta(α, β) on the interval (0,1) with probability density function

$$f(\theta|v_1, v_2) = \frac{\theta^{v_1-1}(1-\theta)^{v_2-1}}{B(v_1, v_2)}, \quad 0 < \theta < 1 \tag{45}$$

where

$$B(v_1, v_2) = \frac{\Gamma(v_1)\Gamma(v_2)}{\Gamma(v_1 + v_2)} \tag{46}$$

and

$$\Gamma(z) = \int_0^{+\infty} t^{z-1}\mathrm{e}^{-t}\, \mathrm{d}t \tag{47}$$

is the gamma function with $\Gamma(z) = (z-1)!$ for integer z. When $v_1 = v_2 = 1$ the beta distribution is a so-called uniform distribution in that $f(\theta|1,1) = 1$, $0 < \theta < 1$. This distribution is often used as a prior distribution for sampling problems; see Ref. (5) for examples.

The expected value and the variance of a beta-distributed random variable θ with parameters v_1 and v_2 are given by the formulae (48) and (49), respectively.

$$E(\theta) = \frac{v_1}{v_1 + v_2} \tag{48}$$

$$\mathrm{var}(\theta) = \frac{v_1 v_2}{(v_1 + v_2)^2(v_1 + v_2 + 1)} \tag{49}$$

Dirichlet Distribution

The generalization of the beta distribution to more than two categories is the Dirichlet distribution. Thus, for $\theta = (\theta_1,\ldots,\theta_k)$, with $\sum_{j=1}^{k} \theta_j = 1$, the Dirichlet probability density function with parameters $(v_1,\ldots,v_k)$ is

$$f(\theta|v_1, \ldots, v_k) = \frac{\Gamma(v_1 + \cdots + vk)}{\Gamma(v_1)\cdots\Gamma(v_k)}\theta_1^{v_1-1}\cdots\theta_k^{v_k-1} \propto \theta_1^{v_1-1}\cdots\theta_k^{v_k-1} = \prod_{j=1}^{k}\theta_j^{v_j-1} \tag{50}$$

The expectation $E(\theta_j)$ of θ_j is

$$\frac{v_j}{\sum_{j=1}^{k} v_j}$$

The variance of θ_j is

$$\frac{E(\theta_j)(1 - E(\theta_j))}{1 + \sum_{j=1}^{k} v_j}$$

The covariance $C(\theta_{j_1}, \theta_{j_2})$, $j_1 \neq j_2$, between θ_{j_1} and θ_{j_2} is given by

$$C(\theta_{j_1}, \theta_{j_2}) = \frac{-E(\theta_{j_1})E(\theta_{j_2})}{1 + \sum_{j=1}^{k} v_j}$$

The beta is a conjugate prior for the binomial distribution, the Dirichlet is a conjugate prior for the multinomial distribution in that the posterior distributions are beta and Dirichlet, respectively, the same form of distribution as the prior distribution. Note that the beta distribution is the special case of the Dirichlet distribution given when $k = 2$.

The posterior distribution for θ, given $\mathbf{x} = (x_1, \ldots, x_k)'$, $\sum_{j=1}^{k} x_j = n$, is a Dirichlet distribution with

$$f(\theta|\mathbf{x}) \propto \prod_{j=1}^{k} \theta^{v_j + x_j - 1} \tag{51}$$

For the posterior distribution, with $v_j = 1$; $j = 1,\ldots,k$, the estimates of the mean, μ, and covariance matrix, V, are given by

$$\begin{aligned}
\hat{\mu} &= (\hat{\mu}_1, \ldots, \hat{\mu}_k)' \text{ with} \\
\hat{\mu}_j &= \frac{v_j + x_j}{\sum(v_j + x_j)} = \frac{x_j + 1}{(n+k)}. \\
\hat{\mathrm{var}}(\theta_j) &= \frac{(x_j + 1)(n + k - 1 - x_j)}{(n+k)^2(n+k+1)} \\
&= \xi_{jj}. \\
\hat{C}(\theta_{j_1}, \theta_{j_2}) &= -\frac{(x_{j_1} + 1)(x_{j_2} + 1)}{(n+k)^2(n+k+1)} \\
&= \xi_{j_1 j_2} = \xi_{j_2 j_1}, \quad j_1, j_2 = 1, \ldots, k, j_1 \neq j_2.
\end{aligned}$$

The covariance matrix V is singular because of the condition that $\sum_{j=1}^{k} \theta_j = 1$. Thus, the inverse does not exist. However, one can use what is known as a *generalized inverse*, denoted V^-. There are $(k-1)$ independent rows in V. Denote the $(k-1) \times (k-1)$ matrix formed by the first $(k-1)$ rows and $(k-1)$ columns of V by V_{11}. This matrix does have an inverse; denote it by V_{11}^{-1}. Let $\boldsymbol{\xi}_1' = (\xi_{k1}, \ldots, \xi_{k,k-1})$ denote the first $(k-1)$ elements of the kth row of V (i.e., the kth row with the last element, ξ_{kk} omitted). Then, by symmetry, ξ_1 denotes the last column of V with ξ_{kk} omitted. Thus, the generalized inverse V^- is given as follows

$$V = \begin{pmatrix} V_{11} & \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_1' & \xi_{kk} \end{pmatrix} \tag{52}$$

$$V^- = \begin{pmatrix} V_{11}^{-1} & \mathbf{0} \\ \mathbf{0}' & 0 \end{pmatrix} \tag{53}$$

where $\mathbf{0}$ is a column vector with $(k-1)$ zeros and $\mathbf{0}'$ is the corresponding row vector.