DRUGS WORKING GROUP

VALIDATION OF THE

'GUIDELINES ON REPRESENTATIVE SAMPLING'

DOCUMENT TYPE :	REF. CODE:	ISSUE NO:	ISSUE DATE:
VALIDATION REPORT	DWG-SGL-001	002	08 DECEMBER 2012

Updates of the last version			
	Date	Update	Page
1.	December-2012	Section 2 “Sampling based on the hypergeometric distribution” and the corresponding tables have been withdrawn.	Section 2, Appendix 1, Appendix 2
2.			
3.			
4.			

1 Introduction

This document presents the results of the validation of the tables and sampling software as published in the 'Guidelines on representative sampling' of the ENFSI. Validation is described by the ENFSI as a process of

1. establishing the performance characteristics and limits of a method and the identification of the influences that may change these characteristics and to what extent, and
2. verifying that a method is fit for purpose, i.e. for use for solving a particular analytical problem.

(See 'Validation and implementation of (new) methods', QCC, ENFSI, 2006).

More specifically this document validates the three statistical sampling methods based on, respectively, the hypergeometric distribution (Section 2), the binomial distribution (Section 3), and the Bayesian approach (Section 3), as well as the method of estimating of the average weight of a drug unit (Section 4). In these sections we will verify, subsequently, whether formulas as described in the '*Guidelines'* are correctly implemented, and whether the tables and subsequently computer software provide expected results. In section 5 we discuss the performance characteristics, limits and influences on these characteristics. We refer to the '*Guidelines'* for a detailed description of these methods, and further considerations and recommendations.

March 2009,

Reinoud Stoel

Netherlands Forensic Institute, The Hague

2 Sampling based on the hypergeometric distribution

Section withdrawn. Please see the new DWG document, reference code: DWG-SGL-002.

3 Sampling based on the binomial distribution

The binomial distribution assumes sampling with replacement. However, when the seizure is very large and the sample is relatively small the hypergeometric distribution can be approximated by the less complex binomial distribution. Similar to the hypergeometric distribution, the binomial distribution can be used to calculate required sample size *n* such that with at least (1-α)100% confidence can be stated that at least a proportion of *k* is positive. Please note that *N*, population size, is not a parameter. It should be kept in mind that the binomial distribution is an approximation, and that the sample size will be (slightly) overestimated.

The first step in the validation process is again to check whether the numbers presented in the table correspond to the results obtained from the computer software. All sample sizes as presented in Table 5.3 correspond to the output of the computer software.

The second step is to verify whether the correct formulas are used, and to validate whether the computer software provides correct results. Determination of the minimum required sample size (*n*) is based on a test the hypothesis

$$H_0 : \theta \leq k \text{ against } H_1 : \theta > k$$

, where $\theta = \frac{N_1}{N}$.

To select *n*, the equation to be solved is

$$P(X \geq x | \theta = k) \leq \sum_{i=x}^{n} \binom{n}{x} \theta^x (1-\theta)^{n-x} \leq \alpha$$

So given that $\theta = k$ the required minimal sample size (*n*) is the smallest value for which $P(X \geq x | \theta = k) \leq \alpha$.

When all sampled drug units are expected to contain drugs (i.e. $x=n$), X is distributed as a binomial random variable:

$$X \sim BIN(x, n, \theta)$$

Resulting in

$$P(X \geq x | \theta = k, x=n) = \theta^n \quad \textbf{(4)}$$

In this case the minimum required sample size is readily obtained by solving

$$n \leq \frac{\log \alpha}{\log \theta} \quad \textbf{(5)}$$

When at most one sampled drug unit is expected not to contain drugs (i.e. $x \geq n-1$), X is distributed as a mixture of two binomial random variables:

$$P(X \geq x | \theta = k, x \geq n\text{-}1) \leq \theta^n + \binom{n}{n-1}\theta^{n-1}(1-\theta)^1 \quad \textbf{(6)}$$

When at most two sampled drug units are expected not to contain drugs (i.e. $x \geq n-2$), X is distributed as a mixture of three binomial random variables.

$$P(X \geq x | \theta = k, x \geq n\text{-}2) \leq \theta^n + \binom{n}{n-1}\theta^{n-1}(1-\theta)^1 + \binom{n}{n-2}\theta^{n-2}(1-\theta)^2 \quad \textbf{(7)}$$

3.1 Results

The software implements Formula 4, 6 and 7 by making use of the standard formulas for the binomial distribution of MS Excel 2003. After having removed the protection of the software it is concluded that the

formulas used for computing the probabilities for $x \geq n$, $x \geq n$-1 and $x \geq n$-2, are equivalent to, respectively Formula 4, 6, and 7. Furthermore, that required sample size (n) is based on the correct cut-off (i.e. $P(X \geq x|\theta = k) \leq \alpha$.).

Further validation of the software is performed by comparing the probabilities of the software ($P_{software}$) with those of hand calculations (P_{hand}) using Formula 4, 6, and 7 as well as calculation by means of the binomial distribution function implemented in the computer software R (P_R).

For a selected number of combinations of k, (1-α)100%, and n-x, the probabilities are computed and compared. Validation is successful if

$$P_{software} = P_{hand} = P_R$$

Table 2 in the Appendix presents the results. As expected, based on the equivalence of the formulas, all computed probabilities are equivalent given the parameters.

3.2 Conclusion

The binomial approach to sampling, as described by the 'Guidelines on representative sampling', is correctly implemented in the software.

4 Sampling based on Bayesian approach.

The Bayesian approach assumes that, although the population proportion is not known, there may be some ideas about the size of this proportion. These ideas are represented by a probability distribution $P(\theta)$. Instead of estimating $P(X \geq x|\theta = k)$, the Bayesian approach directly estimates $P(\theta > k| x,n)$.

4.1 Seizures containing > 50 units (with relative small samples)

Select a sample size n such that the probability that $P(\theta > k| x,n) = (1-\alpha)100\%$.

θ follows a beta distribution with parameters $x+a$ and $n-x+b$:

$$f(\theta \mid x,n,a,b) = \frac{\theta^{x+a-1}(1-\theta)^{n-x+b-1}}{B(x+a,n-x+b)} \quad \textbf{(8)}$$

Selection of the parameters, a and b, should be based on prior information about the content(s) of the units.

4.2 Seizures containing < 50 units

For small seizures it may be better to use the number of positives, Y, in the unexamined units. Select a sample size n such that the probability that

$P(Y > y| x, n, N) = (1-\alpha)100\%$.

Y follows a beta-binomial distribution:

$$f(Y \mid x,n,N-n),a,b) = \frac{\Gamma(n+a+b)\binom{N-n}{y}\Gamma(y+x+a)\Gamma(N-x-y+b)}{\Gamma(x+a)\Gamma(n-x+b)\Gamma(N+a+b)} \quad \textbf{(9)}$$

4.3 Results

4.3.1 > 50 units

Please note that the software uses N≥50 instead of N>50. The software implements Formula 8 by making use of the standard formulas for the Beta distribution of MS Excel 2003.

The first step in the validation process is to check whether the number presented in the table correspond to the results obtained from the computer software. All sample sizes as presented in Table 5.4 correspond to the output of the computer software.

Further validation of the software is performed by comparing the probabilities of the software ($P_{software}$) with those of calculation by means of the Beta distribution function implemented in the computer software R (P_R). For a selected number of combinations of k, n, a, b, the number of negatives, the probabilities are computed and compared. Validation is successful if

$$P_{software} = P_R$$

Table 3 in the Appendix presents the results. As expected based on the equivalence of the formulas, all computed probabilities are equivalent given the parameters.

4.3.2 < 50 units

The software implements Formula 9 by making use of the standard formulas for the Beta distribution of MS Excel 2003.

Further validation of the software is performed by comparing the probabilities of the software ($P_{software}$) with those of calculation by means of the Gamma distribution function implemented in the computer software R (P_R). For a selected number of combinations of k, n, a, b, the number of negatives, the probabilities are computed and compared. Validation is successful if

$$P_{software} = P_R$$

Table 4 in the Appendix presents the results. As expected based on the equivalence of the formulas, all computed probabilities are equivalent given the parameters.

4.4 Conclusion

The Bayesian approach to sampling, as described by the 'Guidelines on representative sampling', is correctly implemented in the software.

5 Estimation of weight

A procedure for estimation of a (1-α)100% confidence interval of the weight of a drug unit, and the total weight in the seizure ($N\bar{X}$) is provided.

$$\left(\frac{n-r}{n}\right)N\bar{X}-\sqrt{\frac{N-n}{N}}\left(\frac{n-r}{n}\right)\frac{Ns}{\sqrt{n}}t_{\alpha} \;\leq\; W \;\leq\; \left(\frac{n-r}{n}\right)N\bar{X}+\sqrt{\frac{N-n}{N}}\left(\frac{n-r}{n}\right)\frac{Ns}{\sqrt{n}}t_{\alpha} \quad \textbf{(10)}$$

where $\sqrt{\frac{N-n}{N}}$ is the finite population correction factor and $\left(\frac{n-r}{n}\right)$ is the correction factor for the number of sampled units without drugs. $\left(\frac{n-r}{n}\right)$ is actually the proportion of drug units in the sample, and $\left(\frac{n-r}{n}\right)N$ is an estimate of the total number of drug units in the population. Please note that the uncertainty in $\left(\frac{n-r}{n}\right)$ is not taken into account (see also Aitken & Lucy, 2002).

Inequality 8 is based on Tzidony & Ravreby (1992, p.1546) (see also Aitken & Lucy, 2002, p.3), however, it seems to contain an error. Since there are *r* units without drugs, the effective sample size on which the estimation of the total weight is based should equal *(n-r)*. This has consequences for both the standard error and the degrees of freedom of the corresponding t-distribution which should equal *(n-r-1)* instead of (*n-1)*, resulting in a slightly higher critical value t_{α}^{*}. Inequality 8 should, therefore, be rewritten into:

$$\left(\frac{n-r}{n}\right)N\bar{X}-\sqrt{\frac{N-n}{N}}\left(\frac{n-r}{n}\right)\frac{Ns}{\sqrt{n-r}}t_{\alpha}^{*} \;\leq\; W \;\leq\; \left(\frac{n-r}{n}\right)N\bar{X}+\sqrt{\frac{N-n}{N}}\left(\frac{n-r}{n}\right)\frac{Ns}{\sqrt{n-r}}t_{\alpha}^{*} \quad \textbf{(11)}$$

Based on the above results it is advised to correct the formula in the booklet and the program (see Stoel & Bolck, in press).

5.1 Conclusion

Although the formulas used in the program are not optimal, they are correctly implemented. That is, the program produces the results as expected from the formulas provided in the booklet. Please note again that the uncertainty in P_{corr} is not taken into account. The confidence interval is, as a consequence, interpreted conditionally on $\hat{P}$. Incorporating the uncertainty in $\hat{P}$ will result in a complex confidence interval with the degrees of freedom of the t-distribution being a random variable. Alberink, Bolck and Stoel (submitted) performed a simulation study and showed that the confidence intervals are indeed not optimal, since they are based on an underestimation of

the variation of the underlying statistical process. They present two other formulas which appear more reliable and asymptotically correct. In particular in situations where the proportion of positives is becoming increasingly smaller than 1. Future research is needed to investigate these new confidence intervals and to contrast them with the Bayesian approach to weight estimation (see Aitken & Lucy, 2002).

It is advised to correct the formulas in the booklet as described by Stoel and Bolck (2009, in press), and to add a note that the resulting interval is not optimal if the proportion of positives is substantively smaller than 1 with. Furthermore, the Bayesian approach to weight estimation could also be considered for inclusion in the software, but this will require a lot of additional effort.

6 Discussion

This document described the validation of the tables and software as presented in the booklet 'Guidelines on representative sampling' of the ENFSI. Based on the results regarding the correctness of the probabilities produced by the program, it is concluded that sample sizes resulting from the hypergeometric, the binomial and the Bayesian approach are also correct.

Regarding the weight estimation some further attention should be given to an error in one of the Formulas. As far as we can judge the error attributed to a paper by Tzidony & Ravreby (1992), and subsequently resisted in the literature. Stoel & Bolck (2009, in press) have proposed a correction to this formula (see also the Section 5). From a statistical point of view it is advised to implement this correction in the booklet as well as in the software, although the consequences in practice will not be large.

One issue remains to be discussed. In the frequentistic approaches sample size is determined based a maximum value of a Type I error. That is, the probability of rejecting the null hypothesis, while it is true, should have a value smaller than, or equal to, α100%. Thus, in selecting an appropriate sample size the focus is on preventing the conclusion that the number of units is greater than some proportion, while it is not in the population. This approach ignores the probability on a Type II error, which is the probability of not rejecting the null hypothesis, while it is false.

In other words, the approach described in the booklet may not be the optimal approach from a statistical point of view, however, it may well be that the Type I error is the most important type of error in drug sampling. We let the answer to this question to be judged be the experts in the field.

As an illustration of the possible consequences of ignoring the Type II error, suppose a very large seizure is encountered, and the binomial approach is used for computing sample size. No negatives are expected, (1-α)100%=95%, and k=.5. The required sample size is than equal to n=5. Table 5 presents subsequent probability of a Type II error for several values of the true proportion of units containing drugs. It becomes clear that the probability of a Type II error is relatively large and much larger than the commonly advised level of .2 (see Cohen, 1988). The Type II error depends on the true amount of units of drugs in the population.

7 Literature

Aitken, C.G.G & Lucy, D (2002). Estimation of the quantity of a drug in a consignment from measurements on a sample. J Forensic Sci 2002; 47;968-975.

Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences Second Edition. New Jersey: Lawrence Erlbaum Associates

Stoel, R.D. & Bolck, A. (2009, in press) A correction to Tzidony and Ravreby (1992): 'A statistical approach to drug sampling: A case study'. J Forensic Sci.

Tzidony D., Ravreby M. (1992) Statistical Approach to Drug Sampling: A Case study. J Forensic Sci;37;1541-1549.

Alberink, I., Bolck, A. & Stoel, R.D. (submitted). Comparison of frequentist methods for estimating the total weight of consignments of drugs. Submitted for publication to J Forensic Sci.

Appendix 1

Table 1: withdrawn

Table 2: probabilities, 1-$P(X \geq x|\theta = k)$, of the binomial distribution as computed with the software, by hand, and with R.

k	***n-x***	***n***	**1-$P_{software}$**	**1-P_{hand}**	**1-P_R**
.5	0	3	.875	.875	.875
.5	1	3	.500	.500	.500
.5	2	3	.125	.125	.125
.7	0	20	.999	.999	.999
.7	1	20	.992	.992	.992
.7	2	20	.965	.965	.965
.9	0	30	.958	.958	.958
.9	1	30	.816	.816	.816
.9	2	30	.589	.589	.589

Note: the tabled probabilities are computed in R by means of the function:
1-pbinom(x-1, n , k)

Table 3: probabilities, $P(\theta>k|\, x,n)$, of the beta distribution as computed with the software and with R.

k	***x***	***n***	***a***	***b***	**$P_{software}$**	**P_R**
.5	3	3	1	1	.9375	.9375
.5	2	3	1	1	.6875	.6875
.5	1	3	1	1	.3125	.3125
.7	3	3	1	1	.7599	.7599
.7	2	3	1	1	.3483	.3483
.5	10	10	.5	.5	.9998	.9998
.5	9	10	.5	.5	.9963	.9963
.9	19	20	1	3	.1927	.1927
.9	18	20	1	3	.0731	.0731

Note: the tabled probabilities are computed in R by means of the function:
1-pbeta(x+a, n-x+b)

Table 4: probabilities, $P(Y>y|\, x,n,N)$, of the beta-binomial distribution as computed with the software and with R.

N	***k***	***x***	***n***	***a***	***b***	**$P_{software}$**	**P_R**
20	.7	3	3	1	1	.8327	.8327
49	.9	18	20	1	3	.1258	.1258

Note: the tabled probabilities are computed in R by means of the function:

Table 5: probability on a Type II error

k	***n***	Type I error (α)	True proportion of drugs	Type II error
.5	5	.05	.51	.97
.5	5	.05	.6	.92
.5	5	.05	.7	.83
.5	5	.05	.8	.67
.5	5	.05	.9	.41
.5	5	.05	.99	.05
.7	9	.05	.75	.96
.7	9	.05	.8	.92
.7	9	.05	.9	.86
.7	9	.05	.99	.06
.9	29	.05	.95	.77
.9	29	.05	.99	.25

Appendix 2

Withdrawn.