Risk, Confidence and Sampling

1. Objectives

Certification systems need an audit sampling strategy that meets two main criteria.

1. Credible.
2. Affordable.

1.1 Credibility

- A credible strategy is one that results in sufficient assurance of compliance, including the identification of non-conformities (NCs). 'Sufficient' assurance can be defined as '*that which ensures the objectives of the certification system are being met and the continued 'buy-in' from clients and customers*'.
- Also important is consistency of the level of assurance across different contexts.
- Credibility also depends on transparency, learning from results, and is openness to improvements proposed both internally and externally.

1.2 Cost, Sampling and Assurance

- 100% assurance (continuous, perfect auditing of everything, everywhere) is unaffordable, hence the need for a sampling approach.
- It is important to appreciate that with any sampling approach, no specific level of assurance or confidence[1] can actually be specified. So the goal is likely to be the maximum level of assurance that meets the two main criteria. The following table sets out three possible sampling approaches (alongside the 100% assurance option) and gives an indication in highlighted columns of the resulting level of assurance, consistency of assurance, and cost.

[1] 'Assurance' and 'confidence' generally have overlapping meanings in this context. 'Confidence' also has a specific meaning in statistical analysis and will be only used in that way in this document.

Sampling approaches	Description	Assurance that all NCs will be identified	Assurance that most NCs will be identified	Consistency of assurance across different contexts	Planning time	Audit cost	Total cost
1. Everything	'Sample' everything	High	Very high	Very high	Low	Extremely High	Extremely High
2. Random square root	Random 'square root' sample size across all sites/criteria	Very Low	Low-Moderate	Low	Low	Moderate	Moderate
3. Random	Random 'theoretical statistical significance' sample size (see below) across all sites/criteria[2]	Very Low	Moderate	Moderate	Low-Moderate	Moderate	Moderate
4. Risk-weighted	Random sample based on weighted risk assessments of sites/criteria	Low	High	High	High	Low-Moderate	Moderate

As can be seen a risk-weighted approach generally offers the best combination of assurance, consistency and cost-control. How this may work in practice is summarised in Section 3. But as all affordable approaches include some element of it is important to understand some of the statistical techniques and constraints of sampling in the 'real world'.

2. Statistical Confidence in Certification

- Statistical tables related to sample size and population size are readily available. These give confidence levels and intervals for different sample sizes, e.g. *'90% confidence that a sample represents the whole population within no more than +/-1%'*.

[2] I.e. an enhanced or better sampling size determinant than the square root 'rule'.

- It is very important to be able to answer the question '*confidence of what?*' Without having an answer to that question, all references to specific confidence levels (or intervals) are meaningless. It is also important to understand the difference between confidence levels and confidence intervals, and the relationship between the two. See box.
- What is less common is an understanding that statistical tables generally only apply to perfectly distributed populations in laboratory conditions. The real world is a very different place and it is all too easy to forget those differences when interpreting statistical tables.
- Annex A shows tables of confidence levels and intervals for different population and sample sizes. It also shows the weaknesses of the so-called square-root 'rule'.
- For any natural product certification, in no case can an absolute confidence level and interval be given[3]. This is because (a) no populations have the perfect 'normal' distribution required by those statistical tables, and (b) there is a whole series of more or less subjective decisions regardless of the sampling approach chosen, as well as external factors that affect results, such as the time of year an audit takes place. And no two auditors will perform identically. In other words in natural product certification we are actually considering real world samples, within real world samples, within real world samples.

Confidence levels and intervals

Say you take a random sample of 100 personnel out of a total workforce of 1,000 and ask them if they have been offered training. If 90 say 'yes' what can you say about training offered to the whole workforce?
Statistical tables will indicate that you could say with 71% confidence that if you asked all 1,000 then the number reporting they were offered training would be between 850 and 950. Or with 95% confidence the number would be between and 816 and 984. I.e. the greater the confidence level, the wider the confidence interval. Simply giving a confidence level without a confidence interval says little or nothing about your confidence.

3. A Risk-Weighted Approach

- One way to even out the quality of assurance provided and likelihood that NCs will be identified across different contexts is to use a risk-based system. Such a system can focus effort on those sites, species, operators, countries, and other factors where NCs are judged more likely to occur. The planning time/cost for such a risk-based system can be significantly more than other approaches, but as fewer/cheaper audits may be required the overall cost need not be any higher. And critically the confidence that most NCs will be identified will be significantly higher and the *consistency of confidence across*

[3] Even if the question 'confidence of what?' were to be precisely answered.

different contexts also significantly higher. In summary, a risk-based system assesses the likelihood that NCs may be found in different situations and then focusses more effort (i.e. more samples) in those situations.

3.1 Assessing Risk Factors

- A set of risk issues relevant to the particular context need to be developed. These might include country/region, scale of operation, length of operation, stability of ownership and management, management structure, diversity of operation and sources, number of sites, history of NCs, and product-related risks (such as novelty, history of NCs, diversity of sources).
- For each of these a simple scoring system may be devised, (e.g. a publicly available Corruption Index for country). If desired a slightly more developed system can consider likelihood and impact.
- A combination of the scores for each issue results in an overall risk rating. This need not be more complex than a 1-5, or a High/Medium/Low rating.

3.2 Focussing the Sampling Effort

There are two main ways that a risk assessment can efficiently focus auditing. The allocation of samples can be 'weighted' according to the risk rating. And the assessment can flag up particular issues that an auditor may be instructed to consider in their work.

- Guidance can be given to certification bodies on risk assessment, risk ratings and risk weighting. However the onus should be on the certification body to make and justify the final sampling decisions and they should be allowed to use professional judgment to adjust the system so long as such adjustments are explicit and justified.

A risk-weighted sample

Say an operation has 25 sites, and the risk assessment results in 2 being assessed as high risk, 10 as medium risk and 13 as low risk.
A simple algorithm could be produced to determine that an audit should span 5 sites, comprising both the 2 high risk sites, 2 of the medium risk sites and 1 low risk site. Note that without a weighted sampling system a random sample of 5 sites might easily have missed both the high-risk sites. A similar approach can be used for choosing samples of documents or personnel within an audit.

3.3 None Ruled In, None Ruled Out , and Rolling Samples

- In any publicly available sampling strategy it is important that no aspect should either be definitively ruled out or ruled in. Otherwise operators would know what they could ignore or would have advance notice of what was being checked.
- Various techniques may be used to incorporate 'none-ruled-in-none-ruled-out' into a risk-based strategy, e.g. by having 'time since last audit' as one of the risk factors, or by a

% of aspects determined by the risk-weighted approach with the balance determined entirely randomly.

- It is also possible to ensure that all sites, documents, criteria, etc. are audited over a specified maximum period of time. This may be achieved by setting a sampling rate of 100% for all aspects that have not been audited for a period equal to that maximum time.

3.4 Learning from Results

A risk-weighted sampling system offers great opportunities to learn and improve from experience. For instance it is possible to monitor the % (not just the absolute number) of NCs identified in High, Medium and Low risk rated samples and adjust the system accordingly.

4. Summary

- No absolute level of confidence can be attributed to a risk-weighted sampling approach, but this is true of *all* 'real-world' certification sampling systems.
- No absolute consistency of confidence can be attained with any real-world sampling system, and this is especially so with a sample size determined by the square root of the population size.
- A risk-weighted approach requires greater investment in planning, but offers greater likelihood of identifying any problems that might exist, and greater opportunity to learn and adapt, and all at moderate overall cost. In other words, focussing efforts on critical and high-risk issues allows greater levels of assurance for lower cost.
- If desired, the risk-weighted approach can facilitate a sampling system based on how many samples can be afforded. Once that number is determined, a sample allocation can be made that best meets the criteria.
- Any sampling technique needs to be tempered by experience and wisdom.

Annex A

Sample Sizes, Confidence Levels and Intervals

Table 1 shows the theoretical sample sizes required for a number of population sizes, to attain the most commonly required statistical levels and intervals of confidence.

Table 1. Sample Sizes, Confidence Levels and Intervals

Population size (e.g. number of sites)	95% confidence		99% confidence		Sample size if using square root 'rule'
	Sample size required for +/-5%	Sample size required for +/-1%	Sample size required for +/-5%	Sample size required for +/-1%	
4	4	4	4	4	2
6	6	6	6	6	3
8	8	8	8	8	3
10	10	10	10	10	4
15	14	15	15	15	4
20	19	20	19	20	5
25	24	25	24	25	5
30	28	30	29	30	6
40	36	40	38	40	7
50	44	50	47	50	8
75	63	74	67	75	9
100	80	99	87	99	10
500	217	475	285	485	23
1,000	278	906	399	943	32
10,000	370	4,899	622	6,239	100
100,000	383	8,762	659	14,227	316
500,000	384	9,423	663	16,055	708

Notes:

- The sample sizes in Table 1 result in theoretical confidence levels and intervals that will generally not be reached in real world situations.
- For these confidence levels and intervals, as Table 1 shows, until population sizes reach over 100, there is very little difference between the necessary sample size and the overall population size.
- For these confidence levels and intervals, as Table 1 shows, the square root 'rule' only provides a sufficient sample size when the population size is in the hundreds of thousands.

The Square Root 'Rule'

The so called square root 'rule' is commonly used to determine sample size, but has simplicity as its only major advantage. Table 2 shows theoretical confidence levels and intervals attained by using the square root 'rule' for different population sizes

Table 2. Confidence Levels Achieved using the Square Root 'Rule'

Population size	Square root 'rule' sample size	Confidence level for +/-10% confidence interval	Confidence level for +/-5% confidence interval	Confidence level for +/-1% confidence interval
4	2	27	13	1
8	3	31	16	3
15	4	30	17	3
25	5	37	19	3
40	7	43	22	4
75	9	47	24	5
500	23	67	37	7
1000	32	74	43	9

Notes:

- As Table 2 shows, the square root 'rule' does not provide high confidence levels, even for large population sizes.
- As the Table shows, the square root 'rule' does not give consistent confidence levels across different population sizes.