Interuniversity Institute for Biostatistics and statistical Bioinformatics

EFSA supporting publication 2015:EN-759

EXTERNAL SCIENTIFIC REPORT

Data Representativeness: Issues and Solutions[1]

Elasma Milanzi*, Edmund Njeru Njagi*, Liesbeth Bruckers, Geert Molenberghs[2, 3]

* Joint first authors

Center for Statistics, Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)
Universiteit Hasselt, Martelarenlaan 42, 3500 Hasselt. Belgium.

ABSTRACT

In its control programmes on maximum residue level compliance and exposure assessments, EFSA requires the participating countries to submit results, from specific numbers of food item samples, analyzed in the countries. These data are used to obtain estimates such as the proportion of samples exceeding the maximum residue limits, and the mean and maximum residue concentration per food item to assess exposure. An important consideration is the design and analysis of the programmes. In this report, we combine elements of survey sampling methodology, and statistical modeling, as a benchmark framework for the programmes, starting from the translation of research questions into statistical problems, to the statistical analysis and interpretation. Particular focus is placed on the issues that could affect the representativeness of the data, and remedial procedures are proposed. For example, in the absence of information on the sampling design, a sensitivity analysis, across a range of designs, is proposed. On the other hand, weighted generalized linear mixed models, and generalized linear mixed models combining both conjugate and normal random effects, are proposed, to address selection bias. Likelihood-based analysis methods are also proposed to address missing and censored data problems. Suggestions for improvements in the design and analysis of the programmes are also identified and discussed. For instance, incorporation of stratified sampling methodology, in determining both the total number, and the allocation of samples to the participating countries, is proposed. All through the report, statistical analysis models which properly take into account the hierarchical (and thus correlated) structure in which the data are collected are proposed.

KEY WORDS

Bias, Censoring, Clustering, Likelihood, Missing data, Stratification, Linear Mixed Models, Generalized Linear Mixed Models.

[1] Question No EFSA-Q-2013-00296.
[2] Elasma Milanzi, Edmund Njeru Njagi, Liesbeth Bruckers, Geert Molenberghs.
[3] Acknowledgement: The authors wish to thank the EFSA staff: José Cortiñas Abrahantes, Jane Richardson, and Daniela Brocca, for the support provided to this scientific output.

Any enquiries related to this output should be addressed to amu@efsa.europa.eu

Suggested citation: Elasma Milanzi, Edmund Njeru Njagi, Liesbeth Bruckers, Geert Molenberghs, 2015. Data Representativeness: Issues and Solutions. EFSA supporting publication 2015:EN-759, 159 pp.

Available online: www.efsa.europa.eu/publications

DISCLAIMER

The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors

SUMMARY

EFSA requires EU member states, as well as two EFTA countries (Iceland and Norway), to provide data on a specific number of food samples analysed within these countries. This is in the framework of EFSA's control programmes on maximum residue level compliance, and exposure assessments. These data are used by EFSA to obtain various estimates, for example the percentage of samples exceeding the maximum residue limits, or the mean and maximum residue concentration of residues for the various food items to assess chronic and acute exposure.

An important consideration in the design and analyses of the programmes is conformity to good statistical practices. This is to ensure that appropriate and representative data are accrued to answer the research questions of interest, and that the data are analysed and interpreted in the context of appropriate statistical tools.

In this report, we consider the design and analysis of the programmes, focusing on the pesticides monitoring data of 2010 as a case study. Of particular focus is to identify issues which could affect the representativeness of the data, hence the results, and propose appropriate remedies. This assessment is guided by good survey sampling practices, and statistical modelling.

After an introduction to sample survey design, we discuss various sampling designs, including simple random, cluster, and stratified sampling. We then illustrate sample size calculations under the various designs, including multistage sampling. The pesticides monitoring programme is then evaluated, in terms of the definition and specification of the target population, the sampling frame, statistical objectives of the programme, the sampling design, and sample size calculations.

Given the multi-national nature of the programme, and the possible intra-cluster correlation arising from clustering of the samples, stratified and clustered designs are highlighted as important frameworks for sample size evaluations. Stratification is discussed as a convenient framework for the allocation of the total number of samples to the various countries. We compare different stratified allocation strategies, including proportional allocation on the basis of population, and food commodity consumption. The need to translate the research objective into either an estimation or hypothesis testing problem is also discussed.

Further, we explore the problems that may affect the representativeness of the sample. Simulation studies are conducted to illustrate the effect of selection bias. In general, when elements with higher/lower values of the outcome are given higher chances of being selected into the sample, and this is not recognized during analysis, positive/negative bias is potentially induced. In addition, selection bias leads to reduced power or inflated type I error (depending on the direction of the alternative hypothesis and bias).

Similar bias occurs when elements with low/high values of the outcome are given zero probability of selection. This non-coverage bias is also illustrated through simulations.

Sample size sufficiency is also discussed. It is illustrated that using an insufficient sample size reduces the precision of the estimates.

The effects of missing data, and left censoring, are also illustrated. It is shown that in the presence of missing data, careful attention needs to be given to the analysis methods, as substantial bias could be induced, depending on the missing data mechanism.

The effects of discrepancies between the sampling design and analyses are also discussed, showing that failure to take into account the appropriate design during analysis affects inference, through impacts on both precision estimates and type I errors.

The pesticides monitoring data are used to explore some of these problems, and statistical methods for dealing with the problems are discussed. These methods include weighted generalized linear mixed models, and generalized linear mixed models with different sets of random effects, to address selection bias. Likelihood-based methods are also proposed to deal with missing data and left censoring problems, and example analyses are provided, based on the pesticides monitoring data.

TABLE OF CONTENTS

BACKGROUND AS PROVIDED BY EFSA

In recent years EFSA has developed its capacity and procedures for receiving monitoring and survey data collected within the European Union (EU). EFSA receives Zoonoses monitoring data under Directive (EC) 2003/99, contaminants monitoring data under Regulation (EC) 2002/32 and pesticides monitoring data under Regulation (EC) 396/2005. Considerable work has been done to standardise and harmonise these programmes and the reporting of the subsequent datasets. However definition of the target population, the sampling unit and the method to select the unit for inclusion in the survey/monitoring programme is critical to ensure a representative sample appropriate for statistical analysis and exposure assessment, for which the results could then be generalizable and repeatable.

Information regarding survey design, randomness, hierarchy and other factors could be influencing the inference process. Representativeness of the data collected is crucial during the assessment process. EUROSTAT has developed a quality framework for the statistical processes and outputs (see Code of Practice[4] and Definition of Quality in Statistics[5]) and identified a set of principles that should guide the process of conducting surveys and in the assessment of their quality. It is important to know the impact of these issues and their influence on the representativeness of the sample taken to answer specific objectives and how such issues could be dealt with to improve EFSA assessments. In this context it is important to explore potential sources of bias, methods to correct for bias and assess the impact of dealing with samples that might not have been selected to address the objective of the assessment. Moreover, in the 2009 European Union Report on Pesticide Residues in Food[6] EFSA made the following recommendation *"To revise the general design of the EU-coordinated multiannual control programme, taking into account the increased number of reporting countries. In particular, a new calculation of the total number of necessary samples to be analysed for each commodity and the allocation to the individual Member States and reporting countries should be performed"*, which is in line with the concerns previously mentioned. Thus pesticide monitoring data will be used as a case study to explore the source of bias, ideal design, as well as samples needed and the possibility of answering more than one objective maintaining valid assessments for each of the specific objectives under consideration.

Pesticide Monitoring Program: A Case Study

According to the EU legislation in place in 2009, EU Member States and two EFTA countries (Iceland and Norway) have to carry out national control programmes on pesticide residues in food commodities and to report the results to the European Commission and EFSA.

General legal provisions for food inspections and monitoring were established by Regulation (EC) No 882/200412 on official controls performed to ensure the verification of compliance with feed and food law, animal health and animal welfare.

Article 15 of Regulation (EC) No 882/2004 lays down that the national competent authority shall carry out regular official controls on feed and food of non-animal origin imported into the territories. They

[4] http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF
[5] http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/ess%20quality%20definition.pdf
[6] European Food Safety Authority; 2009 EU Report on Pesticide Residues. EFSA Journal 2011; 9(11):2430. [226 pp.] doi:10.2903/j.efsa.2011.2430. Available online: www.efsa.europa.eu/efsajournal

shall organise these controls on the basis of the multi-annual national control plan. These controls shall be carried out at appropriate places, including the point of entry of the goods into one of the territories.

To fulfil the requirements of Regulation (EC) No 882/20047 and Regulation (EC) No 396/20058 on pesticide Maximum Residue Levels (MRL), EU Member States perform official controls to ensure the compliance of feed and food samples with regard to the pesticide MRL legislation.

Regulation (EC) 396/2005 requires member states to collect samples under the EU coordinated multiannual control programme and under national control programmes. On the other hand the design of the national control programmes is under the responsibility of the Member States and is therefore not subject to this reopening competition. The EU-coordinated programme aims to provide statistically representative data regarding pesticide residues in food available to European consumers. The lots sampled should be chosen without any particular suspicion towards a specific producer and/or consignment. Thus, the results obtained in the coordinated programme are considered as an indicator for the MRL compliance rate in food of plant and animal origin placed on the European common market and they allow an estimation of the actual consumer exposure (both acute and chronic).

The establishment of a coordinated community programme was initiated in 1996. Since then, the number of participating reporting countries has increased; in 1996, 15 EU Member States and one EFTA State (Norway) reported their control results, whereas in 2009 the number of participating countries was 29: 27 EU Member States and two EFTA countries (Norway and Iceland) who have signed the Agreement on the European Economic Area (EEA agreement). Over time, the programme was also extended with regard to the number of samples, the food commodities and the active substances to be analysed each monitoring year.

Every year, the European Commission prepares a specific control programme (which is published in a Commission Regulation) describing the pesticide/crop combinations that have to be analysed. The programme takes into account food items which are of relevance for human consumption (the major components of the European diet (food of plant origin) are represented by 20 to 30 food products) and pesticides which are of relevance for dietary exposure because of their toxicological profile or specific problems identified in previous years.

Thus, the coordinated EU programme is defined in terms of:

- number of samples to be analysed by each Member State;
- the food items to be sampled and analysed;
- the list of the pesticide to be analysed in each food sample.

The list of EC Regulations laying down the EU monitoring programmes is available at: http://ec.europa.eu/food/plant/protection/pesticides/multi-nnual_control_programmes_en.htm

[7] Regulation (EC) No 882/2004 of the European Parliament and of the Council of 29 April 2004 on official controls performed to ensure the verification of compliance with feed and food law, animal health and animal welfare rules; OJ L 165, 30.4.2004, p. 1–141.
[8] Commission Regulation (EC) No 395/2005 of 9 March 2005 providing for reallocation of import rights under Regulation (EC) No 1206/2004 opening and providing for the administration of an import tariff quota for frozen beef intended for processing; Official Journal L 063 , 10/03/2005 P. 0020 - 0020.

Member States set up **national control programmes** for pesticide residues. Those programmes are often risk-based and focus on commodities and/or pesticides which are considered of particular relevance for consumer safety or MRL compliance. The official controls carried out at national level within the framework of the national control programmes are complementary to the controls performed in the context of the EU-coordinated programme. They are performed to ensure compliance with the provisions established in food legislation regarding pesticide residues. The reporting countries have to define their priorities regarding the design of the national control programmes for pesticide residues in food.

In designing their national control plans, the reporting countries typically take into account the following factors:

- Importance of a commodity in national food consumption;
- Food commodities with high residues/non-compliance rates in previous years;
- Food consumed fresh or in processed form;
- Balance of organic/conventional production;
- Origin of food: domestic, EU or third countries;
- Sampling at different marketing levels: farm gates, wholesalers, retailers, processing industry, schools or restaurants;
- Seasonal availability of food commodities;
- Crops with high RASFF notification rate;
- Food for sensitive groups of the population, e.g. baby food;
- Geographic representatives for the reporting country/cultivation area;
- Food produced by producers with non-compliance in the past;
- Food commodities not included in the EU-coordinated programme.

Regarding the pesticides included in the national control programmes, the reporting countries consider:

- Use pattern of pesticides;
- Pesticides notified in the RASFF
- Toxicity of the active substances;
- Cost of the analysis: single methods/multiple methods;
- Capacity of laboratories.

More details on the design of the national control programmes are reported in Appendix II of the published EFSA Annual Reports on pesticide residues[9]. The number of samples and the analytical scope of the analysis performed by the participating countries are strongly determined by national budgets. Thus, reporting countries have to focus on the specific aspects which are considered most relevant for their national control activities. These results are of value for assessing the MRL compliance at national level; however, due to the variability of the programme designs, the comparison of results from different reporting countries needs to take into account the different focuses of the national programmes.

The sampling strategies for these programmes are specified in Commission Directive 2002/63/EC. This describes the procedures for taking samples of fruit, vegetables and products of animal origin, and a revision of this Directive is not the subject of this reopening competition.

The pesticide programme under Regulation (EC) 396/2005 should be used as a case study considering the different purposes and objectives of the programs, in addition changes to the methodology for the collation of the data has resulted in the availability of a detailed dataset which is comparable at EU level. A review of the survey design for both programmes could identify weaknesses and recommend methods to adjust for sampling bias and lack of accuracy in the data analysis and final exposure assessment. Additionally a proposal could be made for improvements to the survey design to ensure a representative sample is selected from the target population and guarantee robust risk assessments. The findings of this project could be linked to other surveys and monitoring programs operating in EFSA and should potentially develop methodological frameworks that could deal with representativeness issues in other areas.

TERMS OF REFERENCE AS PROVIDED BY EFSA

The overall objective of this reopening competition procedure under EFSA's Framework Contract for provision assistance for statistical analyses, data management and ad hoc consultancy upon request is in relation to: (i) Assessment of the quality of main data sources most commonly used for EFSA risk assessments in term of representativeness and fit for purpose, with special focus on pesticide monitoring data, (ii) Explore and study impact of issues in relation to: design used to collect data that might be used to answer questions for which the collection was not designed for, sample size used and population characteristics that might be ignored during the inference process which could potentially bias the process and (iii) how could bias be dealt with and correct inferences be obtained when the previous issues are encountered.

For the completion of the above objectives, the data from the two pesticide monitoring program under Regulation (EC) 396/2005 will be provided to the selected contractor following specific contract signature and will serve as case study in order to set the methodological framework that could be used to assess other monitoring/surveys data.

Specific objectives

Objective 1: Assessment of the quality of main data sources most commonly used for EFSA risk assessments in term of representativeness and fit for purpose.

[9] The EFSA Reports on the Pesticide Residues are available at: http://www.efsa.europa.eu/en/pesticides/mrls.htm

Objective 2: Assessment of the impact of design, sample size used and population characteristics that might be inappropriate or ignored during the inference process.

Objective 3: Investigate methods to deal with the types of bias and issues identified in the previous objective that could be used to propose potential corrections to ensure reduction or elimination of bias results from the inference process.

This contract/grant was awarded by EFSA to:

Contractor/Beneficiary: Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Martelarenlaan 42, 3500 Hasselt, Belgium.

Contract/grant title: Data Representativeness: Issues and Solutions

Contract/grant number: RC/EFSA/SAS/2013/01

INTRODUCTION AND OBJECTIVES

Quality of data sources used in producing various statistics is crucial in ensuring dissemination of reliable information. The European Food Safety Authority (EFSA) which is mandated to collect data used in risk assessment is thus presented with the challenge to ensure that quality information is used in making such assessments which are essential in policy making. Data representativeness is crucial for such mandates.

In general, data representativeness refers to a phenomenon where the collected data (a sample) accurately reflects the population under study. Ramsey and Hewitt (2005), note that assessment of data representativeness is only possible after clearly stipulating the targeted population and the purpose for collecting the data. Having a large sample does not imply representativeness; rather the manner in which the sample was collected plays a big role in ensuring representativeness. If selection of the sample is biased towards elements that either have the desired characteristic of interested or have similar characteristic, then even a generously large sample will not deliver a representative data. Population characteristics estimated from such data will be biased towards the preferred elements.

Bias is usually difficult to identify and correct, especially when the source is unknown to the investigator. Conventionally, bias is avoided by employing principled methods of sampling design that aim to minimize the risk of bias. With a well-designed probability sample, selection bias is minimal. This is the chief advantage of probability samples over non-probability sampling. In instances where the sample has already been obtained and modifications of the design are no longer possible, correction approaches can be considered. This would, however, require information regarding the existence and nature of the bias in question.

When the bias is unknown to the investigator, no correction-based approaches to the inference process are possible. On the other hand, for surveys conducted regularly, previous surveys provide a good platform to identify possible causes of data non-representativeness or sources of bias. Using data collected through the pesticide monitoring program by EFSA, this work aims at assessing the representativeness of the data EFSA uses in risk assessments through:

(i) Assessment of the quality of main data sources most commonly used for EFSA risk assessments in terms of representativeness and fit for purpose.

Data representativeness is mainly dictated by use of appropriate sampling design; hence this objective will be achieved by appraising current survey design practices against the required basic good survey practices. A brief review of fundamental concepts for survey design practices will be given to provide a general framework for collecting representative data, and to serve as a check-list when evaluating current survey practices in the pesticide monitoring data.

Issues like definition of targeted population, availability of clearly stated objective(s), a recognizable sampling design, a well-defined sampling frame and sample size calculation methods will be examined.

(ii) Assessment of the impact of design, sample size used and population characteristics that might be inappropriate or ignored during the inference process.

Under this objective we will investigate impact of possible causes of non-representativeness identified in Objective **(i),** with specific reference to pesticide monitoring data. It is always reasonably assumed that estimates obtained from samples are associated with ***sampling error***. Sampling error is defined as the statistical imprecision that arises when only a subset (or sample) of the population of interest is used to obtain an estimate of a given population characteristic. It is generally defined as the difference between the actual value of the population characteristic and an estimate obtained from a sample. This estimate is generally not equal to the true value of the characteristic because of sampling variability (i.e. the estimate will vary from sample to sample) and bias. When either or both of the latter are large, sampling error is large, possibly leading to a non-representative sample. To ensure a representative sample, it is essential that sampling error is controlled to be within acceptable limits and a reasonable sampling design is employed to minimize bias.

While the exact quantification of the sampling error is generally not feasible, since the true population values are unknown, the extent of sampling error can often be estimated with knowledge of the probabilistic nature of the sample selection. When probability sampling is employed, the random variation arising from sampling (i.e. observing only a part of the population) can be estimated.

On the other hand if some fundamental concepts in probability sampling are ignored both sampling error and bias may go beyond accepted limits thereby affecting representativeness of the data. In this regard, we will investigate the impact of ignoring or inappropriate use of some survey design concepts namely; ignoring sampling design during data analysis, failure to account for selection probabilities prescribed under the study design during analysis, use of non-exhaustive sampling frame that results into assigning some eligible elements zero probability of being selected, non-response (missing data), measurement errors, e.g., failure to measure concentration of residue because it is below detection limit, and use of insufficient sample size.

A simulation study will be used to investigate the general impact, and the pesticide monitoring data will illustrate the same for EFSA datasets.

In non-probability sampling, the relationship between the target population and the survey sample is immeasurable, making it difficult to identify potential sampling biases. As such, random error cannot be estimated without reference to some probabilistic model that would plausibly describe how the sample was selected.

(iii) Investigate methods to deal with the types of bias and issues identified in the previous objective that could be used to propose potential corrections to ensure reduction or elimination of bias from the inference process.

Methods to deal with bias and inflated sampling error identified in objective **(ii)** can be categorized into pre-survey and post-survey. The former has to do with design modifications or corrections that would help reduce or eliminate bias from the inference process, while the latter includes inferential procedures that provide some degree of robustness against bias.

In general, sampling error can be reduced by increasing the sample size. In well-designed surveys, sample sizes are typically computed to achieve a certain level of precision (i.e. to control the variability component of the sampling error to some tolerable amount). The bias component, however, cannot usually be addressed by an increase in the sample size. Even a

large sample cannot correct for methodological problems like under-coverage, measurement errors, or nonresponse.

MATERIALS AND METHODS

1. Assessment of the Quality of Main Data Sources Most Commonly Used for EFSA Risk Assessments in Terms of Representativeness and Fit for Purpose.

GENERAL FRAMEWORK

To provide cornerstones for appraising representativeness of data collected under current survey practices, a brief overview of fundamental concepts in survey design, and sampling designs are outlined in sections that follow. More information can also be found in Eurostat (2008)[10].

1.1. Introduction to sample survey design

Sample survey design entails all the processes and considerations concerned with obtaining descriptive or inferential statistics of population of interest by studying just a portion of the population instead of the whole population (Kalton, 1983; Kish, 1965). Compared to studying the whole population (Census), a survey has several advantages like cost-effectiveness; indeed studying the whole population will require more financial and human resources than concentrating just on a part of it. Further, a sample survey will require less time than census hence the required statistics are likely to be obtained in good time when they are still relevant. Importantly it is not always feasible to study the whole population. All these advantages apply when the survey is designed in adherence to scientific guidelines which help control some of the errors that may arise due to studying part of the population in place of the whole population (Stopher and Meyburg, 1979). The guidelines are just a collection of interrelated decisions on factors such as mode of data collection, method of processing the data and sample design (Kalton, 1983, pp6). It is vital that every decision is made with the aim of designing a sample survey that is representative of the population understudy.

An initial stage in designing a sample survey is a clear definition of the targeted population, and the elements, i.e., the units that make up the population from which information is sought. For instance, EFSA is mandated to collect data from European Union (EU) member states on a wide range of topics like, pesticide monitoring in food items and monitoring zoonoses and food-borne outbreaks in humans, food and animals, surveillance for residues of chemical elements in foods of animal origin. Although the data for each topic is obtained from EU member states the elements of the targeted population are different. However, the principles of survey designing are universal and some of the recommendations in this report could be applied to other surveys coordinated by EFSA. In the case of pesticide monitoring, the elements are obtained food items only, while for zoonoses and food-borne, humans are also contribute sample elements. In addition to recognizing the elements, a clear definition of the population has to be stated. For example, in the pesticide monitoring study the population can either be defined as, all the apples available for consumption in the EU member states in the year 2010, or all the apples on market in the EU member states. Note that while the former definition includes apples that are still in the farms in the year 2010 the latter does not, hence a careful and specific definition of the targeted population is a crucial starting point in designing a survey.

Logically, the definition of the population should be intertwined with the objectives of the sample survey. Objectives can broadly be divided into two groups: estimation and inferential. Estimation

[10] Eurostat (2008): http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-08-003/EN/KS-RA-08-003-EN.PDF

objectives mainly involve production of quantitative and numerical descriptions (estimation) of relevant aspects of a target population, like the population mean or the population total, mean difference between two groups of the same population, and proportion of the population with a trait of interest, etc. On the other hand, inferential objectives are about testing a particular hypothesis about the population of interest, examples include, testing that the population mean is greater (less) than a certain value, or that means of groups of the same population are not equal. For instance, in the zoonoses and food-borne data collection, a descriptive objective would be stated as, "*the objective is to estimate the number of salmonellosis cases in humans in the EU member states in the year 2010*" and an inferential objective can be formulated, "*the objective is to test the hypothesis that the number of salmonellosis cases in humans in the EU member states in the year 2010, is greater than 100000*". An important difference between the two objectives is that the inferential objective requires specification of power of testing, in addition to level of type I error required in the estimation objective. When a survey is conducted with the aim of estimating a parameter of interest in a population, some level of certainty (usually expressed as a confidence interval) is associated with the estimate. Confidence intervals give a range of values in which we believe the true parameter value lies, and if the true value does not lie in the estimated range, we commit a type I error. The probability of committing this error, is pre-specified in advance and incorporated in sample size calculation during a survey design so as to keep it under control. Similarly, when a survey's objective is to test an alternative against a null hypothesis, type I error is committed when mistakenly reject a true null hypothesis. The power of testing a hypothesis is determined by the probability of correctly rejecting a false null hypothesis. It will later be shown that this affects the sample size needed for the different objectives.

After clearly defining the targeted population and the goals of the survey, issues on how to decide on the portion of the population that needs to be included in the survey can be addressed. Such issues are collectively referred to as, sample design. A choice has to be made between using probabilistic or non-probabilistic sampling methods.

The main characteristic of non-probabilistic sampling methods is that elements are chosen arbitrarily and it is not possible to associate each element with a probability of being selected. Examples include: (i) Convenience sampling, where elements are selected if they can be easily and conveniently accessed, (ii) Volunteer sampling, where elements are included upon volunteering, (iii) Judgement sampling, where the researcher decides on the elements that are likely to be representative of the population and hence selected into the survey (iv) Quota sampling, sampling is done until a specific number of units (quotas) for various sub-populations have been selected. Non-probabilistic methods are prone to subjectivity and may affect the representativeness of the realized sample. Due to arbitrariness in the selection of elements, it is difficult to quantify the impact that non-representative sample would have on survey results. Nevertheless, in some instances non-probabilistic methods may be the only option.

In probabilistic methods, every element in the population has a non-zero probability of being selected thereby minimizing subjectivity, and several choices exist that ensure representativeness of the sample. For example, Czech Republic assigns selection probabilities to commodities by taking into account factors like, consumption of foodstuffs as elaborated by National Institute of Public Health, place of origin (EU, inland or third countries) and their corresponding representation on the market. Similarly, Iceland selects commodities proportional to their place of origin and volume on the market. We therefore focus on probabilistic methods.

All probabilistic methods assume existence of a sampling frame, from which elements can be selected. This can be in form of a list of all elements in the population or some equivalent procedure identifying

the elements in the population. In the example of pesticide monitoring, it is impractical to list all the apples available for consumption in the EU member states, as such a sampling frame can be defined as all areas that can have apples, e.g., supermarkets, farms, open markets, warehouses, etc. Within the sampling frame, sampling units also have to be defined, these are the units that will actually be selected, and these might be the individual elements or groups that contain the population elements. The definition and organization of the sampling frame/units is one of the factors that influence the choice of the sample design.

Other factors that need to be considered in choosing the sample design are objectives orientation, measurability, practicality and cost. Take an example of Ireland; their national pesticide monitoring program also considers the requirements of EU coordinated program, and dietary intake patterns (objective orientation), the residue profile of commodities as established from previous studies (measurability), and capacity of the laboratory (practicality).

The importance of selecting a sample that will achieve the pre-specified goals cannot be overemphasized. Measurability refers to the sample design that will allow computation of valid estimates or approximations of its sampling variability. These are necessary for statistical inference but also allows for assessment of the gap between the values from the sample and those from the whole population which are usually unknown. Practicality of the design is essential to ensuring correct execution of the whole survey. For example, for a chosen survey design one should be able to clearly state the feasible guidelines on how, when or where to collect the sample. The cost of conducting a survey is a major player in many decisions involved in designing a survey. Factors like, objectives, desired precision and/or power of testing a hypothesis can be altered in order to stay within the available budget. Some designs are more costly than others, and usually the costly designs have higher level of precision than their less costly counterparts. In general choosing a sample design will require input from several interested parties and trade-offs are inevitable. These trade-offs should be well documented and be integrated (if possible) in production of the population statistics. Note that estimates of the population characteristics and sampling variability approximation depend on the sample design, thus a survey is basically identified by its sampling design. A more detailed description of the sampling designs will be given in the sections to follow.

1.2. Various Sampling Designs

1.2.1. Simple Random Sampling

Simple random sampling (SRS) is the simplest form of drawing elements from a targeted population. It involves drawing elements successively such that each population member has equal and a non-zero probability of being selected, Barnett, (2002). Assume we have a population with N elements and we would like to draw a sample of n elements. For selection with replacement, i.e., a selected element is returned in the population and thus can be selected more than once, each population element has a selection probability of $\frac{1}{N}$ at each sampling turn. Otherwise if selection is without replacement selection probability changes at each sampling turn, i.e., at first sampling turn each element has $\frac{1}{N}$ selection probability, $\frac{1}{N-1}$ at the second turn, etc. When sampling is without replacement n cannot exceed N, while if sampling is with replacement n can be any value. Many statistical theories assume sampling with replacement, (Kish, 1965). Sampling units are the individual population elements.

Though theory and mathematical properties of SRS are well developed, it is rarely used in practice, mostly because it is not feasible, for example, in the pesticide monitoring study a numbered list of apples would be required to perform a randomized selection process. When the population is too large

and sparse, selected elements may be very far apart thereby decreasing efficiency in executing the survey and increasing the costs. These and other practical considerations make SRS the least popular design in practice. Nevertheless it is the basis of all the other designs such that in some situations computations from SRS can be used to approximate those from other complex designs by adjusting with some known factors, hence its properties are useful. The precision of other sampling designs is usually compared to precision in SRS.

1.2.2. Cluster Sampling

In SRS the sampling units are individual population elements, i.e., each sampling unit has only one element. As noted in the section for SRS, this method of selecting elements is not always viable. This might be due to inexistence of the complete list of the population elements or huge expenses associated with collecting such a sample. In such cases it may be useful to select groups of elements rather than individual elements, such groups are known as clusters. Clusters are a composition of several population elements. For example, in selecting a sample for monitoring zoonoses in animals within a member state (MS), it is more practical to select herds (these can be easily enumerated) and then select holdings/houses within each herd and include all the animals from a selected holding into the survey. In this example, both herds and holdings are clusters since they are both made up of a group of population elements (animals). Note that each population element can only be in one cluster at a particular time. It is important that the defined clusters do not overlap.

The obvious advantage of cluster sampling over SRS is its cost-effectiveness in terms of listing and locating the elements (Kalton, 1983). The major drawback is the increase in element variance. In general cluster sampling should be used when the gain in expense reduction is significantly larger compared to the lower precision.

The nature and size of the selected clusters determines whether all the elements in the selected cluster are included in the survey or further sampling within the cluster is needed. When the clusters are very large, like it would be the case with cities in the example given above, sampling of elements can be done in two phases: firstly, the MS is divided into clusters and SRS is used to select the required clusters. Secondly, within each selected cluster a sample of elements is drawn. This is referred to as sub-sampling and it can be extended to more than two phases when necessary. Re-consider the example of zoonoses monitoring within a MS: herds may be selected as primary sampling units. In the second phase, holdings/houses will be sampled; holdings in this case will be secondary sampling units. Within the selected holdings, a random sample of animals can be selected as tertiary sampling units. When the final cluster size is small all the elements in the selected clusters can be included in the sample, otherwise another stage of sampling may be required.

Impact of Clustering

It is well known that the information contained in such a sample is less than the information that would have been in the same sample assuming independence. Elements within the same cluster are likely to be more similar than elements between clusters. The strength of this similarity is quantified using the coefficient of intraclass correlation(ρ). Regular correlation ranges from -1 to 1 with zero correlation implying independence of the elements, correlation of 1 imply that elements within the same cluster provide exactly the same information. However, negative values of intra-class correlation are not theoretically possible. A regular correlation of -1 implies that elements within the same cluster give exactly opposite information of each other (negative correlations are rare in surveys). Thus amount of information in clustered data depends not only on the cluster size, but also on the structure and strength of the correlations among observations from the same cluster (Faes, et.al, 2009).

The impact of clustering is assessed through the design effect(D), defined as the ratio of variance of the estimate under SRS to variance under the design of interest, cluster sampling in this case. For $\rho > 0, D > 1$ and this means variance from cluster sampling sample is larger than variance from SRS sample with the same sample size. For the maximum value of $\rho = 1$, D equals the cluster size, thus we would require sample size D time larger under cluster sampling than what would be required under SRS, and $\rho = 0$ corresponds to $D = 1$, that is the variance in the cluster sample is the same as that in SRS for the same sample size. In general bigger cluster sizes and large intraclass correction give high values of$D$. Design effect can also be used to obtain the effective sample size, i.e., the sample size one would need in an independent sample to equal the amount of information in the actual correlated sample.

1.2.3. Stratified Sampling

When the population of interest falls naturally into groups, sampling may be organized within each of these groups. Such groups are known as strata. As an example, for a sampling exercise encompassing the EU member states and the 2 EFTA countries, each of these states (countries) could be considered a stratum.

In this type of sampling, the characteristic of interest is surveyed and analyzed within each stratum, after which the results are combined, to provide an overall sample result. Within each stratum, various sampling procedures may be used; for instance, simple random sampling, or cluster sampling.

Apart from administrative convenience, and the ability to make inferences about each stratum, such a design has, under certain conditions, potential for greater statistical precision in estimating the quantity of interest.

Stratification has found applications in many settings. In ecological monitoring, the Countryside Survey in Great Britain has used stratification by environmental factors to capture the land's heterogeneity, and stratification has been proposed as a framework for national, European, and global ecological monitoring (Metzger et al., 2012). Stratification has also been proposed in sampling of *Salmonella* isolates for monitoring of microbial resistance (EFSA, 2014)[11]. Regional and provincial stratification is used in the Belgian Health Interview Survey.

An important consideration in stratified designs is the allocation of the total sample size to the various strata. This can be done using different approaches:

1. Proportional Allocation

In this approach, a uniform sampling fraction is used across the strata. The sample size allocated to each stratum is proportional to the stratum size.

2. Neyman Allocation

Assuming equal costs across strata, the allocation that focuses on minimizing sampling variance is called the Neyman allocation. Strata which have more variability are allocated a larger sample size.

[11] EFSA (European Food Safety Authority), 2014. Technical specifications on randomised sampling for harmonised monitoring of antimicrobial resistance in zoonotic and commensal bacteria. EFSA Journal 2014;12(5):3686, 33 pp. doi:10.2903/j.efsa.2014.3686

As noted from the preceding sample designs, each design has both strong and weak points. In practice a combination of these designs is often used to minimize and maximize the weak and strong points of each design, respectively. For example, it can be shown that proportionate stratification is more beneficial when sampling units are clusters rather than individual population elements (Kish, 1965). In other situations like surveys cutting across geographical boundaries, combination of several designs is mostly inevitable due to factors such as: spatial nature of the population, language and economical differences. This allows for increased design flexibility and minimization of costs although the statistical analysis may become complicated (Harkness, et.al, 2010). Multistage sampling is also convenient where naturally occurring clusters are larger than desired. Going back to the example of monitoring zoonoses, it is more convenient to consider each MS as a stratum, this will give MS the flexibility to design a survey best suited to their populations. Within each strata, cluster sampling and SRS can be employed to come up with the final elements to be included in the sample. Multi-stage design is the most practical design due to its degree of flexibility. It is common to combine stratification, cluster sampling and SRS.

1.2.4. Designs for Measuring Change Over Time

It is usually tempting to compare results of a particular survey to similar surveys from the past with the aim of assessing change over time. This should essentially be possible if the same variable was measured in the different surveys. If measuring change over time is the main objective of the survey, it is important to outline this clearly from the beginning, because measuring change based on surveys designed to measure a different quantity, e.g., population mean, may result into less precise estimates or low power to detect the change.

Distinction is usually made between gross and net change. For example, in the zoonoses and food-borne outbreaks monitoring study, the change in number of salmonella cases in humans from 2007 to 2010 may be measured as follows: select human elements into the 2007 survey and estimate the number of salmonella cases, follow the same elements at some pre-specified time intervals for the whole period 2007 to 2010. At each time interval, estimate the number of salmonella cases. By the end of study period (2010), the evolution of the number of salmonella cases between 2007 and 2010 can be estimated. The crucial characteristic of this method of measurement is that it allows tracking of individual elements' changes. Alternatively, after estimating the number of salmonella cases in 2007, we can collect another independent sample in 2010 and compute the required estimate. The change in the number of salmonella cases is obtained as the difference between the estimates from the two years. The first method measures the gross change while the second measures net change. Choice of which measure to use totally depends on the objective(s) of the survey.

In general repeated survey designs are recommended for measuring change. These can either be panel designs or repeated cross-sectional surveys. Panel designs allow measurement of both net and gross change while repeated cross-sectional surveys only allow for gross change.

A longitudinal survey is a well-known form of panel designs where the initial selected sample is followed for the whole period of the survey and they produce precise net change estimates. To put things in perspective, consider survey conducted at time t and $t+1$, where interest is in estimating change in the mean of variable y $(\bar{y})$ between time t and $t+1$, $\Delta = \bar{y}_{t+1} - \bar{y}_t$. It can be shown that

$$\mathrm{Var}(\Delta) = \mathrm{Var}(\bar{y}_{t+1}) + \mathrm{Var}(\bar{y}_t) - 2\sqrt{\mathrm{Var}(\bar{y}_{t+1})\mathrm{Var}(\bar{y}_t)}\,\mathrm{Corr}(\bar{y}_{t+1}, \bar{y}_t).$$

It follows that the Δ will be estimated more precisely if $\mathrm{Corr}(\bar{y}_{t+1}, \bar{y}_t)$ is high and positive. The best way to attain high and positive correlations is to use the same elements both at time t and $t+1$, this is

achieved with longitudinal survey. Medium and positive correlations can be obtained if there is some level of overlapping between elements at time t and $t+1$, which can be realized through another form of panel designs, referred to as rotating panel surveys. In this design, the sample at $t+1$ will partially be composed of elements from the sample at t, hence change will be estimated with medium precision. Given that negative correlations are very rare in surveys, $\mathrm{Var}(\Delta)$ will be the highest when$\mathrm{Corr}(\bar{y}_{t+1}, \bar{y}_t) = 0$, and this corresponds to repeated cross-sectional surveys where the elements for the sample at t are different from (independent of) the elements in the sample at $t+1$. Note that while repeated cross-sectional surveys will lead to less precise change estimates than longitudinal surveys, the former produces highly precise population mean estimates ($\bar{y}_t$ and $\bar{y}_{t+1}$) than the latter. Indeed note that,

$$\mathrm{Var}(\bar{y}_{t+1}) = \mathrm{Var}(\Delta) - \mathrm{Var}(\bar{y}_t) + 2\sqrt{\mathrm{Var}(\bar{y}_{t+1})\mathrm{Var}(\bar{y}_t)}\,\mathrm{Corr}(\bar{y}_{t+1}, \bar{y}_t), \text{ and}$$

$$\mathrm{Var}(\bar{y}_t) = \mathrm{Var}(\Delta) - \mathrm{Var}(\bar{y}_{t+1}) + 2\sqrt{\mathrm{Var}(\bar{y}_{t+1})\mathrm{Var}(\bar{y}_t)}\,\mathrm{Corr}(\bar{y}_{t+1}, \bar{y}_t),$$

will have low values when $\mathrm{Corr}(\bar{y}_{t+1}, \bar{y}_t)$ is close to zero. Thus if main interest is on the individual population estimates at each time point, then repeated cross-sectional surveys are recommended, otherwise panel designs should be used.

In longitudinal surveys, both $\bar{y}_t$ and $\bar{y}_{t+1}$ estimate population mean for the population defined at time t since same elements are followed for the whole survey period. If the population is dynamic $\bar{y}_{t+1}$ does not estimate the population mean for the population at time $t+1$. Populations elements included in survey at time t are likely to be selected such that the resulting sample is representative for the population at that time which might not necessarily be representative of the population at time $t+1$.

On the other hand, in repeated cross-sectional surveys, each estimate, i.e., $\bar{y}_t$ and $\bar{y}_{t+1}$ estimate the population mean for the population at that particular time. This is because at each time point a fresh sample is selected. This important difference between the two types of surveys and practicality are important determining factors for deciding which type of survey to use. For example, longitudinal surveys are impractical for the pesticide monitoring study since it is not possible to follow samples of commodities over a period of time.

An example of repeated cross-sectional surveys is evident in the sampling plan for Denmark. In addition to commodities required for EU coordinated program, Denmark also includes 25 commodities which, based on analysis of pesticide residues in food items over several years, were found to account for more than 95% of pesticides residues intake through food items. Indeed, if the same residues are measured on these commodities each year, the data collected would be useful in measuring the gross change in, say proportion of samples above MRL, over the years.

1.3. Sample Size Calculation Under Various Designs

Sample size calculation formulas presented in this section are based on estimation of mean of variable of interest. The rationale behind the derivation of the formulas is to first fix the desired margin of error $d = z\sqrt{\mathrm{Vary}(\bar{Y})}$. Often, $\mathrm{Vary}(\bar{Y})$ is a function of sample size (n), hence by fixing all other parameters except n in the expression of d, we can easily obtain the required sample size. Further, z is the quantile for the normal distribution implying that sample sizes obtained from formulas below are based on the assumption that for a considerably large sample, the estimated mean has a normal distribution. The normality assumption which implicitly assumes use of large sample size allows us to drop the small population finite correction in the expressions for variance of mean that were obtained from Kish,

(1965). In cases where the normality assumption is not viable exact expressions can be used in place of the normal approximation, more details on exact sample size calculation, especially for estimating a proportion can be found in Fleiss, (1981).

1.3.1. Simple Random Sampling

Determining the sample size is another crucial part in designing a survey. Sample size can be calculated with the aim of achieving some desired characteristics like, precision, power of hypothesis testing, and type I error. The purpose for which the sample is being collected also plays a big role in determining the size of the sample, the sample size needed for inferential objectives is not necessarily the same as that under estimation objectives. We will therefore present the sample size calculations for each of the objective types.

1.3.1.1. Sample Size Calculation Under Estimation

When estimation of some characteristics of the targeted population is of interest, it is important that the estimate be obtained with the highest precision practically possible. Sample size is thus calculated with the aim of obtaining a desired level of precision. Let $Y_1,\ Y_2, \dots, Y_i \dots, Y_N$ be elements from the targeted population with variance, $\text{var}(Y_i) = S^2,\ let\ \bar{Y}$ be the population mean and $\text{var}(\bar{Y}) = S^2/N$. Likewise, let $y_1, y_2, \dots, y_i, \dots, y_n$, $\bar{y}$, s^2, and $\text{var}(\bar{y}) = s^2/n$ be the corresponding quantities from the sample (s^2 is the element variance). It can be shown that the sample size n can be obtained as:

$$n = \frac{4\, z^2 s^2}{w^2} = \frac{z^2\, s^2}{d^2}, \text{where } s^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

Where w^2 is the desired width of the confidence interval for the estimated mean, d is the margin of error, defined as the error which the researcher is willing to accept in estimating the statistic of interest and z is the normal quantile of $\alpha\ (type\ I\ error)$, the risk that a researcher is willing to accept that the true margin of error exceeds the acceptable margin of error, (Bartlett, et.al, 2001). It is clear that a smaller sample size correspond to large margin of error. While d and z are usually fixed, $\text{var}(\bar{y})$ depends on the sampling design.

Example 1.3.1.1: Assume we would like to draw a sample from the 29 participating EU member states to study a particular binary trait, like whether a food sample has residues above MRL or not, it follows that $\bar{y}$ will be a proportion (p). Let the desired margin of error(d), type I error(α), and the element variance $s^2 = p(1-p)$, be 0.05, 0.05, 0.2 and 0.16, respectively. The margin of error defines our desired level of accuracy, and the value of 0.05 in this case simply means that we desire that the difference between the true proportion (p) and the estimated proportion ($\hat{p}$) should not exceed 0.05. If the actual difference exceeds this value, we commit a type I error. For $\alpha = 0.05$, $z = 1.96$ and $d = 0.05$ the total sample size (all MS together) required will be:

$$n = \frac{1.96^2 \times\ 0.16}{0.05^2} \approx 246$$

1.3.1.2. Sample Size Calculation Under Hypothesis Testing

In this scenario interest is in testing a formulated hypothesis with a desired level of power of testing $(1-\beta)$) at an acceptable level of type I error (α). More details on setting the acceptable level for type

I error can be found in FASFC[12]. In the paradigm of pesticide monitoring program, a null hypothesis (H_0) can be formulated as: proportion of samples above MRL for a particular food is equal or below a pre-specified "safe" proportion against an alternative hypothesis(H_1): proportion exceeds the pre-specified proportion. A hypothesis can also be formulated to compare group means in the population. In general a hypothesis can be formulated as either of the following:

$$H_0: \bar{Y} \leq \bar{Y}_0 \quad or \quad H_0: \bar{Y} \geq \bar{Y}_0 \ or \quad H_0: \bar{Y} - \bar{Y}_0 = \delta$$

$$H_1: \bar{Y} > \bar{Y}_0 \quad H_1: \bar{Y} < \bar{Y}_0 \quad H_1: \bar{Y} - \bar{Y}_0 \neq \delta$$

For the formulation on the far left, the alternative hypothesis (H_1) tests that the population mean ($\bar{Y}$) is greater than the pre-assumed mean ($\bar{Y}_0$), while for the formulation in the middle, H_1 tests that population mean ($\bar{Y}$) is less than the pre-assumed mean ($\bar{Y}_0$). The formulation on the right is a two sided hypothesis, H_1 tests that the population mean different from the pre-assumed mean, it can be rejected in both when the population mean is smaller or greater than the pre-assumed mean, i.e., ($\bar{Y} < \bar{Y}_0$ or $\bar{Y} > \bar{Y}_0$). Rejecting H_0 leads to adoption of H_1. In addition to variance (s^2), and type I error(α), the desired power of testing $(1-\beta)$ is needed for calculating the sample size. Sample size n is given by:

$$n = \frac{s^2[\,\Phi^{-1}(1-\alpha) + \Phi^{-1}(\,1-\beta)]^2}{\delta^2},$$

(Jennison and Turnbull, 2000), where Φ^{-1} is the inverse standard normal cumulative distribution function. Notably, the ratio s^2/δ^2 influences sample size for each combination level of desired power to test the hypothesis and the type I error.

Example 1.3.1.2: Continuing with Example 1.3.1.1, further assume that for hypothesis testing we need the power of testing, $1-\beta = 0.80$ and we would like to detect a difference of $\delta = 0.05$ for a two sided alternative hypothesis. The required sample size will be:

$$n = \frac{0.16(\,1.96 + 0.84)^2}{0.05^2} \approx 502$$

The values for $\Phi^{-1}(1-\alpha)$ and $\Phi^{-1}(1-\beta)$ can be easily obtained through the normal distribution tables or statistical software, like SAS using the "quantile" function.

1.3.2. Cluster Sampling

The sample size calculations presented in SRS assume independence among the sampled elements. Sometimes it is desirable to sample group of elements rather than individual elements such that elements in the same cluster are usually more similar than elements between clusters. In the pesticide data monitoring program this can be the case for samples that were obtained from the same food item: it is more likely that residues found in samples obtained from apples will be more similar than those found in samples obtained from tomatoes. Kish (1965), notes that the main advantage of this design is its cost effectiveness. Otherwise it is associated with decrease in precision hence it should be used when the gain in costs is considerably larger than the loss in precision.

[12] FASFC: http://www.favv-afsca.fgov.be/publicationsthematiques/food-safety.asp

1.3.2.1. Sample Size Calculation Under Estimation

Assume we have a targeted population for which sampling in groups (clusters) would be cost effective enough to subdue the loss in precision. Define $Y_{it,}$ as the i^{th} element in the t^{th} cluster and denote $\bar{Y}_c$ as the population mean. It can be shown (Kish, 1965) that $\text{var}(\bar{Y}_c) = S_c^2/A$, where A is the total number of clusters in the population. Likewise, let, y_{it}, $\bar{y}_c$, and $\text{var}(\bar{y}_c) = s_c^2/a$ be the corresponding quantities for the sample. Selection of elements proceeds by selecting a out of A clusters using SRS and including all elements from the selected clusters into the survey. It follows that the total sample size, $n = aB$, where B is the total number of elements in each cluster. The number of clusters to be selected can be obtained as:

$$a = \frac{4z^2 s_c^2}{w^2} = \frac{z^2\, s_c^2}{d^2},$$

Where

$$s_c^2 = \sum_{t=1}^{a} \frac{(\bar{y}_t - \bar{y}_c)^2}{a-1}$$

Where $\bar{y}_t$ is the estimated mean in t^{th} cluster, the rest of the parameters are as defined in SRS scenario. Note that s_c^2 computes the variability of cluster means from the overall mean or alternatively the variance between clusters. It can be likened to the SRS scenario by considering the clusters as elements (recall that we are randomly sampling the clusters). However, s_c^2 cannot be estimated before the survey since it requires the value of a which we want to compute, it is either obtained from pilot studies or based on expert opinion. When all elements from the selected clusters are included in the survey, total number elements sampled is obtained as $n = aB$.

Example 1.3.2.1: Let $s_c^2 = 0.1$, d = 0.05, z = 1.96, and B = 10. The number of clusters needed to achieve these desired characteristics and the total sample size are,

$$a = \frac{1.96^2 \times 0.1}{0.05^2} \approx 154$$

Alternatively, the sample size can be obtained by adjusting the SRS variance with the design effect. Design effect was defined as the ratio of variances under SRS design and design of interest (i.e., cluster sampling), and for the mean estimate, this implies:

$$D = [1 + \rho(B-1)] = \frac{s_c^2/a}{s^2/n}$$

$$\Rightarrow \frac{s_c^2}{a} = [1 + \rho(B-1)] \times \frac{s^2}{n}.$$

Margin of error for the mean estimate under cluster sampling is given by:

$$d = z\sqrt{\frac{s_c^2}{a}} \Rightarrow d^2 = z^2 \frac{s_c^2}{a}$$

$$\Longrightarrow d^2 = z^2[1+\rho(B-1)] \times \frac{s^2}{n}$$

$$\Longrightarrow n = [1+\rho(B-1)] \times \frac{z^2 s^2}{d^2}.$$

Thus the total sample size under cluster sampling is simply the product of the design effect and sample size. Notice that under cluster sampling we have two options: we can calculate either the number of clusters (a) and fix the number of elements selected from each cluster or calculate the total sample size (n) and fix the number of clusters from which to select all the elements. The two quantities (number of clusters and total sample size), cannot be estimated simultaneously. In the first part of the Example **1.3.2.1** we estimated a, the number of clusters and the second part estimates n the total sample size. For the latter, prior knowledge of intra-class correlation (ρ), variance of the mean under SRS and average cluster size are required. These quantities can be obtained from a pilot study.

Example 1.3.2.2: For $B = 10$, $s^2 = 0.16$ and some given values of ρ, values of total sample size required (n) are:

$\boldsymbol{\rho}$	$\boldsymbol{D}$	$\boldsymbol{n}$
0	1	246
0.02	1.18	290
0.04	1.36	334
0.06	1.54	379
0.08	1.72	423
0.1	1.9	467
0.2	2.8	688
0.4	4.6	1131
0.6	6.4	1574
0.8	8.2	2016
1	10	2459

Even for a weak intra-class correlation of 0.1, the sample size almost doubles and for the strong intra-class correlation of 1 the sample size required is B times the samples size under SRS ($\rho = 0$).

1.3.2.2. Sample Size Calculation Hypothesis testing

Similarly, under hypothesis testing the number of clusters is obtained as:

$$a = \frac{s_c^2[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)]^2}{\delta^2}.$$

The total sample size will be similarly obtained as $n = aB$.

Example 1.3.2.3: Lets$_c^2 = 0.1$,$\delta = 0.05$,$\alpha = 0.05$, $\beta = 0.2$ and $B = 10$. The number of clusters needed to test a two sided hypothesis to achieve these desired characteristics is,

$$a = \frac{0.1(\,1.96 + 0.84)^2}{0.05^2} \approx 314.$$

Alternatively, values of sample size required for some given values of ρ and $s^2 = 0.16$ are:

ρ	D	n
0	1	502
0.02	1.18	592
0.04	1.36	682
0.06	1.54	773
0.08	1.72	863
0.1	1.9	953
0.2	2.8	1405
0.4	4.6	2308
0.6	6.4	3211
0.8	8.2	4114
1	10	5018

The relationship between ρ and n is similar to that observed in Example 1.3.2.2.

The examples and formulae given above, assume cluster of equal sizes, and we acknowledge that this is rarely the case in reality. However, when dealing with unequal cluster sizes, quantities like, sample size and mean estimate are no longer fixed quantities. The former becomes a random variable and the latter, a ratio estimator. Sample size calculations in such setting become complicated as it can be observed from variance expressions for mean for unequal cluster sizes in Kish, 1965 pp. 190--193. However, the presented methods can serve as good approximations especially in cases where the cluster size and the mean are not correlated. For example, the cluster size B in the above formulas can be replaced by a reasonable average cluster size like $B_{avg} = \frac{N}{A}$.

1.3.3. Stratification

Given a specified allocation scheme, and the desired precision (hereunder represented by the margin of error), the required overall sample size, as well as the allocation to strata, can be determined.

Suppose a population of size N is stratified into H strata, each of size N_h , $h = 1, \ldots, H$. The "weights" $W_h = \frac{N_h}{N}$ denote the population proportion of the strata. Simple random samples are drawn separately within strata.

For the estimation of the population mean $\bar{Y}$, the stratified estimator is given as

$$\bar{y}_{st} = W_1\bar{y}_1 + W_2\bar{y}_2 + \cdots + W_H\bar{y}_H = \textstyle\sum_{h=1}^{H} W_h\,\bar{y}_h\ ,$$

with $\bar{y}_h$ the stratum sample means. The variance of this estimator, ignoring the finite population correction factor, can then be expressed as

$$Var(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \frac{S_h^2}{n_h},$$

with n_h the stratum-specific sample size; $\sum_h n_h = n$, where n is the "total" sample size, and S_h^2 the population variance in stratum h. The finite population correction factor can be incorporated as in Barnett, 1991, p110; Kalton, 1983, p20; and Groves *et. al*, 2004, p112.

Estimators for the population variances in the strata are given as

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$$

A proportion is just a special case of a mean, and, therefore, estimation of a population proportion P follows similar logic:

$p_{st} = W_1 p_1 + W_2 p_2 + \cdots + W_H p_H = \sum_{h=1}^{H} W_h\, p_h$,

with p_h the stratum sample proportions. An estimator for the variance of p_h, ignoring the finite population correction factor, can then be expressed as

$$var(p_{st}) = \sum_{h=1}^{H} W_h^2 \frac{p_h(1-p_h)}{n_h - 1}.$$

This is usually approximated to

$$var(p_{st}) = \sum_{h=1}^{H} W_h^2 \frac{p_h(1-p_h)}{n_h}.$$

To estimate the mean with a margin of error of size d, the sample size required is derived as follows. A sample size n is required, such that

$$z\sqrt{Var(\bar{y}_{st})} = d.$$

Substituting for $Var(\bar{y}_{st})$, we have that

$$z\sqrt{\sum_{h=1}^{H} W_h^2 \frac{S_h^2}{n_h}} = d,$$

$$z^2 \sum_{h=1}^{H} W_h^2 \frac{S_h^2}{n_h} = d^2.$$

Now, $w_h = \frac{n_h}{n}$, the sample proportion of the stratum. Note that this is different from W_h, the population proportion of the stratum. From the sample proportion of the stratum, $n_h = nw_h$. Substituting for n_h above, we get

$$z^2 \sum_{h=1}^{H} W_h^2 \frac{S_h^2}{n w_h} = d^2.$$

We solve for n in this equation, obtaining

$$n = \frac{z^2}{d^2} \sum_h \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{w_h} .$$

1. Proportional Allocation

Under proportional allocation, the proportions of the sample in the stratum, w_h, are set equal to the proportions of the population in the stratum, W_h; i.e. $w_h = W_h = \frac{N_h}{N}$. The formula to calculate the sample size is then

$$n = \frac{z^2}{d^2} \sum_h \frac{N_h}{N} S_h^2.$$

2. Neyman Allocation

Neyman allocation, (Groves et. al, 2004, p117; Som, 1996, p211; Kalton, 1983, p24; Barnett, 1991, p120), is the allocation that minimizes sampling variance, assuming equal costs across strata. It is sometimes referred to as the optimum allocation (Som, 1996, p211).

Neyman allocation requires the following:

$$n_h = \frac{W_h S_h}{\sum W_h S_h} n$$

This implies that for a margin of error of size d, the following sample size is required:

$$n = \frac{z^2}{d^2} \left(\sum W_h S_h\right)^2.$$

Example 1.3.3.1

1. Proportional Allocation

Suppose a certain proportion of interest is to be estimated from a population of size $N = 1000000$. The population is grouped into 10 strata of the sizes shown in Table 1. Further, sampling will be organized within the strata themselves, where simple random samples will be drawn separately in each stratum. An overall population estimate is required, with a margin of error of 5%.

To calculate the sample size required, and the allocation to the various strata, the weights W_h, and the variability within the strata S_h, will be required. In Table 1, the population proportions W_h are provided. For the variability within strata, conservative estimates of 0.25 are uniformly assumed across strata. This assumption is for convenience of illustration; strata variances may vary. When dealing with proportions, 0.25 is the maximum element variance possible.

Applying the corresponding formula above, the overall sample size is $n = 384$. Sample allocation to the various countries is provided in Table 1. Large strata receive large allocation; the reverse holds for small strata.

Table 1 Proportional Allocation Example (Uniform Strata Variances of 0.25)

Stratum	Population	Proportion of Population	Sample Size Allocation
1	20000	0.0200	8
2	30000	0.0300	12
3	40000	0.0400	15
4	80000	0.0800	31
5	100000	0.1000	38
6	110000	0.1100	42
7	120000	0.1200	46
8	130000	0.1300	50
9	170000	0.1700	65
10	200000	0.2000	77

Table 2 Sample Sizes for Smaller Uniform Strata Variances

Variance	Sample Size Required
0.1250	192
0.0500	79
0.0250	38
0.0100	15
0.0020	3

As mentioned, variances of 0.25 in the case of proportions provide the most conservative scenario. Smaller uniform variances would require smaller sample sizes, as shown in Table 2 above.

2. Neyman Allocation

Assume the population mentioned above is now stratified into 10 equal sized strata (i.e. each of size 100000). Equal sizes are now assumed, for convenience of illustration. Suppose the variances in stratum 1 to 10 range from 0.0250 to 0.2500, with increments of 0.0250 (stratum 1=0.0250, stratum 2=0.0500,…, stratum 10=0.2500), and a similar margin of error as above is required. Applying the corresponding formula above, the "total" sample size $N = 194$. The allocation to the various strata would be as shown in Table 3. Those strata with more variability receive larger allocations.

Table 3 Neyman Allocation Example (Equal-Sized Strata Assumed, Different Variances)

Stratum	Variance	Sample Size Allocation
1	0.0250	9
2	0.0500	12
3	0.0750	15
4	0.1000	17
5	0.1250	19
6	0.1500	21
7	0.1750	23
8	0.2000	24
9	0.2250	26
10	0.2500	27

1.3.4. Multi-stage Sampling

For multi-stage sampling a sample size calculations are not entirely theoretical, some expert input is required to fix some quantities. Multi-stage sampling suits a lot of practical situations due to its flexibility, for example, it is used in a lot of cross-border surveys like the European Social Survey13.

Sample Size Calculation Under Estimation

Assume a population that is divided into H strata, and within each strata there are A_h clusters $h = 1, \dots, H$, each of size B. The sample collection proceeds as follows: within the h^{th} stratum, a_h clusters are selected using SRS, and within each of the selected clusters, all the B elements are included in the sample. Denote as y_{ith} an i^{th} element sampled from t^{th} cluster, $t = 1, \dots, a_h$, within h^{th} stratum. Additionally, let $\bar{y}_{th}$ be the estimated mean in cluster t within h^{th} stratum, and $\bar{y}_h$ be the stratum mean. Further, let $\bar{y}_{ms}$ be the overall mean estimate obtained under this multistage sampling design and $\text{var}(\bar{y}_{ms}) = s^2_{ms}$ its corresponding variance. Sample size in h^{th} stratum is given by $n_h = a_h B$, and the total sample size , $n = \sum_{h=1}^{H} a_h B$. Variance is given by,

$$s^2_{ms} = \sum_{h=1}^{H} W_h^2 \sum_{t=1}^{a_h} \frac{1}{a_h}\left[\frac{(\bar{y}_{th} - \bar{y}_h)^2}{a_h - 1}\right] = \sum_{h=1}^{H} W_h^2 \frac{s^2_{ch}}{a_h},$$

Where W_h stratification weight and $\frac{s^2_{ch}}{a_h}$ is the variance in the h^{th} stratum (note that this is simply the variance under cluster sampling design). Recall that margin of error (d) specifies half of the desired confidence interval width, i.e.,

$$d = z \times \sqrt{s^2_{ms.}}$$

For proportionate allocation,

$$s^2_{ms} = \sum_{h=1}^{H} W_h^2 \frac{s^2_{ch}}{a_h} = \sum_{h=1}^{H} \left[\frac{A_h B}{N}\right]^2 \frac{s^2_{ch}}{a_h} = \sum_{h=1}^{H} \left[\frac{A_h B}{N}\right]^2 \frac{s^2_{ch}}{\frac{a_h B}{n} \times \frac{n}{B}}.$$

Since $\left[\frac{A_h B}{N}\right] \approx \frac{a_h B}{n}$ it follows that,

$$s^2_{ms} = \frac{B}{n} \sum_{h=1}^{H} \left[\frac{A_h B}{N}\right] s^2_{ch}\,. \implies d = z \times \left(\frac{B}{n}\right)^{0.5} \sqrt{\sum_{h=1}^{H} \left[\frac{A_h B}{N}\right] s^2_{ch}}\,,$$

Hence,

$$n = \frac{z^2 B}{d^2} \sum_{h=1}^{H} \left[\frac{A_h B}{N}\right] s^2_{ch}\,.$$

[13] http://www.europeansocialsurvey.org/methodology/sampling.html

To obtain s_{ch}^2, number of clusters, a_h is required and this can be fixed by researchers. Obviously the formulas will become more complicated if either clusters are of unequal sizes or the number of elements sampled from the selected clusters is different for different clusters. A workable solution is to use average values, e.g., replace B with the average cluster size. Using these values will lead to approximation whose performance will later be studied through simulations.

It can also be shown that by using the design effect concept, total sample size can be obtained as

$$n = \frac{z^2}{d^2} \sum_{h=1}^{H} D_h \left[\frac{A_h B}{N}\right] s_h^2 ,$$

where s_h^2 and D_h are the element variance and design effect for the h^{th} stratum, respectively.

A good overview for calculating sample size for multistage designs for a binary characteristics is given in EFSA, 2013[14] where it is also noted that to obtain the final sample size, sample size for some levels has to be fixed.

Example 1.3.4.1: Consider a population stipulated in the Example 1.3.2.1 under cluster sampling, and additionally the elements are divided into H strata. We would like to select a_h clusters from each stratum from which $B = 10$ elements will be selected. Our interest is to have a total sample size that will achieve a 0.05 margin of error (d) at 0.05 type I error (α). Let $s_{ch}^2 = 0.1$ for each stratum, it follows that, $s_{ms}^2 = 0.1 \sum_{h=1}^{k} W_h = 0.1$.

Total sample size can be obtained as:

$$n = \frac{1.96^2 \times 0.1 \times 10}{0.05^2} \approx 1537.$$

The sample size for stratum h with $W = 0.02$ will be obtained as

$$n_h = 0.02 \times 1537 \approx 31.$$

Table 4 shows the allocation to other strata;

[14]European Food Safety Authority, 2013. Sample Size Considerations for Hierarchical Population. EFSA Journal 2013;11(7):3292, 47 pp. doi:10.2903/j.efsa.2013.3292

Table 4 Proportional Allocation Example Under Multi-stage Sampling (Uniform Strata Variances of 0.25), for Estimation Objective .

Stratum	Population	Proportion of Population	Sample Size Allocation
1	20000	0.0200	31
2	30000	0.0300	46
3	40000	0.0400	61
4	80000	0.0800	123
5	100000	0.1000	154
6	110000	0.1100	169
7	120000	0.1200	184
8	130000	0.1300	200
9	170000	0.1700	261
10	200000	0.2000	307

1.3.4.1. Sample Size Calculation Hypothesis testing

For hypothesis testing we have

$$n = \frac{B \times s_{ms}^2[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)]^2}{\delta^2}$$

Example 1.3.4.2: If interest is in testing a two sided alternative hypothesis with a power of $1-\beta = 0.8$ to detect a difference of $\delta = 0.05$ with type I error $\alpha = 0.05$ for $B = 10$. The total sample size is obtained as:

$$n = \frac{10 \times 0.1(1.96 + 0.84)^2}{0.05^2} \approx 3136.$$

Table 5 Proportional Allocation Example Under Multi-stage Sampling (Uniform Strata Variances of 0.25), for Hypothesis Testing Objective.

Stratum	Population	Proportion of Population	Sample Size Allocation
1	20000	0.0200	63
2	30000	0.0300	94
3	40000	0.0400	125
4	80000	0.0800	251
5	100000	0.1000	314
6	110000	0.1100	345
7	120000	0.1200	376
8	130000	0.1300	408
9	170000	0.1700	533
10	200000	0.2000	627

1.3.5. Designs for Measuring Change Over Time

In addition to precision and representativeness considerations, the power to detect expected change is also a crucial factor in surveys that are meant to measure change. Reconsider Example 1.3.4.1. Say we need power of $1-\beta = 0.80$ to detect a change $\delta = 0.05$ (in proportions) for a two sided alternative hypothesis at $\alpha = 0.05$ type I error level. Further, let $\text{Var}(\bar{y}_t) = \text{Var}(\bar{y}_{t+1}) = 0.25$.

Let the survey elements at the two time intervals be the same such that $\text{Corr}(\bar{y}_{t+1}, \bar{y}_t) = 0.95$, if follows that

$$\text{Var}(\Delta) = 0.25 + 0.25 - 2 \times 0.25 \times 0.95 = 0.025,$$

The required sample size is

$$n = \frac{0.025(\ 1.96 + 0.84)^2}{0.05^2} \approx 78$$

Let the survey elements at the two time intervals overlap such that $\text{Corr}(\bar{y}_{t+1}, \bar{y}_t) = 0.5$, if follows that

$$\text{Var}(\Delta) = 0.25 + 0.25 - 2 \times 0.25 \times 0.95 = 0.25,$$

The required sample size is

$$n = \frac{0.25(\ 1.96 + 0.84)^2}{0.05^2} \approx 784$$

Let the survey elements at the two time intervals are independent such that $\text{Corr}(\bar{y}_{t+1}, \bar{y}_t) = 0.01$, if follows that

$$\text{Var}(\Delta) = 0.25 + 0.25 - 2 \times 0.25 \times 0.01 = 0.495,$$

The required sample size is

$$n = \frac{0.495(\ 1.96 + 0.84)^2}{0.05^2} \approx 1552$$

Thus a repeated cross-sectional survey will require about 20 times larger sample size than a longitudinal survey to detect a change of 0.05 with 80% power.

1.3.6. Complex Survey Analysis

More often than not practical considerations do not allow for implementation of exclusively one kind of sampling design. Issues like available budget and physical location of the sampling units may dictate that a combination of all sampling designs discussed in Section 0 be combined in designing one survey. Consider again the pesticide monitoring survey that combines data on food residues from different member states. As stated before, the member states form natural strata hence stratification would be beneficial. Further, it may not be possible to enumerate all available food items in all member states. On the other hand, it is relatively easy to enumerate the food categories, and these may be regarded as clusters or strata depending on the objective of the survey. Importantly, when selecting food items within the food categories, it may be helpful to assign higher probability of selection to food items that are highly consumed, making sampling probability proportion to size part of the design. Thus the design would combine stratification, cluster, and probability proportion to size. While such a combination makes the survey practical, it complicates estimation of parameters of interest especially variance parameters.

Complex surveys may also arise in the context of cluster sampling if cluster sizes and number of elements to be sampled from the selected clusters differ. Kish, (1965), clearly outlines complications arising from such settings which occur more often in reality than the assumed equal cluster sizes in Section 1.3.2.

The need to combine designs is inevitable in many situations, and so is the need for estimation methods that can accommodate such complicated designs. Methods for analyzing complex surveys fall into two groups: design and model based. Literature has shown that the two methods usually produce similar results (Lehtonen, et. al, 2004; Ghosh and Pahwa, 2006). Model based estimation is preferred for: (i) its flexibility in accommodating several design aspects, and (ii) availability of software that aids their application. For sample size calculations, models accommodating various design aspects can be used to analyze data from pilot studies to get appropriate variance estimate which can be plugged in formulas given in Section 1.3.4.

RESULTS: PESTICIDE MONITORING DATA

1.4. Review of Survey Practices for Pesticide Monitoring Data

1.4.1. Definition of Target Population

A clear definition of targeted population is important to determine the extent to which results from the survey can be generalized. It is clear that the survey seeks to assess MRL compliance among food items available to European consumers.

1.4.2. Sampling Frame

While definition of the sampling frame is critical to the design of the survey and especially to selecting a representative sample, it is not always possible to have an unambiguous sampling frame. For example, in the pesticide monitoring study, food items available to EU consumers include food items that are imported from other non-EU countries hence a sampling frame can be defined to include such food items too. In this case, food items available to consumers through markets, retailers and other selling point would define the sampling frame. On the other hand, the sampling frame can be extended to include food commodities that are not yet available to consumers, i.e., all food items produced or imported to the EU. In such a case, food items that are still on the farm would be part of the sampling frame. Obviously the extended definition is likely to be more representative than the other one, however, practical considerations may limit its practicality.

Different member states have different definitions of sampling frame, for example, Latvia's sampling frame includes domestic fresh plant products from conventional farms and sampling points includes farm gets. This may not be practical for all member states hence the sampling frame for EU coordinated program may be defined by bringing together definitions sampling frames from all MS and taking a subset of definitions that are practical for all member states.

In general, expert input would be required to decide on a more representative and practical sampling frame. Trade-offs between representativeness and practicality are almost inevitable. It is vital that information about compromises that have been made is available to the researcher so that it can be incorporated in analysing and/or interpreting the results from the survey.

1.4.3. Objective of the Study

A clearly stated objective of the survey is the main determining factor of other decisions made in conducting a survey, like sample size calculation and the type of sampling design. The study is conducted to assess MRL compliance rate for food items available to EU consumers. Laboratory protocols are used to determine residue levels in food item samples. A sample is non-compliant if the residue level is above MRL for the particular residue. To assess MRL compliance for a food item, proportion of non-compliant samples is determined and an "acceptable" proportion can be determined to help classify food item as compliant or non-compliant, suggesting a hypothesis testing objective. Ideally, the sample should serve also to estimation of European exposure to pesticide residues, suggesting the need for estimation of the residues level in the samples taken.

1.4.4. Sampling Design

As noted earlier, multistage sampling would be preferred in this setting due to its flexibility. Specifically, it allows each member state to develop a sampling design relevant to its needs while at the same time allowing for a valid overall analysis of results from all MS. Many members states, like, Belgium, Netherlands, and Austria use both objective and targeted sampling of food commodities. The targeted sampling is directed towards food items that are relevant to the specific MS in terms of consumption and non-compliance rate from previous years. These designs can easily be brought together in the framework of a multi-stage sampling design.

Analysis of the data from the 2010 EFSA report on pesticide monitoring reflects a SRS design. On the other hand, the multi-country nature and sampling protocols in the study support a multistage design. The two designs can be presented as follows:

1.4.4.1. Design Reflected by Analysis

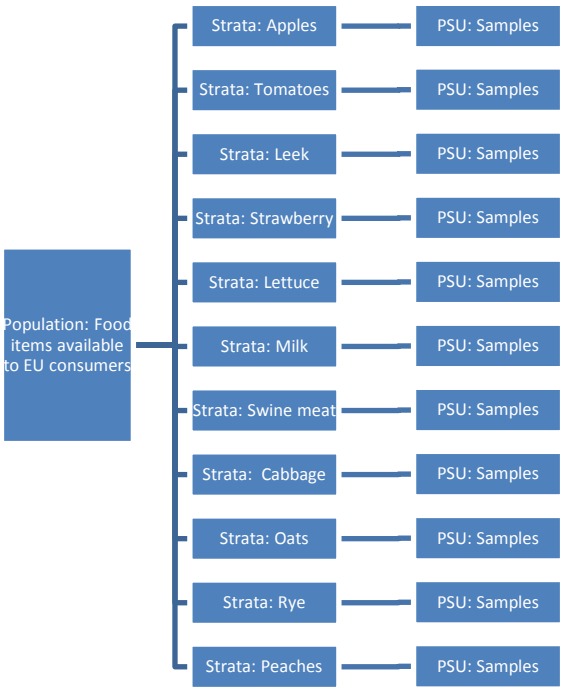

Figure 1 Simple Random Sampling Design as Reflected by the Analysis.

Note that the sampling frame is reduced to only the food items under study and thus the results cannot be generalized to other food items. Importantly, it ignores the geographical spread of the data by not taking into account the country from which the food items are sampled from. This design is obviously not feasible since EFSA relies on MS to collect the food samples from the selected food products.

1.4.4.2. Design Reflected by Nature of the study

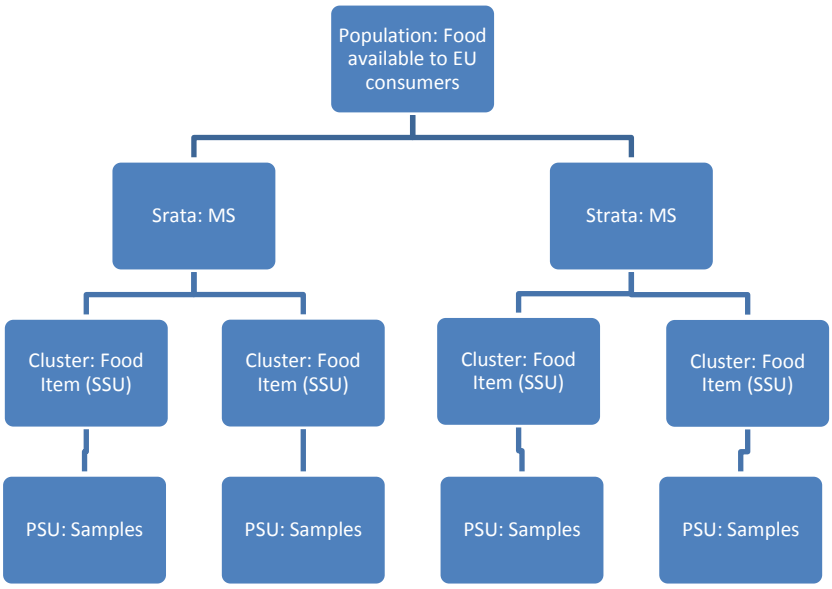

Figure 2 Multi-Stage Sampling Design.

This design rightly recognizes the geographical structure of the data by considering MS as strata. It further allows for generalization of results to all food items in the sampling frame since food items are randomly selected, not fixed as in the previous design. In this design, food items are secondary sampling units (SSU) and food samples are primary sampling units (PSU).

1.4.5. Sample Size Calculation

The sample size used in the survey was calculated based on a different objective from what seems to be the objective of the survey. According to Article 3 of Regulation (EC) No 901/2009, the total number of samples to be analyzed was derived on the basis of a binomial probability distribution, which estimated that the examination of 642 samples allows the detection of a sample containing pesticide residues above the limit of determination (LOD), with a certainty of more than 99%, provided that no less than 1% of products of plant origin contain residues above that limit. This sample size calculation model is stipulated in the Codex Alimentarius[15], (1999) and was meant for collecting primary samples from a bulk sample.

According to the assessment carried out of the EU monitoring program and reports presented based on this program, the objective of the survey could be rephrased as the estimation of MRL compliance rates in food of plant and animal origin or consumer exposure, implying that sample size calculations should be revised. Using sample size calculation formulas given in Section 1.3, we illustrate possible

[15] Codex Alimentarius, recommended methods of sampling for the determination of pesticide residues for compliance with mrls cac/gl 33-1999.

sample sizes that may be required under different designs based on the objective: Estimation of proportion of samples that have pesticide residues above MRL.

1.4.5.1. Simple Random Sampling

Margin of Error and Sample Size Calculation

Firstly we illustrate the sample size that would be required to obtain a pre-specified margin of error for different values of population variability s^2.

The simple inverse relationship between the margin of error and sample size is clear from Figure 3 where higher level of precision corresponds to larger samples size. For example, for the variance of 0.05 a sample size of 1920 is required to achieve 0.01 level of margin of error compared to 76 required to ensure a margin of error of 0.05. On the other hand, a bigger sample size is required to achieve the same margin of error for larger values of variance than smaller values. A sample size of 384 is required to attain a 0.01 margin of error for $s^2 = 0.01$ compared to more than 2882 needed for $s^2 = 0.075$. A margin of error of 0.01 means we are willing to be precise within the range of 1%, i.e., we want the maximum difference between the true mean $\bar{Y}$ and sample mean $\bar{y}$ to be 0.01, i.e., $\bar{Y}_i = \bar{y}_i \pm 0.01$.

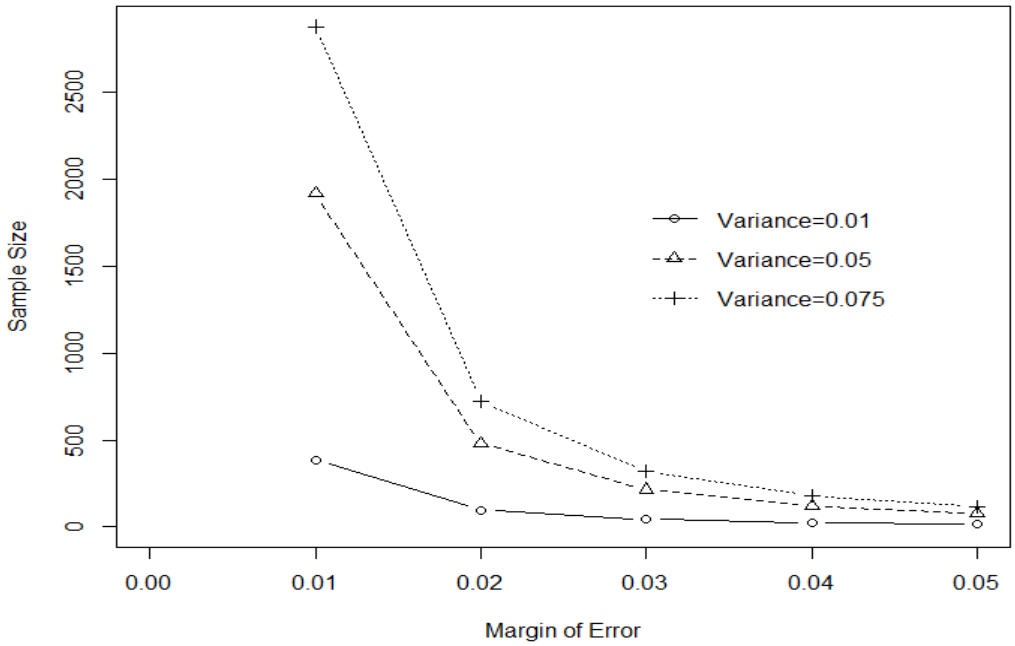

Figure 3: A Plot Showing the Inverse Relationship Between Sample Size and Margin of Error for Estimating a Proportion.

We further explore the margin of error achieved by the samples reported in the 2010 pesticide monitoring programme. We assume that the survey population was designed with the following parameters: type I error $\alpha = 0.05$, element variances $s^2 = (0.002, 0.05)$. The aim is to investigate the margin of error that is achieved by sample sizes for the different food items. The margins of error for a particular food item is given by

$$d = \sqrt{\frac{z^2 s^2}{n_i}},$$

where n_i is the sample size for i^{th} food item and Table 6 presents the results.

Table 6 Margin of Error for the Food Items in 2010 Pesticide Monitoring Study.

Food Item	Number of Samples Reported	Margin of error	
		Var=0.002	Var=0.05
Apples	2057	0.002	0.010
Head cabbage	999	0.003	0.014
Leek	961	0.003	0.014
Lettuce	1568	0.002	0.011
Milk and milk products	654	0.003	0.017
Oats	246	0.006	0.028
Peaches	1200	0.003	0.013
Pears	388	0.004	0.022
Rye	406	0.004	0.022
Strawberries	1272	0.002	0.012
Swine meat	623	0.004	0.018
Tomatoes	1794	0.002	0.010

Some food items would be estimated with higher precision than others. For variance of 0.002, the margin of error is small for all food items but notable differences can be seen at the possible variance of 0.05. Oats, pears and rye have the highest values of margin of error while apples, tomatoes, lettuce, peaches and strawberries have the lowest values. As expected the latter group also has the highest sample sizes and the former has the lowest sample sizes.

Power of hypothesis testing ($1-\beta$) and sample size:

In this scenario, we explore the relationship between power for testing a two sided hypothesis, type I error and sample size. Sample size is obtained for different combinations of power, $(1-\beta) = (0.7, 0.8\ 0.9)$ and the difference to be detected, $\delta = (0.01,\ 0.05\ ,0.075)$, type I error, $\alpha = 0.05$. The values of element variance given by $s^2 = \delta(1-\delta)$ were motivated by the observed range of proportion of samples above MRL for the 2010 pesticide monitoring study for the various food items. Using the terminology in Section 1.3.1.2, the hypothesis can be formulated as:

$$H_0: \bar{Y} - \bar{Y}_0 = \delta \quad \text{or} \quad H_0: \bar{Y} - \bar{Y}_0 \leq \delta \quad \text{or} \quad \bar{Y} - \bar{Y}_0 \geq \delta$$

$$H_1: \bar{Y} - \bar{Y}_0 \neq \delta \quad \text{or} \quad H_1: \bar{Y} - \bar{Y}_0 > \delta \quad \text{or} \quad H_1: \bar{Y} - \bar{Y}_0 < \delta$$

Figure 4 indicates that for the given value of α the sample size increases with increase in power, this trend is evident for both two-sided and one-sided hypotheses. In general two sided hypothesis requires more samples than one-sided hypothesis (keeping all other parameters the same).

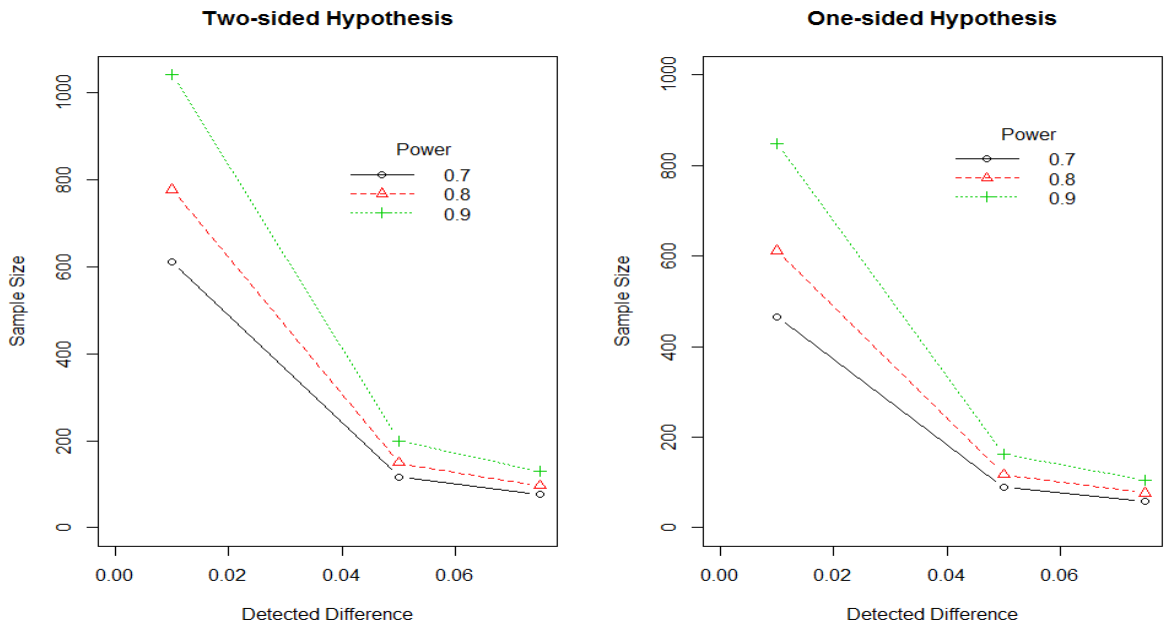

Figure 4: The Relationship Between Power for Hypothesis Testing (Power), and Sample Size for Some Values of the Difference to be Detected (**δ**).

Similar to the margin of error section, we gauge the power that could be attained by the samples size for the various food items in the 2010 pesticide monitoring study. The values of α and δ are as given above, and the power was obtained as

$$(1-\beta) = \Phi\left(\sqrt{\frac{n_i\delta^2}{\delta(1-\delta)}} - \Phi^{-1}(1-\alpha/2)\right).$$

Figure 5, Figure 6 and Figure **7** suggest that if hypothesis testing can be inferred to be the main goal of the survey for the pesticide monitoring programme, for some food items the test would have a power would range between 20% and 80% depending on the specified detectable difference. For example, the estimated proportions for MRL compliance for peaches, apples, pears, tomatoes, leek, lie between 0.01 and 0.018, and a useful detectable difference would be 0.005. The corresponding power at $\alpha = 0.05$ is above 80% for apples, lettuce and tomatoes. The rest would have power below 80% for the considered detectable difference values. In general, large values of detectable difference imply high power of testing.

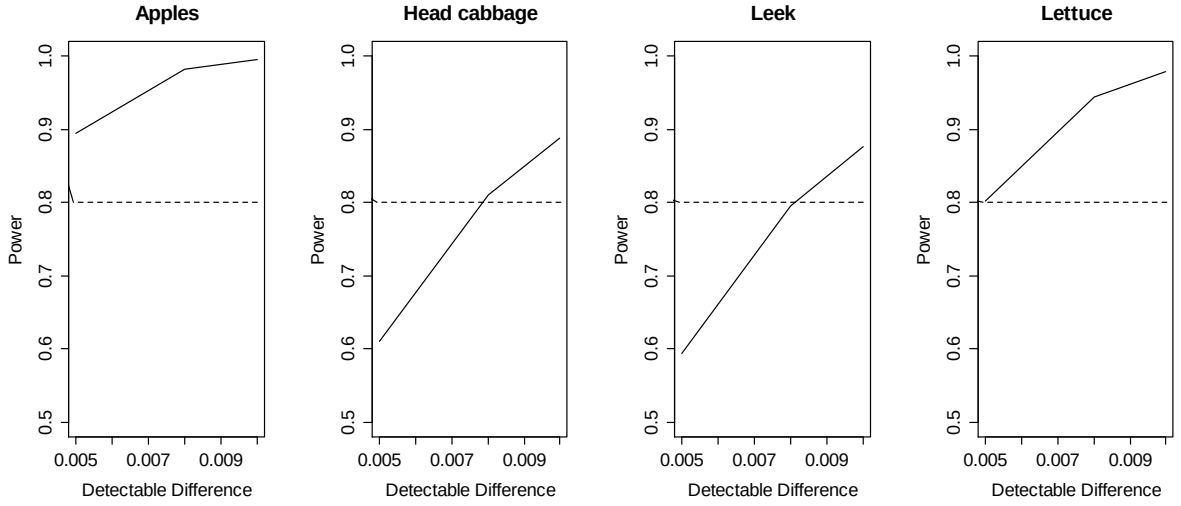

Figure 5: Power of Hypothesis Testing (Power) at $\boldsymbol{\alpha = 0.05}$ and Detectable Difference ($\boldsymbol{\delta}$) for the Various Food Items in the 2010 Pesticide Monitoring Study.

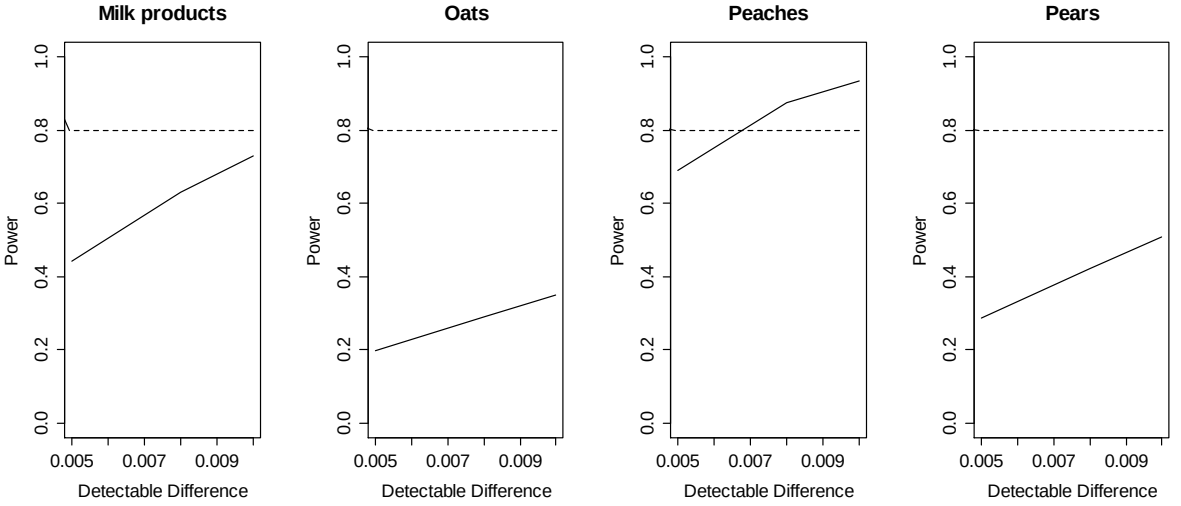

Figure 6: Power of Hypothesis Testing (Power) at $\boldsymbol{\alpha = 0.05}$ and Detectable Difference ($\boldsymbol{\delta}$) for the Various Food Items in the 2010 Pesticide Monitoring Study.

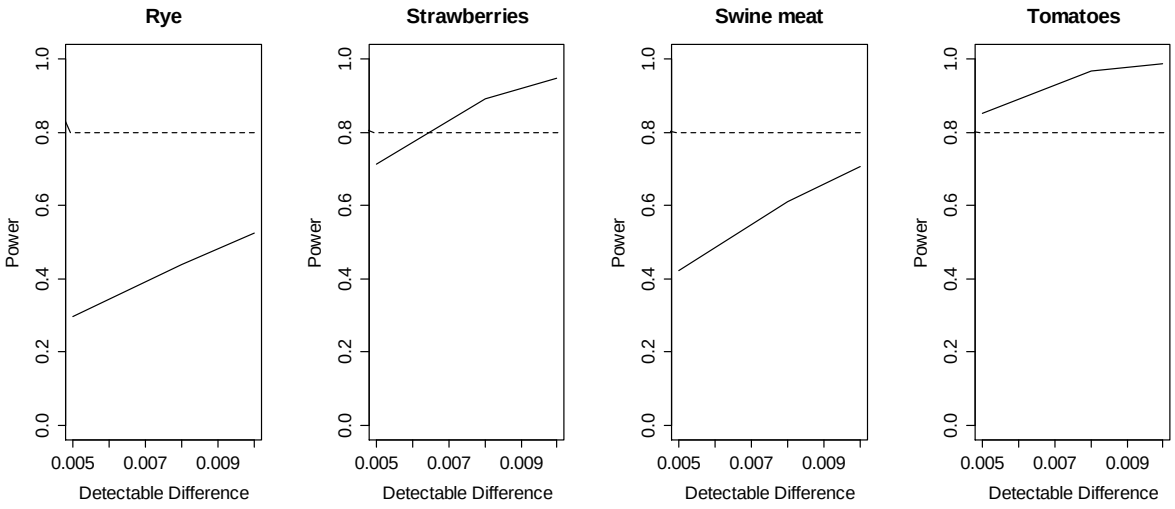

Figure 7: Power of Hypothesis Testing (Power) at $\boldsymbol{\alpha = 0.05}$ and Detectable Difference ($\boldsymbol{\delta}$) for the Various Food Items in the 2010 Pesticide Monitoring Study.

1.4.5.2. Cluster sampling

Impact of clustering on the margin of error:

This setting studies the impact of clustering on the margin of error for the estimated proportion. It is assumed that one obtains a sample size under SRS yet the samples collected are made up of clusters, such that elements within the same cluster are correlated. For given values of clusters size, $n_a = (5, 50, 95)$ and correlation, $\rho = (0.1, 0.5, 0.9)$, the effective sample size (ess) can be obtained as

$$ess = \sum_{a=1}^{N_e} \frac{n_a}{1 + \rho(n_a - 1)},$$

where N_e is the number of clusters, ρ is the correlation within a cluster, n_a is the size of a^{th} cluster and $N_e = \frac{N}{n_a}$, where N is the total sample size. Note that the denominator is the design effect.

The values of n_a and ρ are chosen to represent, small, moderate and high values of cluster size and strength of correlation, respectively. The margin of error is then obtained as

$$d = \sqrt{\frac{4}{ess}},$$

Which assumes the maximum variance of 0.25

Figure 8 illustrates the change in the margin of error between two samples of the same size, one with independent observations (sample=1) and the other with clustered observations (sample=2). The difference is more prominent at the combination of strongest correlation (0.9) and largest cluster size

(95) and the difference is almost non-existent for the combination of weak correlation (0.1) and smallest cluster size (5).

Impact of clustering on the power for hypothesis testing:

Similar to the immediate section above, this simulation study explores the impact of clustering on the power for hypothesis testing. With the values of n_a, ρ and ess the same as those in the section "Impact of clustering on the margin of error", power of hypothesis testing is obtained as,

$$(1-\beta) = \Phi\left(\sqrt{\frac{ess\ \delta^2}{\delta(1-\delta)}} - \Phi^{-1}(1-\alpha)\right).$$

Figure 9 shows the impact of clustering on the power for testing. The sample size used in obtaining the effective sample size was chosen such that power for hypothesis testing when elements are independent is 80% for the given type I error and detectable difference of $\delta = 0.005$. This sample size can be deduced from **Figure 4**, for example, at $\alpha = 0.05$ and $\delta = 0.005$, the sample size needed to achieve 80% power under SRS is about 2500. This implies that in Figure 9, the number of clusters for the setting "alpha=0.05" is obtained as $N_e = \frac{2500}{n_a}$. When the elements are correlated, it can be seen that the power reduces greatly and never reaches 80% in any of the considered settings. For the highest level of correlation and largest cluster size the power for hypothesis testing is almost zero.

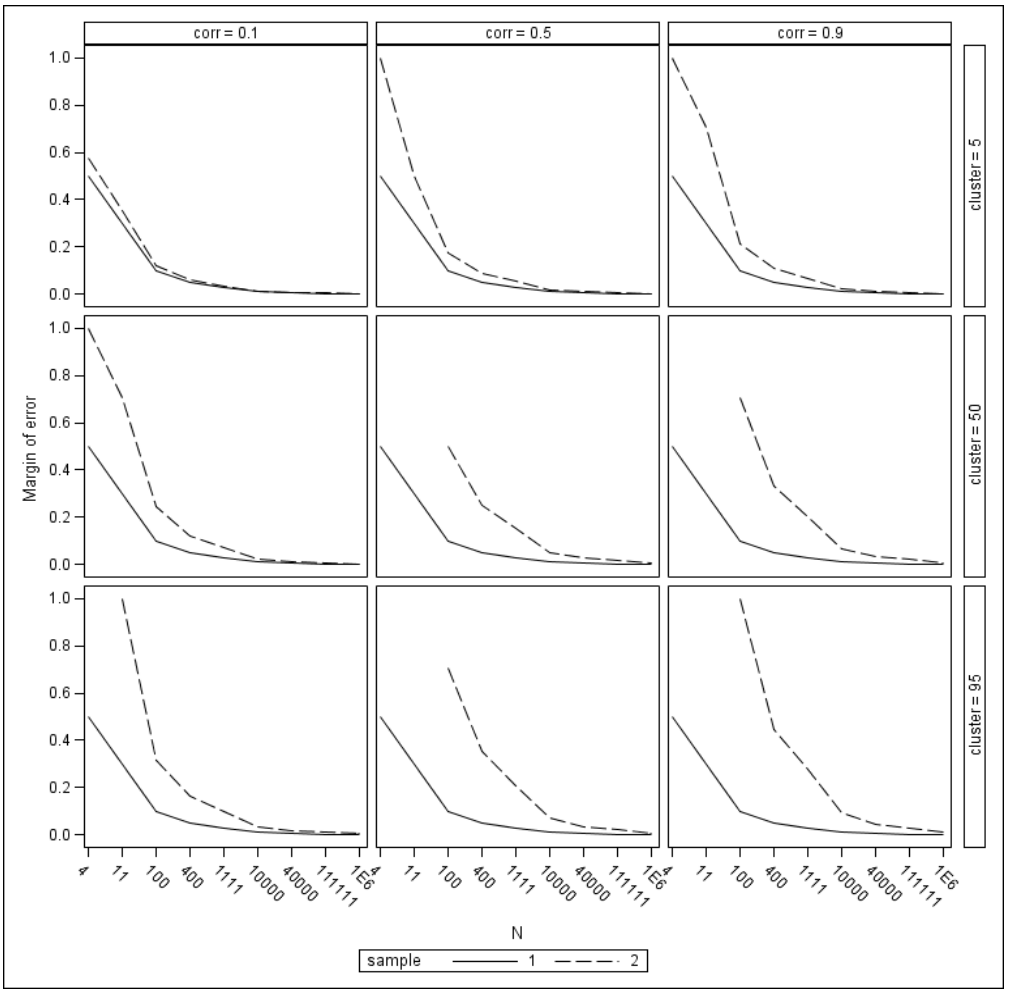

Figure 8 Comparison of the Margin of Error for Clustered Observation (Sample=2) and Independent Observations (Sample=1) for Various Cluster Sizes (Cluster) and Correlation (Corr). N is the Total Sample Size Assuming Independent Observations.

1.4.5.1. Stratification

The participating countries can be considered to constitute strata; sampling is organized within the countries themselves.

In the following, we will consider sample size calculations, and the allocation of the sample to the various countries. The assignment of the sample to the various countries will be based on proportional allocation, as well as Neyman allocation. For proportional allocation, the proportions will first be based on the current EFSA approach (by population size); later, proportions based on the food consumption will also be considered.

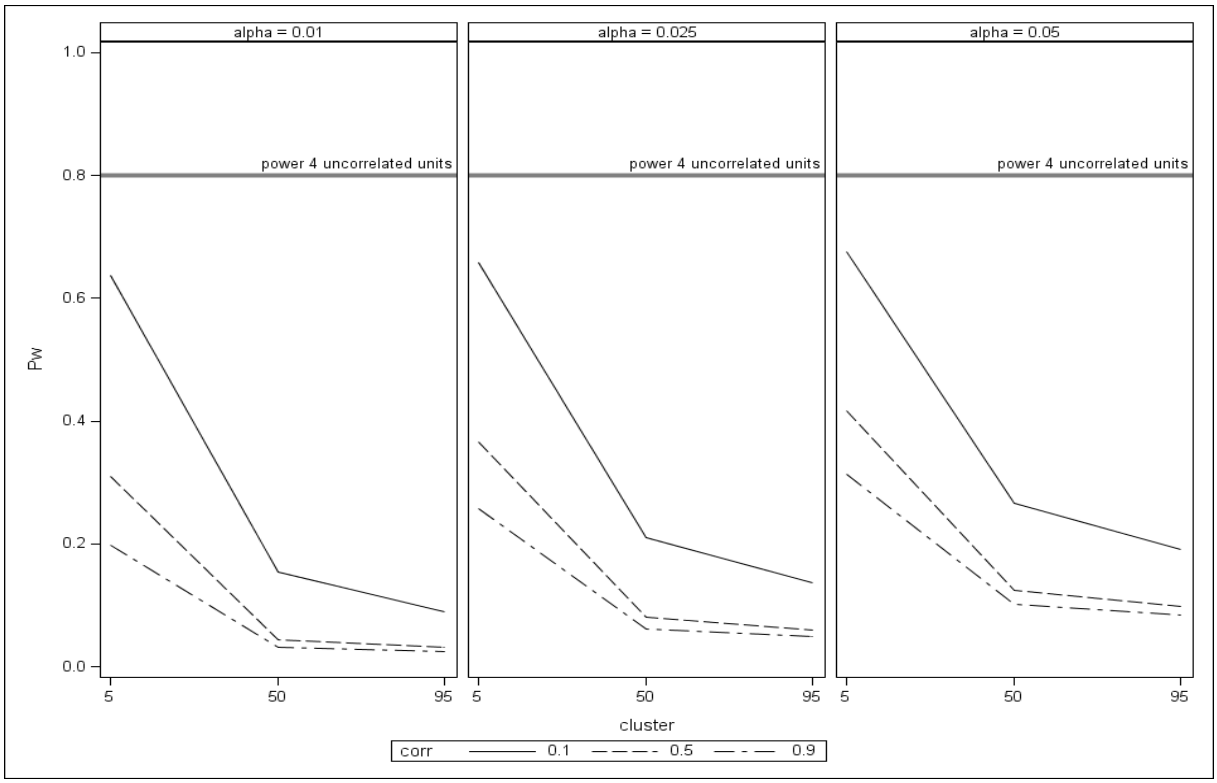

Figure 9: Comparison of the Power of Hypothesis Testing (PW) Between Samples With Clustered and Independent Observations for Various Cluster Sizes (Cluster), Correlation (Corr) and Type I Error (Alpha). The Power for Independent Observation Was Fixed At 80%.

As noted earlier, variance estimates are needed for the sample size calculations. The pesticides monitoring data will therefore be used to calculate these estimates. These will then be used to calculate sample sizes needed to estimate the proportion of samples exceeding MRLs, within specified margins of error, under the allocation schemes mentioned above.

For each stratum (country, in this case), h, and for each food item, f, the variance is estimated as follows $p_{hf}(1-p_{hf})$, where p_{hf} is the proportion of samples of the corresponding food item, exceeding MRL, in the respective country.

Therefore, to calculate the variance estimates, we first compute the proportion of samples exceeding MRLs within the various countries, for the various food items. A remark here is that these computations are solely for purposes of obtaining variance estimates for use in sample size calculations; further analyses related to obtaining overall EU proportions, as well as proportions at the analytical determination level, will be considered in Section 2, and Section 3.

Table 53 - 0 in 0 show, for each country, the proportion of samples exceeding MRL, and the estimated variances. The proportions are generally low, hence low variances. In majority of the cases, the variances are actually zero, due to zero number of samples exceeding MRL. Variance estimates above zero range from 0.0064 to 0.1600, with only three cases where the country-specific variance for a certain commodity is at least 0.1000. The average variance estimate is 0.0115.

We now consider proportional allocation, by country population. Table 7 below shows the population proportion of the 27 EU member states, and Iceland and Norway. The variance estimates computed

above will be used as rough guidelines for sample size calculation; sample sizes assuming uniform strata variances of 0.0064, 0.0100, and 0.1600, will be computed.

In Table 8, the resulting sample sizes are shown; these calculations can be performed using for instance R software. Variances larger than 0.1600 would require larger sample sizes to achieve the same margins of error, while the reverse would hold for variances less than 0.0064.

In Table 9, we for instance consider the allocation of the $n = 1537$ obtained under the assumption of uniform variances of 0.0100, in the case where a margin of error of 0.0050 is required. Overall sample size and allocation are determined simultaneously, and, due to rounding off of the allocation figures, the total of the allocation figures may differ, minutely though, from the determined overall sample size. That is why for this case, while the overall determined size is 1537, the sum of the allocation figures gives 1535.

Table 7 Population and Proportion by Country, 1 January 2010

Country	Population	Proportion of Total Population
Austria	8375290	0.0166
Belgium	10839905	0.0214
Bulgaria	7421766	0.0147
Cyprus	819140	0.0016
Czech Republic	10462088	0.0207
Denmark	5534738	0.0109
Estonia	1337666	0.0026
Finland	5351427	0.0106
France	64658856	0.1279
Germany	81802257	0.1618
Greece	11305118	0.0224
Hungary	10014324	0.0198
Iceland	317630	0.0006
Ireland	4549428	0.0090
Italy	60340328	0.1194
Latvia	2120504	0.0042
Lithuania	3141976	0.0062
Luxembourg	502066	0.0010
Malta	414027	0.0008
Netherlands	16574989	0.0328
Norway	4858199	0.0096
Poland	38167329	0.0755
Portugal	10573479	0.0209
Romania	20294683	0.0401
Slovakia	5390410	0.0107
Slovenia	2046976	0.0040
Spain	46486619	0.0920
Sweden	9340682	0.0185
United Kingdom	62471264	0.1236

Table 8 Sample Sizes for Various Uniform Variances

Variance	Margin of Error	Sample Size
0.0064	0.0010	24585
	0.0020	6146
	0.0030	2732
	0.0040	1537
	0.0050	983
	0.0060	683
	0.0070	502
	0.0062	**642**
	0.0080	384
0.0100	0.0010	38415
	0.0020	9604
	0.0030	4268
	0.0040	2401
	0.0050	1537
	0.0060	1067
	0.0070	784
	0.0077	**642**
	0.0080	600
0.1600	0.0010	614633
	0.0020	153658
	0.0030	68293
	0.0040	38415
	0.0050	24585
	0.0060	17073
	0.0070	12544
	0.0080	9604
	0.0309	**642**

Table 9 Allocation to Countries (Uniform Strata Variances of 0.0100, Margin of Error 0.005)

Country	Allocation
Austria	25
Belgium	33
Bulgaria	23
Cyprus	2
Czech Republic	32
Denmark	17
Estonia	4
Finland	16
France	197
Germany	249
Greece	34
Hungary	30
Iceland	1
Ireland	14
Italy	183
Latvia	6
Lithuania	10
Luxembourg	2
Malta	1
Netherlands	50
Norway	15
Poland	116
Portugal	32
Romania	62
Slovakia	16
Slovenia	6
Spain	141
Sweden	28
United Kingdom	190

We now compare proportional and Neyman allocation. This will be illustrated using the data on apples. We will consider the 11 countries for which there was at least one sample exceeding the MRL. For these countries, we use the estimated variances as input for the sample size calculations. The respective country populations are again used for the stratum sizes. The margin of error is specified as 0.008.

In Table 10, we show the sample size required under each criterion, as well as the allocation to the countries. Note that the computed total sample size under proportional allocation is 1006, but the sum of the allocations is 1007, due to the rounding off explained above. That is why 1006 is also indicated, in brackets.

Noteworthy is for instance the effect of the country variance on the sample allocation. While the Czech Republic and Portugal get an equal allocation through proportional allocation (owing to their fairly equal population sizes), Portugal gets a noticeably larger allocation through the Neyman allocation, owing to the larger variance.

It is also noteworthy that a smaller sample is required in the case of Neyman allocation. The gains from using Neyman allocation over proportional allocation depend on the variability between the stratum variances; the greater the variability, the larger the gains.

Note that a margin of error of 0.008 has been used in this case. Reducing the margin of error will increase the sample size requirement; for instance, a margin of error of 0.005 would remarkably increase the required sample size to 2575 and 2223, under proportional and Neyman allocation, respectively.

In multinational surveys, a combination of different allocation strategies may be considered. In the World Fertility Survey, conducted in 44 developing countries between 1974 and 1982, allocation of the sample was done not only on the basis of population, but also, for instance, on "ethnic and regional heterogeneity" (Harkness et. al, 2010).

Table 10 Proportional and Neyman Allocation

Country	Population	Variance	Allocation Scheme	
			Proportional	Neyman
Cyprus	819140	0.0663	3	6
Czech Republic	10462088	0.0363	**42**	**57**
Spain	46486619	0.0222	184	197
France	64658856	0.0074	257	158
United Kingdom	62471264	0.0069	248	148
Greece	11305118	0.0322	45	58
Luxembourg	502066	0.0475	2	3
Netherlands	16574989	0.0075	66	41
Portugal	10573479	0.0731	**42**	**81**
Romania	20294683	0.0199	81	82
Sweden	9340682	0.0197	37	37
			Sample Size	
			1007 (1006)	868

Finally, we compare proportional allocation on the basis on member state population, and on the basis of member state food consumption. Food consumption data available on the EFSA[16] website related to 19 member states. For these member states, the available information included the average consumption, the respective number of subjects, and the number of consumers out of the subjects.

To illustrate and compare proportional allocation based on the two criteria, a food category, containing no sub-categories, was selected at random. The selected category was "Coffee, tea, cocoa (expressed as liquid)". This category's average consumption per member state (based on the consumers) was multiplied by the "estimated consumer population" of the member state, where the "estimated consumer population" was computed as the member state population multiplied by the proportion of consumers out of the subjects in the food consumption data. Table 11 below gives the estimated consumption per member state.

A margin of error of 0.005 was fixed, and constant variances of 0.0100 across strata were assumed. The sample size as well as the allocation to the member states was conducted, using proportional allocation on the basis of both the member state population, and the consumption of this particular food category. The "overall" sample size was similar in both cases (n=1537). Table 12 shows the allocation to the various member states. There are noticeable differences; for instance, 227 samples would be required from Italy under allocation by population, but only 55 would be required under

[16] Concise Database summary statistics - Total population; http://www.efsa.europa.eu/en/datex/docs/datexfooddbstatistics1.xls (Download: 27-04-2014 19:38:54) 27

allocation by consumption. Note that the remark made earlier, in relation to rounding off of the allocations, vis-à-vis the computed total sample size, also applies to Table 12.

As another example, and using the same data, we consider the category "Starchy roots or potatoes". For the same member states above, 0 shows the estimated consumption, computed as described above. The allocation based on the consumption of this category, is provided in Table 14, together with that based on population. The note about rounding off of allocations, made above, also applies here.

Table 11 Consumption Estimates for "Coffee, tea, cocoa (expressed as liquid)"

Member State	Consumption
Austria	3713768488
Belgium	3839700013
Bulgaria	893439674
Czech Republic	5846326505
Denmark	4625192183
Estonia	453350381
Finland	3102223456
France	18234475363
Germany	56642382985
Hungary	1759519968
Iceland	136327387
Ireland	3248364102
Italy	7474964303
Netherlands	14693853463
Norway	2934035084
Poland	26408449900
Slovakia	2487471586
Sweden	5374678214
United Kingdom	45586528017

Table 12 Allocation by Population and by Consumption of Coffee, tea, cocoa (expressed as liquid)"

Member State	By Population	By Consumption
Austria	32	28
Belgium	41	28
Bulgaria	28	7
Czech Republic	39	43
Denmark	21	34
Estonia	5	3
Finland	20	23
France	244	135
Germany	308	420
Hungary	38	13
Iceland	1	1
Ireland	17	24
Italy	227	55
Netherlands	62	109
Norway	18	22
Poland	144	196
Slovakia	20	18
Sweden	35	40

United Kingdom	235	338

Table 13 Consumption Estimates for "Starchy roots or potatoes"

Member State	Consumption
Austria	493535089
Belgium	1052285353
Bulgaria	617692431
Czech Republic	1080628722
Denmark	619949219
Estonia	268361155
Finland	507951620
France	4313120121
Germany	10240882245
Hungary	1099627870
Iceland	25010038
Ireland	1041846451
Italy	2902974630
Netherlands	2124432273
Norway	647814149
Poland	11601433491
Slovakia	518936821
Sweden	1286267561
United Kingdom	6969138420

Table 14 Allocation by Population and by Consumption of "Starchy roots or potatoes"

Member State	By Population	By Consumption
Austria	32	16
Belgium	41	34
Bulgaria	28	20
Czech Republic	39	35
Denmark	21	20
Estonia	5	9
Finland	20	16
France	244	140
Germany	308	332
Hungary	38	36
Iceland	1	1
Ireland	17	34
Italy	227	94
Netherlands	62	69
Norway	18	21
Poland	144	376
Slovakia	20	17
Sweden	35	42
United Kingdom	235	226

1.4.5.2. Multistage Sampling

As per earlier observation, sample elements are likely selected using a multistage design. This section therefore explores sample size calculations under this design. It is assumed selection proceeds as follows: EU is subdivided into strata (MS) which are further subdivided into food categories strata (a food category is a collection of several food items). The size of each stratum is determined by the volume of consumption of a food category in a particular MS. Food items are considered as clusters, and are randomly selected from each MS-food category stratum. Finally, food samples are randomly selected from the selected food items as illustrated in Figure 10. Our goal is to obtain the total number of food samples required to estimate proportion of samples with residues above MRL with a pre-specified margin of error and type I error.

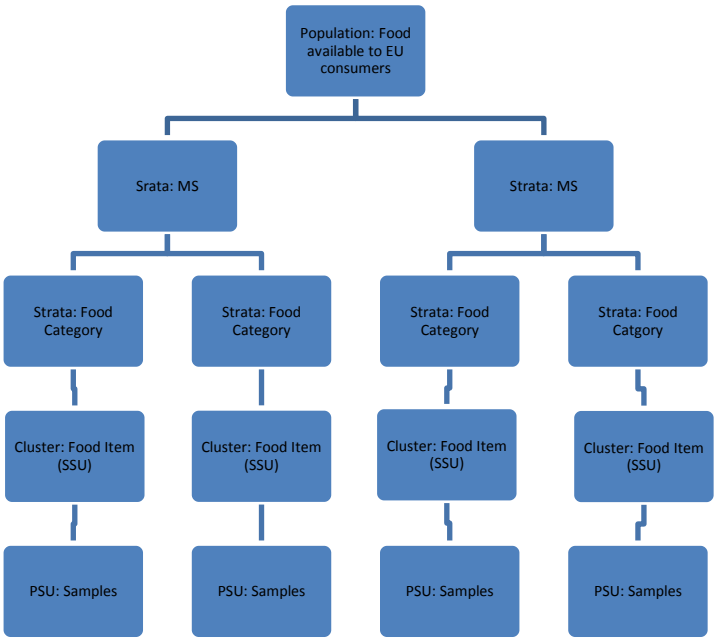

Figure 10: Multistage Design Used in Sample Size Calculation

According to formulas in Section 1.3.4 to obtain the total sample size: we need variance within each strata, and their corresponding weights(W), and the number of clusters (SSUs) or average size of cluster (number of PSUs) has to be fixed in advance. Variance within each strata is obtained by multiplying variance under SRS (S^2) by the design effect D. The values assigned to these parameters are given in Table 15. Further, values of intra-class correlation for samples within a food item were estimated from the 2010 pesticide monitoring data for each residue per sampling country. The values ranged from 0.0001 to 0.02, however, we consider values of 0.01, and 0.05 to illustrate the impact of weak and medium, respectively. Very small values of intra-class correlation are represented by the SRS design. For illustration purposes the sampling frame is simplified to include only food categories to which food items from 2010 pesticide monitoring study belong. That is we have five food category strata namely: 01=" Cereals & cereal products", 04="Vegetables, nuts, pulses including carrots, tomato", 06="Fruits", 10= "Meat and meat products, offal" 13= "Milk and dairy based products" Due to small number of clusters in each stratum, all clusters were included in the sample.

Consumption figures were extracted from consumption data summary statistics available on EFSA[17] website which has data from 19 out of 29 countries, as such countries with no data on a particular food category were assigned the minimum of the respective food category.

Consumption figures are given in grams/person/day, and we obtain the total consumption figures of a food category in a specific country (T_{ch}) by multiplying this figure with the population of the specific country (as of 1 st January 2010) as $T_{ch} = C_{ch} \times P_h$ where C_h is the c^{th} food category consumption per person for the h^{th} member state and P_h is the corresponding population. Total food consumption for all member states T_s is simply, $T_s = \sum_{h=1}^{H} T_{ch}$. Finally, each food category-members state stratum is assigned the weight $W_h = \frac{T_{ch}}{T_s}$.

The procedure to obtain sample size has been summarized in Figure 11 and Table 15 summarizes the parameter used in different scenarios. The first two scenarios aim at investigating the impact of correlation, i.e., all other settings are the same except for correlation which increases to 0.05 in scenario 2 from 0.01 in scenario 1. As expected, sample size required increases when correlation increases. Note that a correlation of 0.05 is still considerable weak yet the sample size increases by approximately 900. Scenario 3 was included to study the effect of cluster size by keeping all other settings similar to scenario 1 except for cluster size, which is increased to 20. The sample size required is larger for the bigger cluster size than the smaller cluster size although the difference is considerably small. The last two scenarios study the impact of margin of error. Indeed, the setting for scenarios 4 and 5 are similar to scenarios 1 and 2, respectively, except for margin of error which has been increased to 0.0077 in scenarios 4 and 5. The required sample size decreases with increase in margin of error. This is logical since smaller margin of error implies high level of precision for the estimates and this requires large sample size. Correlation and variance can be estimated from previous similar surveys while type I error, margin of error and cluster size have to be fixed by the researcher.

Table 15 Parameters Used in Multistage Design Sample Size Calculation.

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Type I error ($\boldsymbol{\alpha}$)	0.05	0.05	0.05	0.05	0.05
Margin of error ($\boldsymbol{d}$)	0.005	0.005	0.005	0.0077	0.0077
Correlation ($\boldsymbol{\rho}$)	0.01	0.05	0.01	0.01	0.05
Average cluster size ($\boldsymbol{b}$)	15	15	20	15	15
Stratum variance ($\boldsymbol{s_h^2}$)	0.01	0.01	0.01	0.01	0.01
Design effect ($\boldsymbol{D}$)	1.14	1.7	1.19	1.1.4	1.7
Total sample size ($\boldsymbol{n}$)	1752	2612	1829	739	1101

Results for allocation to different food items are shown in Table 65, Table 66, and 0. Though we assumed cluster sizes of 15 and 15 for estimating the design effect, the realized cluster sizes range from 0 to 190. This is a perfect example of a realistic situation since it is usually not possible to determine the cluster sizes of all clusters in a sampling frame when strata and/or clusters have different sizes. The safest way is to assume as large cluster sizes as possible though this may sometimes lead to having larger samples than necessary. Alternatively, sample a large number of clusters to make sure

[17] Concise Database summary statistics - Consumers only;

http://www.efsa.europa.eu/en/datex/docs/datexfooddbstatistics2.xls (Download: 27-04-2014 19:39:14)

that cluster sizes are as small as possible. This is perfectly illustrated by the food category "Milk and dairy based products" which has high consumption figures for many countries implying that it was allocated a relatively large sample size. However there is only one food item in the whole category hence all samples are allocated to this category will be taken from one cluster, and the cluster size will be very large. Selecting more food items from the category can reduce the average cluster size and in the process subduing the loss of precision arising from bigger cluster sizes.

Specification of cluster size and correlation can be avoided by just specifying the design effect. Recall that design effect is simply the ratio of variance under the design of interest (multi-stage design in this case) and the variance under SRS. This can be estimated from previous studies. Small values of design effect lead to smaller sample size than large values.

Most countries with no consumption data were allocated small samples since they were given the minimum consumption figures for the specific category.

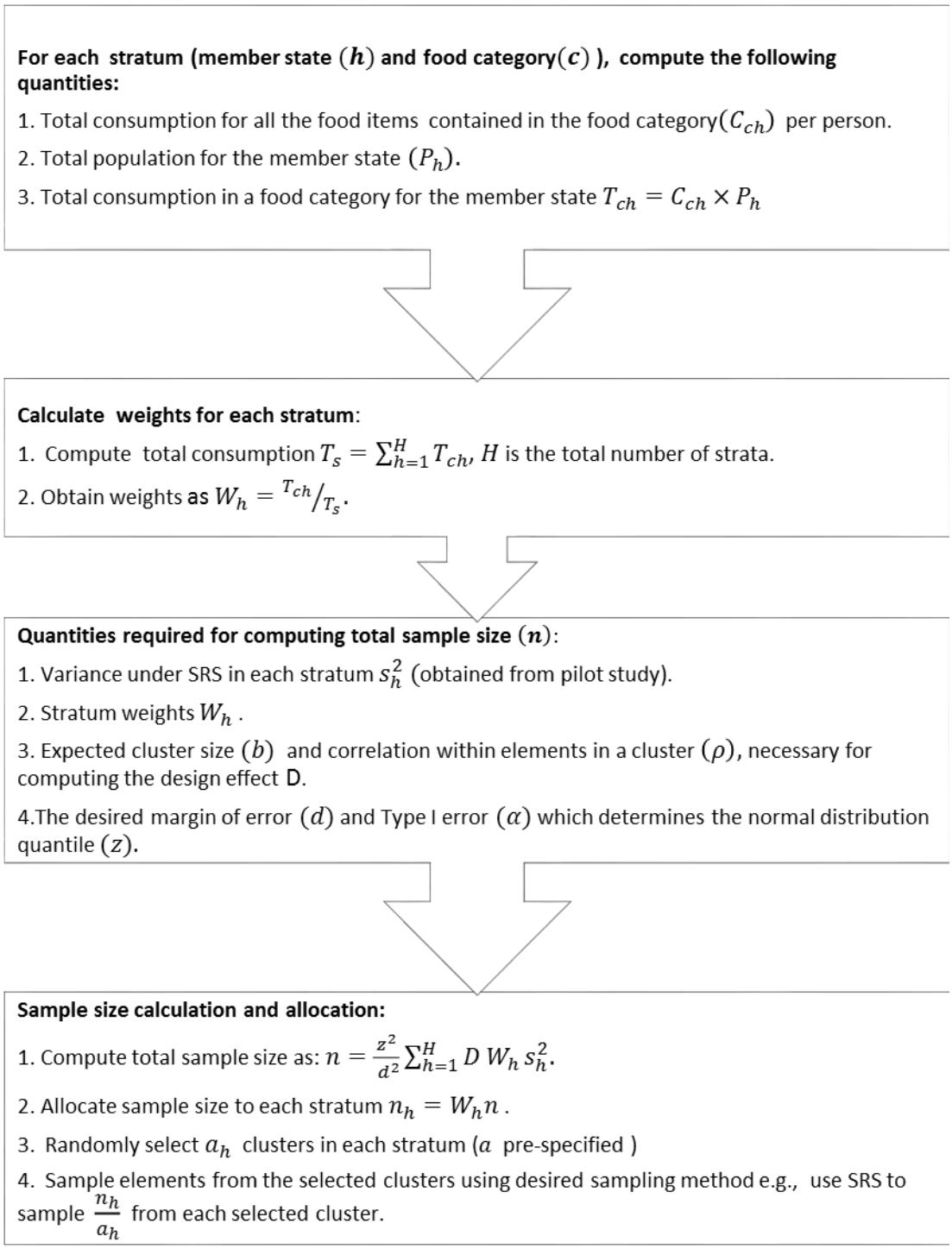

Figure 11: Procedure for Sample Size Calculation in a Multistage Design.

Table 16 Multistage Allocation for Design Effect of 1.14

Country	Oats	Rye	Head Cabbage	Leek	Lettuce	Tomatoes	Strawberries	Apples	Peaches	Pears	Swine meat	Milk	total
Austria	5	5	2	2	2	2	2	2	2	2	6	6	38
Belgium	5	5	3	3	3	3	2	2	2	2	5	8	43
Bulgaria	4	4	2	2	2	2	1	1	1	1	3	5	28
Cyprus	1	1	1	1	1	1	1	1	1	1	1	1	12
Czech republic	6	6	2	2	2	2	2	2	2	2	7	7	42
Germany	42	42	19	19	19	19	14	14	14	14	48	91	355
Denmark	3	3	1	1	1	1	1	1	1	1	3	8	25
Estonia	1	1	1	1	1	1	1	1	1	1	1	2	13
Spain	13	13	5	5	5	5	3	3	3	3	18	15	91
Finland	2	2	1	1	1	1	1	1	1	1	3	9	24
France	37	37	12	12	12	12	8	8	8	8	47	61	262
United Kingdom	28	28	9	9	9	9	6	6	6	6	36	56	208
Greece	4	4	2	2	2	2	1	1	1	1	5	4	29
Hungary	5	5	2	2	2	2	2	2	2	2	7	10	43
Ireland	2	2	1	1	1	1	1	1	1	1	3	5	20
Iceland	1	1	1	1	1	1	1	1	1	1	1	1	12
Italy	29	29	14	14	14	14	11	11	11	11	30	46	234
Lithuania	1	1	1	1	1	1	1	1	1	1	2	2	14
Luxembourg	1	1	1	1	1	1	1	1	1	1	1	1	12
Latvia	1	1	1	1	1	1	1	1	1	1	1	1	12
Malta	1	1	1	1	1	1	1	1	1	1	1	1	12
Netherlands	7	7	3	3	3	3	2	2	2	2	9	23	66
Norway	2	2	1	1	1	1	1	1	1	1	2	9	23
Poland	21	21	10	10	10	10	10	10	10	10	35	25	182
Portugal	3	3	2	2	2	2	1	1	1	1	5	4	27
Romania	6	6	3	3	3	3	2	2	2	2	8	7	47
Sweden	5	5	1	1	1	1	1	1	1	1	5	13	36
Slovenia	1	1	1	1	1	1	1	1	1	1	1	1	12
Slovakia	4	4	1	1	1	1	1	1	1	1	4	2	22
Total													1944

Table 17 Multistage Allocation for Design Effect of 1.7

Country	Oats	Rye	Head Cabbage	Leek	Lettuce	Tomatoes	Strawberries	Apples	Peaches	Pears	Swine meat	Milk	total
Austria	8	8	3	3	3	3	3	3	3	3	9	8	57
Belgium	7	7	4	4	4	4	2	2	2	2	7	12	57
Bulgaria	6	6	3	3	3	3	1	1	1	1	5	7	40
Cyprus	1	1	1	1	1	1	1	1	1	1	1	1	12
Czech republic	8	8	2	2	2	2	2	2	2	2	11	11	54
Germany	62	62	28	28	28	28	21	21	21	21	72	135	527
Denmark	4	4	2	2	2	2	2	2	2	2	4	12	40
Estonia	1	1	1	1	1	1	1	1	1	1	2	3	15
Spain	19	19	8	8	8	8	5	5	5	5	27	23	140
Finland	3	3	1	1	1	1	1	1	1	1	4	13	31
France	54	54	18	18	18	18	12	12	12	12	69	90	387
United Kingdom	41	41	14	14	14	14	8	8	8	8	53	83	306
Greece	5	5	2	2	2	2	2	2	2	2	7	6	39
Hungary	7	7	3	3	3	3	3	3	3	3	10	14	62
Ireland	3	3	2	2	2	2	1	1	1	1	4	8	30
Iceland	1	1	1	1	1	1	1	1	1	1	1	1	12
Italy	43	43	20	20	20	20	17	17	17	17	44	68	346
Lithuania	2	2	1	1	1	1	1	1	1	1	2	2	16
Luxembourg	1	1	1	1	1	1	1	1	1	1	1	1	12
Latvia	1	1	1	1	1	1	1	1	1	1	2	2	14
Malta	1	1	1	1	1	1	1	1	1	1	1	1	12
Netherlands	10	10	5	5	5	5	3	3	3	3	14	34	100
Norway	3	3	1	1	1	1	1	1	1	1	3	14	31
Poland	32	32	15	15	15	15	15	15	15	15	52	37	273
Portugal	5	5	2	2	2	2	1	1	1	1	7	6	35
Romania	9	9	4	4	4	4	2	2	2	2	12	10	64
Sweden	8	8	2	2	2	2	2	2	2	2	8	19	59
Slovenia	1	1	1	1	1	1	1	1	1	1	2	1	13
Slovakia	5	5	2	2	2	2	1	1	1	1	6	3	31
Total													2815

2. Assessment of the Impact of Design, Sample Size Used and Population Characteristics That Might be Inappropriate or Ignored During the Inference Process

GENERAL FRAMEWORK

Various types of errors in designing a survey can introduce bias rendering a sample non-representative and therefore not fit for the purpose, (Cochran, 1977; Kish, 1965). This section aims at illustrating the impact of some commonly committed errors during survey design analysis. Where possible, simulations will be used for illustrations.

2.1. Discrepancies in Sampling Design and Analysis

The rule of thumb in analysing survey data is to ensure that analysis reflects the sampling design used in selecting the sample (Kish, 1965). Ignoring the sampling design may introduce bias and/or reduce precision of estimates due to various reasons. In line with observations made in Section 1.4.5 where the sample size needed to achieve the same margin of error is larger under cluster sampling than simple random sampling, repercussions can be expected if data collected under cluster sampling is treated like SRS sample during analysis. Through a simulation study, we give a detailed overview of things that can go wrong when sampling design is ignored during analysis.

2.1.1. Simulation Study

2.1.1.1. Data Simulation.

The study was designed to investigate the impact of analysing survey data that was collected using a multistage design by methods that ignore some, or all of multistage sampling design aspects. To this end, a population was generated such that it has a multistage structure as follows: The sampling frame was defined to contain 29 strata (resembling member states within EU), and within each strata are 150 clusters (equivalent to food items in the pesticide monitoring study). Elements within each stratum are the sampling units and are equivalent to samples taken from food items in the pesticide monitoring study. Our interest is to estimate the mean of a binary trait (Y) in the population and also test the hypothesis that the mean is greater than a certain value. A good example of a binary trait is testing whether a sample taken from a food item is compliant with MRL or not.

Notably, we only have one level of stratification, contrary to two levels of stratification as illustrated in Section 1.4.5.2. This does not obstruct applicability of results from this simulation study to the above mentioned setting. Indeed, even in settings with more than one stratification level, allocation is always done at one level (the smallest level), hence it suffices to use one level of stratification for illustration purposes.

Characteristics of the population were decided as follows: each stratum was assigned the mean of the binary trait ($\bar{Y}_h$). The values for $\bar{Y}_h$ were obtained as the overall mean non-compliance rate for the 29 MS in the 2010 pesticides monitoring study. Further, cluster specific means ($\bar{Y}_{hc}$) were allocated to each of the clusters such that:

$$\bar{Y}_c = \frac{\exp(c)}{1+\exp(c)} \quad \text{where } c \sim N(0,0.25).$$

The mean for the c^{th} cluster from stratum h was obtained as,

$$\bar{Y}_{hc} = \frac{\exp(s_h + c)}{1 + \exp(s_h + c)} \qquad where\ s_h = \log\left(\frac{\bar{Y}_h}{1 - \bar{Y}_h}\right).$$

Next, the cluster size for the c^{th} cluster from stratum h, (N_{ch}) was randomly generated as

$$N_{ch} \sim \text{Lognormal } (5,2),$$

So as to obtain clusters of varying sizes ranging from one to thousands.

Finally, the value of the binary trait for the k^{th} element in the c^{th} cluster and h^{th} stratum (Y_{hck}) was generated such that outcomes from the same cluster were more similar than those from a different cluster. A method by Lunn and Davies, (1998) for generating correlated binary data was utilized. It basically involves three stages: (i) generate random variables $O \sim \text{Bernoulli}(1, \bar{Y}_{hc})$ and $P \sim \text{Bernoulli}(1, \bar{Y}_{hc})$, (ii) Generate $Q \sim \text{Bernoulli}(1, \rho_{hc})$ where ρ_{hc} is the desired intra-class correlation coefficient. (iii) Compute the outcome as

$$Y_{hck} = O(1 - Q) + PQ.$$

Outcomes from the same cluster are thus correlated with a correlation coefficient equal to ρ_{ch}.

The reasonable assumption of exchangeable correlation within a cluster was used, hence the subscripts for ρ_{ch} will be dropped (since there is no particular order within the cluster it is reasonable to assume that correlation is the same for any two members of the same cluster). Two values of $\rho = (0.02, 0.3)$ were considered to help illustrate the impact of weak and strong correlation. Table 18 gives some characteristics of the generated population. A total of 4949407 elements were generated with strata means ranging from 0.00004 to 0.0588, and the mean of outcome was 0.0152 which (similar to the overall mean observed in 2010 pesticides monitoring study).

2.1.1.2. Aim

The objective is to illustrate the impact of ignoring crucial aspects of a sampling design during analysis. This will be achieved by selecting a sample from the population generated above using a multistage sampling design and analyze it assuming a multistage, SRS, cluster, and stratified SRS sampling designs.

2.1.1.3. Sample Selection

The selected sample is to be used in

1. Estimating the overall mean of the outcome of interest($\bar{Y}$), in the targeted population.
2. Testing hypotheses:
   a. $H_0: \bar{Y} \leq \kappa$ vs $H_1: \bar{Y} > \kappa$ where κ is a pre − specified value
   b. $H_0: \bar{Y} \leq \kappa - (\delta + 0.001)$ vs $H_1: \bar{Y} > \kappa - (\delta + 0.001)$.

Where (a) and (b) are relevant for estimating type I & II errors, respectively. The 0.001 is added to the detectable difference δ to ensure that the alternative hypothesis holds such that not rejecting the null hypothesis would lead to type II error.

Table 18 Some Characteristics of the Simulated Population for $\rho = 0.3$.

Stratum	Population Size (N_h)	Percent of Population ($W_h * 100$)	Mean of Outcome($\bar{Y}_h$)	Weighted Mean
1	112054	0.0226	0.0385	0.0009
2	330239	0.0667	0.0001	0.0000
3	239181	0.0483	0.0060	0.0003
4	133632	0.0270	0.0505	0.0014
5	91033	0.0184	0.0372	0.0007
6	137573	0.0278	0.0126	0.0003
7	131353	0.0265	0.0037	0.0001
8	153567	0.0310	0.0001	0.0000
9	152501	0.0308	0.0258	0.0008
10	225273	0.0455	0.0031	0.0001
11	91567	0.0185	0.0525	0.0010
12	91479	0.0185	0.0165	0.0003
13	482158	0.0974	0.0179	0.0017
14	102412	0.0207	0.0001	0.0000
15	214884	0.0434	0.0037	0.0002
16	168728	0.0341	0.0001	0.0000
17	124053	0.0251	0.0057	0.0001
18	116437	0.0235	0.0087	0.0002
19	180427	0.0365	0.0198	0.0007
20	137219	0.0277	0.0000	0.0000
21	91508	0.0185	0.0588	0.0011
22	156528	0.0316	0.0157	0.0005
23	160214	0.0324	0.0047	0.0002
24	129281	0.0261	0.0137	0.0004
25	191252	0.0386	0.0282	0.0011
26	164729	0.0333	0.0089	0.0003
27	183368	0.0370	0.0160	0.0006
28	122026	0.0247	0.0163	0.0004
29	334731	0.0676	0.0270	0.0018
Total	4949407	1.0000		0.0152

Though we have considered a one-sided hypothesis only, impact of ignoring some aspects of the design on type I error and power of testing observed from these simulation studies directly apply to two-sided hypothesis as well. Further, the results do not depend on the magnitude of the estimate, i.e., the simulation results can be generalized to cases where the estimate is larger than 0.015.

Similar to procedure followed in Section 1.3.4, the initial step involves calculating the total sample size required for each objective. Population variance S^2 was estimated as,

$$S^2 = \sum_{h=1}^{29} W_h \bar{Y}_h(1 - \bar{Y}_h)D,$$

where $D = 1 + \rho(10 - 1)$, is the design effect and 10 was the assumed average cluster size in the selected sample. Other parameters are as indicated in Table 18. Total sample size was obtained as

$$n = \frac{[\Phi(1-\alpha)]^2}{d^2} S^2, \text{and } n = \frac{[\Phi(1-\alpha) + \Phi(1-\beta)]^2}{\delta^2} S^2,$$

for estimation and hypothesis testing, respectively. $d, \delta, \alpha, \beta,$ and Φ, represent the margin of error, detectable difference, type I error rate, type II error rate, and the inverse standard normal cumulative distribution function, respectively. The parameters were assigned the following values $d = 0.0077, \delta = 0.0077, \alpha = 0.05,$ and $\beta = 0.2$.

Sample size allocation to strata follows the proportional allocation as illustrated in Section 1.2.3, with the weights provided in column "Percent of the Population" of Table 18.

Recall that we cannot compute total samples size, total number of clusters required, and number elements to be sampled from a cluster simultaneously; hence the number of clusters was fixed to 348 out of 4350 clusters (348 would correspond to 12 clusters [number of food items in the 2010 pesticides monitoring study] from each stratum if the strata were of equal size). Stratified simple random sampling with proportional allocation was used to allocate the number of clusters to be sampled from each stratum with W_h as weights. This was to ensure that large sized strata contributed larger number of clusters than smaller ones.

Number of elements to be selected from each of the selected clusters within each stratum was determined through stratified systematic sampling. This is a form of sampling that is equivalent to SRS if elements to be sampled do not have a monotone order (Kish, 1965). Basically, all elements from selected clusters within a stratum are brought together to form a sub-population. Let n_{hs} and n_h represent total number of elements in the sub-population and the sample size allocated to h^{th} stratum, respectively. The sampling probability for each element becomes $C = n_h/n_{hs}$. Selection of the required sample proceeds by taking a random number between1 and $\xi = 1/C$, and taking the ξ^{th} element after (Kalton, 1983). The simulation of 200 replicate samples of size n were selected in all considered scenarios.

All selection procedures can be easily done with statistical software like SAS or R. For example a stratified design with systematic sampling can be achieved by the following SAS code:

```
proc surveyselect data=inputdata seed=1
      method=sys  sampsize=348 out=out_data;
strata sid/alloc=allo_wt;
run;
 where
```

Data specifies the sampling frame in form of a SAS dataset.
Seed ensures that the same sample is selected every time the code is run.
Method indicates the preferred method of sampling the elements and sys refers to systematic sampling.
Sampsize is the required sample size.
Out specifies the name of the dataset to store the selected sample.
Strata specifies the stratifying variable and *alloc* names the dataset with stratification weights.

2.1.1.4. Estimation:

Mean:

Estimates for the population mean and variance are obtained as

$$\bar{y} = \frac{1}{V}\sum_{v=1}^{V=200} \bar{y}_v, \quad \text{and} \quad \text{Var}(\bar{y}) = \frac{1}{V}\sum_{v=1}^{V=200} Var(\bar{y}_v),$$

where $\bar{y}_v$ is the estimated mean in the v^{th} replicate. Estimation of $\bar{y}_v$ and $Var(\bar{y}_v)$ depends on the method of analysis.

Relative bias and margin of error are obtained as

$$Rbias = \frac{\bar{y} - \bar{Y}}{\bar{Y}}, \qquad and \qquad \hat{d} = z\sqrt{\text{Var}(\bar{y})}.$$

Four analyses, presented in Table 19, were conducted. The "Surveylogistic" procedure in SAS was used to account for survey design aspects, like stratification and clustering.

Table 19 Performed Survey Sample Analysis

	Design used in Analysis			
Design Aspect Accounted for	**Multistage Sampling**	**Cluster Sampling**	**Simple Random Sampling**	**Stratified simple Random Sampling**
Stratification	Yes	No	No	Yes
Clustering	Yes	Yes	No	No

Type I Error:

For each sample replicate 95% confidence intervals are constructed around the sample estimate with

$$lcl = \bar{y} - \hat{d}, \qquad \text{and}$$

$$ucl = \bar{y} + \hat{d}$$

as lower and upper limits, respectively and

$$\hat{d} = z\sqrt{Var(\bar{y})}.$$

Type I error is estimated by the proportion of times that the confidence interval does not include the true mean, i.e., let

$$f_v = \begin{cases} 0, & lcl \le \bar{Y} \le ucl \\ 1, & \text{otherwise} \end{cases}, \qquad \Rightarrow \hat{\alpha} = \frac{1}{V}\sum_{v=1}^{V=200} f_v$$

Hypothesis Testing:

Under the hypothesis testing objective $\bar{y}$, Var $(\bar{y})$, $Rbias$, and $\hat{\alpha}$, are estimated like in the estimation objective above. Type II error is estimated by the proportion of times we fail to reject the null hypothesis when the alternative is true, e.g., for $H_0: \bar{Y} \le 0.005$ $H_1: \bar{Y} > 0.005$, let

$$c_v = \frac{\bar{y} - 0.005}{\sqrt{Var(\bar{y})}}, \qquad \text{and } f_v^* = \begin{cases} 1, & z \ge c_v \\ 0, & \text{Otherwise} \end{cases}.$$

Type II error was computed as

$$\hat{\beta} = \frac{1}{V}\sum_{v=1}^{V=200} f_v^*,$$

And power was estimated as $\hat{p} = 1 - \hat{\beta}$.

2.1.1.5. Results

Table 20 gives results for the scenario where sample size is calculated based on the estimation objective. In this scenario differences in the variance of the estimated mean can be among the various analyses there are observed differenced in the variances of the estimated mean, where designs that do not take into account clustering (SRS and stratified SRS) have smaller variances than cluster and multistage sampling which take clustering into account. This is in line with our expectations since the former designs assume more information than there actually is. This is also reflected in the level of type I error. For the setting where $\rho = 0.3$ type I error is below the pre-specified level (5%) for multistage and cluster sampling, and it is doubled to 11.5% for SRS and stratified SRS. The power for estimation is always above the pre-specified power of 80%. This may not be the case where lack of precision (e.g., large standard errors) is severe, the power would be greatly reduced. When intraclass correlation is reduced to 0.02, a scenario close to having independent elements, differences among the various design almost disappear.

Table 20 Impact of Ignoring Aspects of Sampling Design During Analysis of a Sample Collected for Estimation Objective: Recall that ρ is the intraclass correlation, $\bar{Y}$ is the true overall mean, $\bar{y}$ is the estimated overall mean, R.bias is the relative bias, $\hat{\alpha}$ is the estimated type I error and $\hat{d}$ is the estimated margin of error.

Method of Analysis	ρ	n	$\bar{Y}$	$\bar{y}$	Var($\bar{y}$)	R.bias	$\hat{\alpha}$	$\hat{d}$
Cluster Sampling	0.3	3531	0.015	0.014	.0000077	0.007	0.02	0.005
Multistage Sampling			0.015	0.014	.0000072	0.007	0.025	0.005
SRS			0.015	0.014	.0000043	0.007	0.115	0.004
Stratified SRS			0.015	0.014	.0000042	0.007	0.115	0.004
Cluster Sampling	0.02	1130	0.015	0.015	.0000015	-0.053	0.045	0.008
Multistage Sampling			0.015	0.015	.0000015	-0.053	0.050	0.008
SRS			0.015	0.015	.0000013	-0.053	0.045	0.007
Stratified SRS			0.015	0.015	.0000013	-0.053	0.050	0.007

This implies that if we draw many sample replicates and analyze the data assuming a stratified SRS, 11.5% of the 95% confidence intervals of the sample mean will not include the true mean. Usually in practice only one sample is collected and if the chance that the confidence interval will not include the true mean is large, results from such a sample can hardly be reliable. The situation is similar for hypothesis testing where type I error is inflated to 8.5%, and 6% for interclass correlation coefficient of 0.3 and 0.02, respectively. On the other hand, SRS and stratified SRS have smaller variances, and narrower margin of error lengths than the other designs. This phenomenon is related to the concept of effective sample size discussed in Section 1.2.2.

Table 21 Impact of ignoring aspects of sampling design during analysis of a sample collected for hypothesis testing objective. Recall that ρ is the intraclass correlation, $\bar{Y}$ is the true overall mean, $\bar{y}$ is the estimated overall mean, R.bias is the relative bias, $\hat{\alpha}$ is the estimated type I error and $1-\hat{\beta}$ is the estimated power of hypothesis testing.

Method of Analysis	ρ	n	$\bar{Y}$	$\bar{y}$	Var($\bar{y}$)	R.bias	$\hat{\alpha}$	$1-\hat{\beta}$
Cluster Sampling	0.3	5683	0.015	0.015	.0000059	-0.009	0.010	0.000
Multistage Sampling			0.015	0.015	.0000055	-0.009	0.015	0.000
SRS			0.015	0.015	.0000026	-0.009	0.085	0.000
Stratified SRS			0.015	0.015	.0000026	-0.009	0.085	0.000
Cluster Sampling	0.02	1818	0.015	0.015	.0000074	-0.033	0.035	0.100
Multistage Sampling			0.015	0.015	.0000069	-0.033	0.045	0.090
SRS			0.015	0.015	.0000044	-0.033	0.060	0.075
Stratified SRS			0.015	0.015	.0000043	-0.033	0.060	0.075

Outcomes of elements within the same cluster are correlated which implies that clusters contain less information than if outcomes were independent. The two SRS based analyses assume independence, and wrongly assumes that the sample has a lot of information which leads to smaller estimated variances for the sample mean. Smaller variances implies extremely narrow confidence intervals with a reduced chance of containing the true mean value despite the average estimate ($\bar{y}$) being almost unbiased. This is better illustrated in Figure 12 where the confidence limits (the dotted lines) are narrower for SRS and stratified SRS than cluster and multistage sampling design analysis. Larger estimated variances observed for cluster sampling are well reflected in the overly conservative confidence intervals. While chances of committing type I error are minimal, unnecessarily wide confidence interval do not provide meaningful inferences.

Multistage design analysis which accounts for all aspects of sampling design used in selecting the sample, gives expected results. This underscores the importance of having a recognizable sampling design since it plays a crucial role during analysis.

2.2. Selection Bias

This type of bias mainly arises from ignoring unequal sampling probabilities. Several scenarios lead to elements in a sample being selected with unequal probabilities. Examples include; oversampling from a population subgroup to be able to draw some inferences about it, non-proportionate stratification, and selecting same number of elements from clusters with different sizes. Selection bias often results when the selection probabilities depend on the outcome of interest and it can be positive or negative depending on the elements that were given higher selection probabilities. For example, if food items originating from third world countries are known to have high non-compliance rate and these are given high probability of selection, positive bias is expected. This type of selection bias can only be corrected for, if the exact relationship between the outcome and sampling probability is known, otherwise some untestable assumptions have to be made.

On the other hand unequal sampling probabilities that do not depend on the outcome, e.g., in the case where food items are considered as clusters, this would result when same number of samples are collected from food items with differing volume of consumption (i.e., selecting the same number of elements from clusters with different size). This kind of bias can be easily corrected for by taking into account the sampling probability weights during analysis. If analysis ignores the sampling probabilities, variance of the estimate is likely to be underestimated and this may impact quantities like type I error. Through a simulation study we illustrate impact of unequal sampling probabilities resulting from: i). sampling same number of elements from clusters with different sizes, ii). unequal sampling probabilities that depend on the outcome of interest.

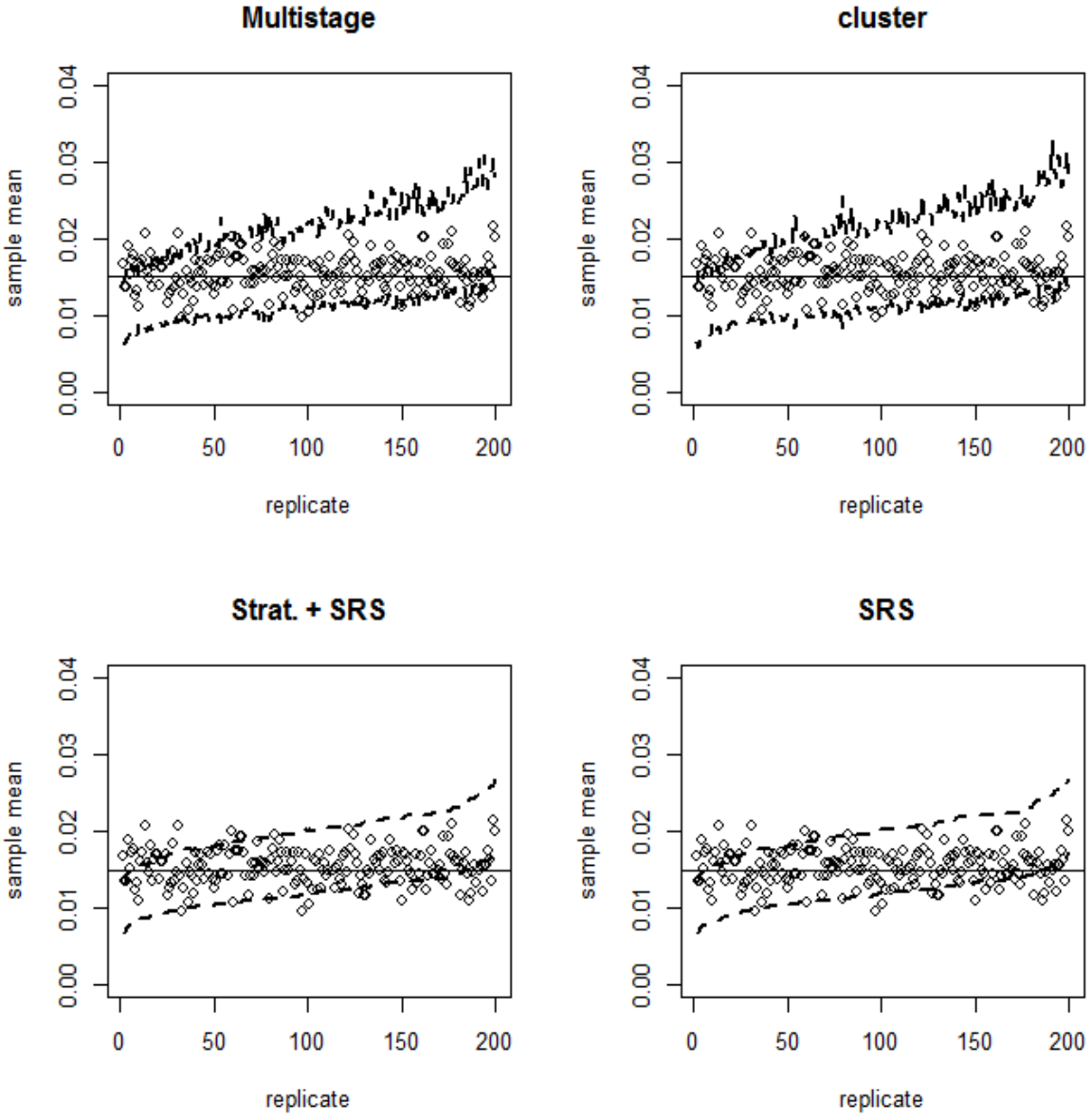

Figure 12 Scatter plot for estimates of the mean obtained from 200 replicates (sorted from lowest to highest) for the estimation objective. The horizontal line indicates the true mean and helps in showing the variability of sample mean estimates from the true mean. The dotted line represent the confidence limits.

2.2.1. Simulation Study

Objective of this simulation study is to illustrate the impact of using unequal selection probabilities. Study design and estimation procedures are similar to those in Section 2.1.1, and the main difference is in the selection of the sample.

Two settings were considered: the first setting investigate the impact of unequal selection probabilities that are related to the outcome of interest. Specifically, consider a likely situation in the pesticide monitoring study where food items with reasonably high consumption are given high probability of selection. In addition, assume that highly consumed food items are more likely to have non-compliant samples than lowly consumed food items. Thus selection probabilities based on food consumption will implicitly give non-compliant samples high chance of being selected.

A second setting studies the impact of selecting the same number of elements from clusters with different sizes and ignoring the resulting unequal sampling probabilities. Only the multistage design is considered in this case.

2.2.1.1. Data Simulation

Data generation process was almost similar to that in Section 2.1.1.1 with a slight changes in computation of the mean. Cluster sizes and cluster means were generated as

$$N_{ch} \sim Lognormal\ (5,2),\ \text{ and }\ \bar{Y}_c = \frac{\exp(c)}{1+\exp(c)} \quad \text{where } c \sim N(0,0.25), \text{respectively,}$$

and these were sorted such that larger sized clusters were allocated high cluster mean values i.e., the mean for a cluster with size $N_{ch(i)}$ was $\bar{Y}_{c(i)}$ where (i) indicates the i^{th} smallest value. This also meant that the first stratum had clusters with smallest means and the 29th stratum had clusters with the highest means. Only the case with $\rho = 0.3$ and the assumed cluster size for sample size calculations was reduced to five (to reduce sample size for faster computations) and all other settings were similar to those in Section 2.1.1.1. Some characteristics of this populations are presented in Table 22.

2.2.1.1. Sample Selection

Unequal selection probabilities that depend on the outcome:

Sample selection proceeded as follows: proportional allocation was used to allocate the total sample size to the strata with weights as given in the column $\boldsymbol{W_h}$ of Table 22. Next, clusters were selected within each stratum where the probability of selection for each clusters was set to $\frac{N_{ch}}{N_h}$, where N_h is the total size for h^{th} stratum as given in the second column of Table 22, and N_{ch} is the cluster size. This ensured that large sized clusters were given high probability of selection. Next, all elements from the selected clusters were brought together to form a subpopulation from which final elements were selected with probability equal to $\frac{N_{ch}}{N_{hs}}$, where N_{hs} is the size of the h^{th} stratum in the subpopulation. The implication was that elements from the same clusters were given the same probability of selection and elements from larger clusters had higher probability of selection than elements from small clusters.

Analyses taking into account different attributes of design will be considered, further the multistage design will take into account the sampling weights defined as the inverse of selection probability.

These are commonly used to restore representativeness of the sample in face of differing selection probabilities.

Table 22 Some Characteristics of the Simulated Population Where Large Clusters Have High Means.

Stratum	Population Size (N_h)	Proportion of Population (W_h)	Mean of Outcome($\bar{Y}_h$)	Weighted Mean
1	125021	0.0253	0.0733	0.0019
2	126182	0.0255	0.0002	0.0000
3	128255	0.0259	0.0099	0.0003
4	132140	0.0267	0.1061	0.0028
5	133271	0.0269	0.0726	0.0020
6	133700	0.0270	0.0314	0.0008
7	136849	0.0276	0.0095	0.0003
8	137406	0.0278	0.0002	0.0000
9	139168	0.0281	0.0469	0.0013
10	139510	0.0282	0.0074	0.0002
11	140996	0.0285	0.0895	0.0025
12	144549	0.0292	0.0347	0.0010
13	145343	0.0294	0.0590	0.0017
14	148070	0.0299	0.0003	0.0000
15	149138	0.0301	0.0083	0.0002
16	149994	0.0303	0.0003	0.0000
17	151386	0.0306	0.0121	0.0004
18	157859	0.0319	0.0181	0.0006
19	161907	0.0327	0.0532	0.0017
20	164717	0.0333	0.0003	0.0000
21	169652	0.0343	0.1240	0.0043
22	180176	0.0364	0.0478	0.0017
23	187454	0.0379	0.0153	0.0006
24	190503	0.0385	0.0393	0.0015
25	201215	0.0407	0.0769	0.0031
26	241596	0.0488	0.0302	0.0015
27	260232	0.0526	0.0542	0.0029
28	286724	0.0579	0.0692	0.0040
29	386394	0.0781	0.1207	0.0094
Total	4949407	1.0000		0.0468

Unequal Selection Probabilities independent of the outcome:

For this setting we studied the impact of selecting the same number of elements from clusters with different sizes. Selection of the sample proceeded as follows: The total sample size was allocated to strata through proportional stratified sampling with weights as given in the second column of Table 22. Next, stratified SRS sampling was used to select clusters within each stratum and SRS was used to select $\frac{n_h}{a_h}$ elements from each of the selected clusters, where n_h is the sample size allocated to h^{th} stratum and a_h is the corresponding number of selected clusters. Since clusters within a stratum had different sizes, selecting the same number of elements from each selected cluster implied that elements from smaller clusters would have higher probability of selection than elements from bigger clusters. Elements from small sized clusters would therefore be over represented.

On the other hand, since stratified SRS was used for selecting clusters, the dependence between selection probability and the outcome was minimize.

We compare two kinds of analysis, one that takes into account the different sampling probability as weights hence making the sample representative and one that ignores them, but both analyses take the multistage design into account.

2.2.1.2. Results

Unequal selection probabilities not related to the outcome:

There was considerable bias when unequal selection probabilities were not taken into account even when they were not related to the outcome as observed in Table 23 under the column "Not weighted". Bias for the weighted analysis was 14.5% compared to 63% for the analysis with no weights for the estimation objective. Figure 14 illustrates the differences in the distribution of the mean estimates from different replicates, where the median of the weighted analysis is very close to the true mean (the dotted line) than the median of un-weighted analysis.

Due to negative bias, the power of hypothesis testing was greatly reduced in both cases, 13% and 0% for the weighted and un-weighted analysis respectively. The hypothesis tested in this case was:

$$H_0: \bar{Y} \leq 0.047 - (0.009) \text{ vs } H_1: \bar{Y} > 0.047 - (0.009),$$

Hence the correct conclusion was H_1, and concluding H_0 would lead to type II error. Since the mean was mostly underestimated H_0 was concluded for a large number of replicate samples thereby increasing type II error and reducing power in the process.

Another notable impact was in the underestimation of variance in the un-weighted analysis leading to a highly inflated type I error of 100%.

Table 23 Impact of Selecting a Sample With Unequal Selection Probabilities That are not Related to the Outcome.

Method of Analysis	Objective	n	$\bar{Y}$	$\bar{y}$	Var($\bar{y}$)	R.bias	$\hat{\alpha}$	$\hat{d}$	$\hat{\beta}$
Weighted	Estimation	4008	0.047	0.040	.0001670	-0.145	0.175	0.025	
Not weighted			0.047	0.017	.0000035	-0.630	1.000	0.004	
Weighted	Hypothesis	9159	0.047	0.040	.0001315	-0.139	0.025		0.870
Not weighted	Testing		0.047	0.018	.0000021	-0.619	0.000		1.000

Unequal selection probabilities related to the outcome:

Simulation results are presented in Table 24. It is clear that when selection probabilities and the outcome are related even taking into account sampling probability weights does not reduce the bias. Type I error was inflated to 98% for multistage analysis, implying that 95% confidence interval will almost never include the true outcome mean. Type I error was highly inflated for all analyses. This is because the crucial information about the relationship between selection probabilities and the outcome was not accounted for during analysis. Notice the increase in relative bias of the mean, which is a direct result of selecting more elements from clusters with higher means. Interestingly, bias was positive for all other analyses except for the multistage design which took into account sampling probability weights. In the scenario where selection probabilities and the outcome are not related,

sampling probability weights can reduce bias yet they have an opposite effect in this case. Recall that a cluster was selected with probability $\frac{N_{ch}}{N_h}$ and the probability of selecting an element given that the cluster it belongs to was selected was $\frac{N_{ch}}{N_{hs}}$ hence the marginal probability of selecting an element was

$$\frac{N_{ch}}{N_h} \times \frac{N_{ch}}{N_{hs}}.$$

Note that for each replication and the same stratum, N_h and N_{hs} will remain constant hence the marginal probability will vary depending on the cluster size N_{ch}. Using inverse probability as sampling weights would always result in large weights for small clusters and this may lead to their over representation. Since small sized clusters were allocated smaller means, underestimation of the mean would be likely.

Due to the negative bias, power of hypothesis testing was reduced to 2% in multistage sampling. Figure 13 further illustrates the impact of unequal selection bias where is clear that sample means were consistently under or over estimated.

Table 24 Impact of Selecting a Sample With Unequal Selection Probabilities That Depend on the Outcome.

Sampling Design	Objective	N	$\bar{Y}$	$\bar{y}$	Var($\bar{y}$)	R.bias	$\hat{\alpha}$	$\hat{d}$	$\hat{\beta}$
Cluster	Estimation	4008	0.047	0.062	.0000188	0.317	1.00	0.008	
Multistage			0.047	0.028	.0001036	-0.409	0.99	0.020	
SRS			0.047	0.062	.0000226	0.317	1.00	0.009	
Stratified SRS			0.047	0.062	.0000216	0.317	1.00	0.009	
Cluster	Hypothesis Testing	9159	0.047	0.062	.0000133	0.327	1.000		0.000
Multistage			0.047	0.034	.0000830	-0.282	0.000		0.980
SRS			0.047	0.062	.0000064	0.327	1.000		0.000
Stratified SRS			0.047	0.062	.0000061	0.327	1.000		0.000

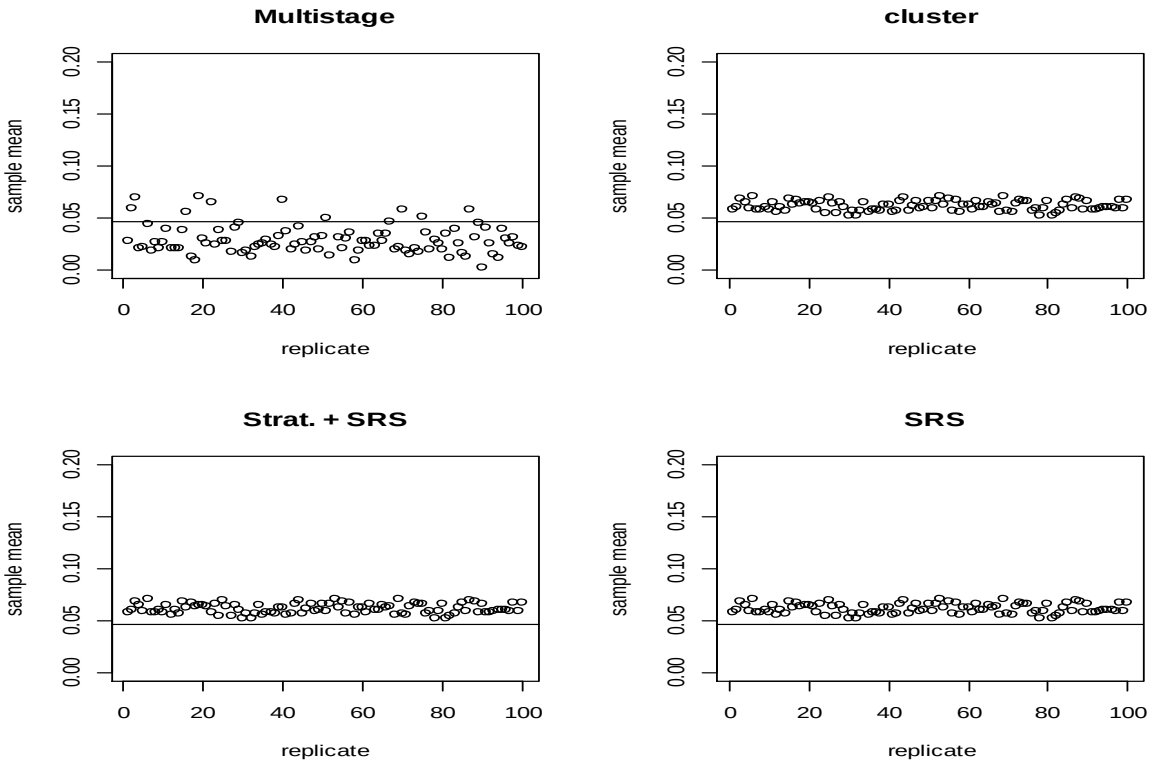

Figure 13 Scatter Plot for Mean Estimates From the 200 Replications to Illustrate Variability of Sample Means From the True Mean for the Selection Bias Scenario Where Selection Probabilities are Related to the Outcome.

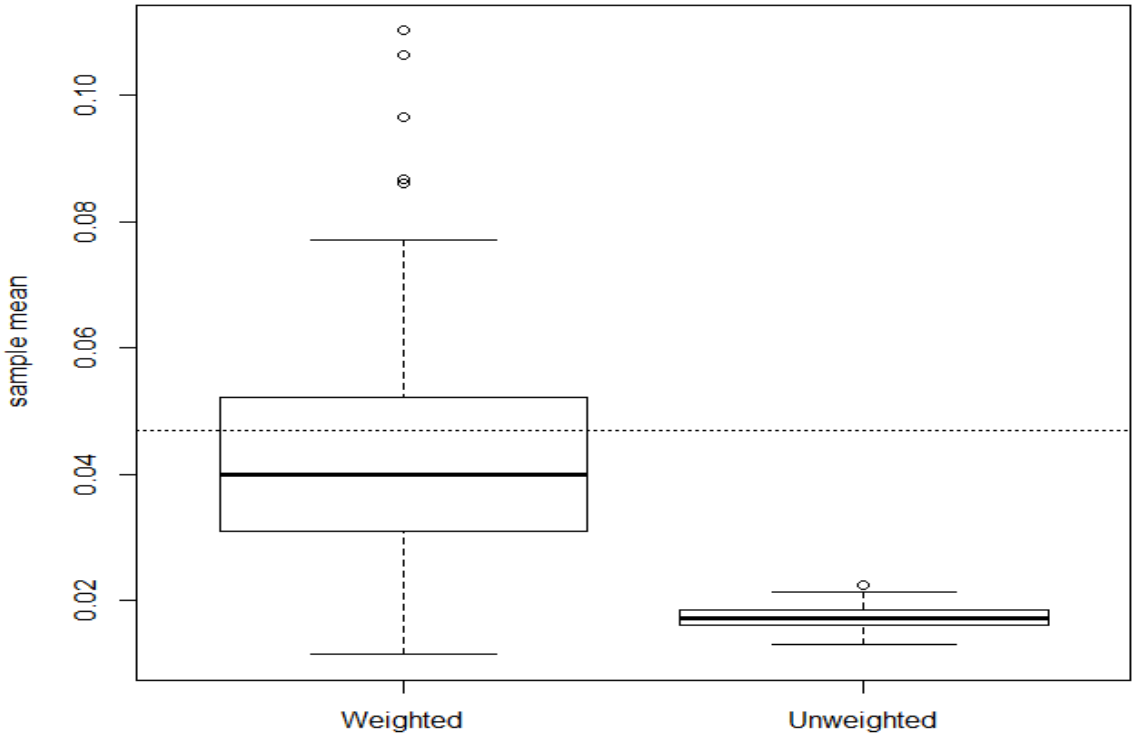

Figure 14 Boxplot Weighted and Un-Weighted Analysis for the Selection Bias Scenario Where Selection Probabilities are not Related to the Outcome.

2.3. Non-coverage Bias

This bias arises when there are discrepancies between the target population and the frame population (or sampling frame, i.e. the population from which the sample is actually selected). This can be due to an inadequate sampling frame or flaws in the implementation of the data collection. When elements of the target population are excluded, there is under-coverage, while inclusion of elements erroneously results in over-coverage. Over-coverage typically results in increased cost but no significant bias, while under-coverage often results in biases, which are difficult to detect and evaluate. Coverage errors largely affect the representativeness of the sample; for instance, under-coverage of certain sectors of the target population will result in a sample that is not representative. For example when a food item is widely consumed but only used in processed products but the consumer is unaware of this and this ingredient is not identified in consumption surveys and thus not included in the pesticides monitoring study.

The degree of the impact of coverage errors depends largely on the gap between the target population and the sampling frame. Appropriateness, relevance and timeliness of the sampling frame are therefore critical in minimizing the risk of a non-representative sample. We illustrate the impact of under-coverage using a simulation study.

2.3.1. Simulation Study

The study was executed in a similar way to Section 2.1.1 for data generated with $\rho = 0.3$ and a few modifications to suit objectives of this section, which is to show the impact of using a non-exhaustive sampling frame. To achieve this objective the following scenarios and modifications are considered:

a. Sampling frame is reduced by including only clusters with mean greater than the 25th percentile of all cluster means. These is referred to as "No lowest means" scenario. For the pesticides monitoring study, this is similar to omitting food items that are suspected to have low exceedance rates.

b. Sampling frame is reduced by including only clusters with mean less than the 75th percentile of all cluster means. These is referred to as "No highest means" scenario. For the pesticides monitoring study, this is similar to omitting food items that are suspected to have high exceedance rates.

c. Sampling frame is reduced by including only clusters with means between 25th and 75th percentile of all cluster means. These is referred to as "No highest & lowest means" scenario. For the pesticides monitoring study, this is similar to omitting food items that are suspected to have lowest and highest exceedance rates.

Estimation and hypothesis testing proceeds as in Section 2.1.1.4 and only multistage analysis was used. Results are presented in Table 25 and Figure 15.

Table 25 Impact of Using a Non-Exhaustive Sampling Frame

Coverage	Objective	n	$\bar{Y}$	$\bar{y}$	Var($\bar{y}$)	R.bias	$\hat{\alpha}$	$\hat{d}$	$\hat{\beta}$
Full Coverage	Estimation	3531	0.015	0.015	.0000072	0.007	0.025	0.005	
No highest & lowest means			0.015	0.013	.0000053	-0.128	0.060	0.005	
No highest means			0.015	0.012	.0000047	-0.230	0.225	0.004	
No lowest means			0.015	0.017	.0000078	0.119	0.070	0.005	
Full Coverage	Hypothesis Testing	5683	0.015	0.015	.0000055	-0.009	0.015		0.000
No highest & lowest means			0.015	0.013	.0000039	-0.123	0.005		0.005
No highest means			0.015	0.012	.0000034	-0.231	0.000		0.025
No lowest means			0.015	0.017	.0000059	0.129	0.170		0.000

The impact of non-coverage is highest when clusters with mean less than the 25th percentile of all cluster means are excluded from the sampling frame. Type I error was inflated to 22% and relative bias was as high as 23% for the estimation objective. Type I error is small for the hypothesis testing due to negative bias. Certainly, a sampling frame that attains almost a full coverage is crucial in ensuring representativeness of the data. Since all scenarios have the same sample size, the impact on variance is almost negligible when compared to the impact on bias. The direction of bias depends on

excluded clusters, it is positive when clusters with low means (or low risk food items) are excluded and negative high means clusters (high risk food items). This is clearly illustrated in Figure 15

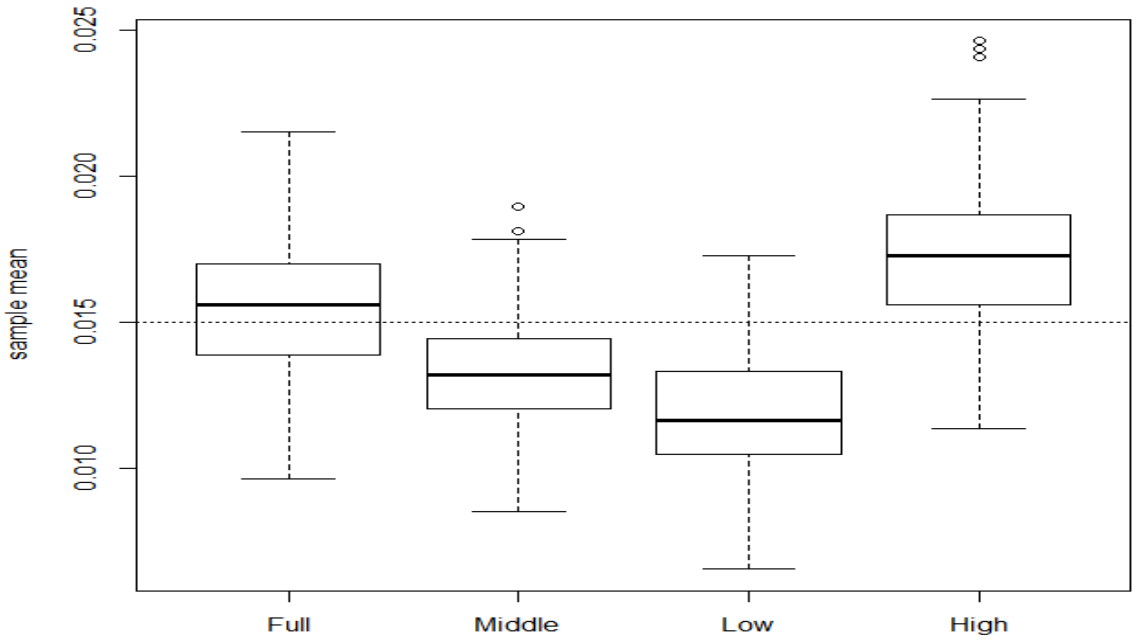

Figure 15: Boxplot for Sample Means Under Various Scenarios of Non-Coverage. Full, is when the sampling frame correctly represents the target population, High, omits clusters with low means from the sampling frame, Low, omits cluster with high means from the sampling frame, and middle omits clusters with both highest and lowest means.

2.4. Sufficiency of Sample Size

Having enough sample size is essential in achieving some acceptable level of accuracy, and drawing relevant inferences from the sample. The impact of sufficiency of sample size was investigated by drawing different sample size under multistage design as explained in Section 2.1.1.1, and analysis also takes into account the multistage design.

Additional scenarios were included to investigate the impact of the choice of number of clusters to be selected. For the same sample size, choosing smaller number of clusters to be selected implies bigger cluster sizes which may reduce precision depending on the strength of intraclass correlation. In food surveys consumption of a food item plays a crucial role in deciding the number of samples to be collected. However, consumption and granularity of the food item may not always match, for example, for the same consumption volume, berries would have lower level of granularity than oranges and thus more samples may be required from berries. Assuming size of a cluster is defined by consumption, this would imply that equally sized clusters may end up with different sizes in the sample, bigger/or small depending on the granularity of the food item. We therefore investigate the impact of the size of cluster on estimation and hypothesis testing.

Results of this investigation for the estimation objective are presented in Table 26. As expected, precision reduced with decrease in sample size and this was evidenced by the increase in estimated variance for small sample sizes. This is also reflected in margin of error.

Table 26 Investigating Impact of Sample Size Sufficiency and Choice of Number of Clusters

# of clusters	n	$\bar{Y}$	$\bar{y}$	Var($\bar{y}$)	R.bias	$\hat{\alpha}$	$\hat{d}$
696	3531	0.015	0.015	0.000006	-0.013	0.025	0.005
	1766	0.015	0.015	0.000010	-0.033	0.035	0.006
	955	0.015	0.014	0.000016	-0.047	0.050	0.008
	484	0.015	0.014	0.000031	-0.071	0.065	0.011
348	3531	0.015	0.015	0.000007	0.007	0.025	0.005
	1766	0.015	0.015	0.000011	0.005	0.035	0.006
	955	0.015	0.015	0.000018	-0.004	0.050	0.008
	484	0.015	0.014	0.000031	-0.065	0.025	0.011
174	3531	0.015	0.015	0.000008	-0.025	0.035	0.005
	1766	0.015	0.015	0.000011	-0.021	0.045	0.007
	955	0.015	0.015	0.000019	-0.032	0.045	0.008
	484	0.015	0.014	0.000030	-0.063	0.070	0.011

On the other hand, decreasing the number of clusters to be sampled, which in turn increases the cluster size seem to be associated with decrease in precision as well. The difference is not huge for the studied number of clusters probably due to the weak correlation within clusters; nevertheless it confirms that larger clusters may reduce precision.

2.5. Non-response (Missing data) and Non-response Bias

The problem of missing data is common in surveys. All the data for a selected survey element may be missing, or only part of this data may be available.

Apart from a reduction in the sample size, missing data has potential for creating bias in the results.

It is important to distinguish between the three mechanisms that cause missing data (Little and Rubin, 2002; Molenberghs and Verbeke, 2005; Rubin, 1976; Verbeke and Molenberghs, 2000). The missingness mechanism is termed as missing completely at random (MCAR) if the probability of missingness depends on neither the observed nor the missing data.

On the other hand, data are referred to as missing at random (MAR) if the probability of missingness depends on the observed data, but not on the unobserved data.

Data are then termed as missing not at random (MNAR) if the probability of missingness depends on the unobserved data.

Consider the following two examples from the pesticides monitoring programme. First, each country is supposed to report a certain number of samples for each food item. However, some countries do not meet the number requirement for the various food item; either they provide fewer samples, or no samples at all, for particular items. The samples not provided can be termed missing.

Now, suppose that the missing samples would systematically have been MRL non-compliant, had they been reported. Alternatively, suppose the reverse were true; that the missing samples would systematically have been MRL compliant, had they been reported. It's clear that under such circumstances, the probability that the samples are missing would be dependent on their "unobserved" MRL compliance status. This is what MNAR mechanisms entail; that the probability that the data are missing depends on the data values that would have been observed, had these data not been missing.

Though countries may fail to meet the number requirement for the various food items, at least there is partial information which is available, coming from the samples reported for the particular, and/or other, food items. This available information constitutes the observed data. If the probability that the samples are missing does not depend anymore on the unobserved data, given that the observed data has been taken into account, then we have a MAR mechanism operating. The question as to how this observed data is properly taken into account, is addressed in Section 3.

If the missing samples are purely a random subset of all the samples that would have been realized (had all the reporting countries completely complied), such that the fact that they are missing is neither related to what would have been observed, nor what has been observed, then the mechanism is MCAR.

As a second example, consider the following. For most of the food items, the participating countries are expected to analyze and report results on multiple prescribed residues from each laboratory sample. However, not all residues are tested/reported. Therefore, the analytical determinations related to the corresponding non-tested/non-reported residues are not available. These analytical determinations can be considered missing data.

Clearly, if the missing analytical determinations would have systematically turned positive or negative, had they been done, then the MNAR mechanism applies. On the other hand, if the probability that these analytical determinations are missing is unrelated to the results that would have been observed, given that the information available (observed) on the rest of the determinations is controlled for, then the MAR mechanism applies. Finally, if the missing analytical determinations are truly a random subset of all "potential" determinations, then the MCAR mechanism applies.

As mentioned above, methods for tackling missing data will be discussed in Section 3.

Now, if the MCAR assumption holds, then standard methods of data analysis can be used. However, this assumption is generally unrealistic, and application of standard methods when this assumption is violated leads to bias.

To illustrate the bias that results from standard analyses when the MCAR assumption does not hold, we first generate a sample of $n = 1000$ independent trials from a Bernoulli distribution with success probabilities $\pi = 0.001, 0.500,$ and 0.900. For a number of times $B = 1000$, subsamples of size $m = 500$ are created. These subsamples are generated by deleting 500 observations according to the following 5 different schemes, which represent different missingness mechanisms:

Mechanism 1 (MNAR 1): Probability that 0 values are missing is 90%, whereas the probability that 1 values are missing is 10%. This is a missing not at random mechanism, with higher missingness probability for the 0 values.

Mechanism (MNAR 2): Probability that 0 values are missing is 75%, whereas the probability that 1 values are missing is 25%. This is again a missing not at random mechanism.

Mechanism 3 (MCAR): The 0 values and the 1 values have the same probability of missingness; 50%, 50%, respectively. This is a missing completely at random mechanism.

Mechanism 4 (MNAR 3): Probability that 1 values are missing is 90%, whereas the probability that 0 values are missing is 10%. This is a missing not at random mechanism, with higher missingness probability for the 1 values.

Mechanism 5 (MNAR 4): Probability that 1 values are missing is 75%, whereas the probability that 0 values are missing is 25%. This is again a missing not at random mechanism.

Practically, each subsample is created by independently (with replacement) sub-sampling the values, and using the respective probabilities. For each of the $B = 1000$ subsamples, and for each of the missingness mechanisms, the success probability was estimated by simply taking the mean of the 500 available values. An overall estimate for the mean was obtained as

$$\hat{\pi} = \frac{1}{B}\sum_{B=1}^{1000} \hat{\pi}_B$$

The bias for each scenario was then estimated as the difference between the true mean and the overall estimate:

$$\widehat{Bias} = \hat{\pi} - \pi$$

The results are summarized in Table 27 below.

Table 27 Bias Induced by Simple Analyses Under Missingness

Missingness Mechanism	π	$\hat{\pi}$	$\widehat{Bias}$
MNAR 1	0.0010	0.0089	0.0079
MNAR 2		0.0029	0.0019
MCAR		0.0009	-0.0001
MNAR 3		0.0003	-0.0007
MNAR 4		0.0001	-0.0009
MNAR 1	0.5000	0.8972	0.3972
MNAR 2		0.7453	0.2453
MCAR		0.4934	-0.0066
MNAR 3		0.2459	-0.2541
MNAR 4		0.0976	-0.4024
MNAR 1	0.9000	0.9887	0.0887
MNAR 2		0.9668	0.0668
MCAR		0.9078	0.0078
MNAR 3		0.7665	-0.1335
MNAR 4		0.5226	-0.3774

Noteworthy is the bias under the MNAR mechanisms. As expected, positive bias is induced by a higher probability of missingness in the 0 values, while negative bias is seen in the case where the 1 values have a higher probability of missingness. The magnitude of the bias increases as the imbalance between the respective probabilities of missingness increase. There is minimal bias when the mechanism is MCAR.

The estimated means under these mechanisms, for the three values of the true mean, are presented in Figure 16, Figure 17, and Figure 18, together with the true mean.

The message in this section is that whenever some of the data, which were supposed to be available, are unavailable, caution needs to be exercised with standard analysis methods.

Now, sometimes, data may be unavailable, but some "partial" information about them may be known. For instance, a data value may be "missing", but it may be known to be, say, below a certain value. This is the topic of the next section, namely, that of left censoring.

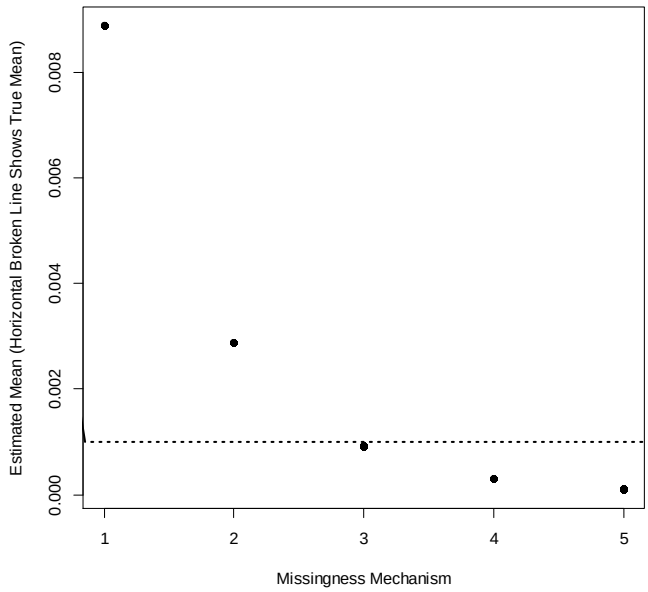

Figure 16 Estimated mean under the different missingness mechanisms, for $\boldsymbol{\pi = 0.001}$.

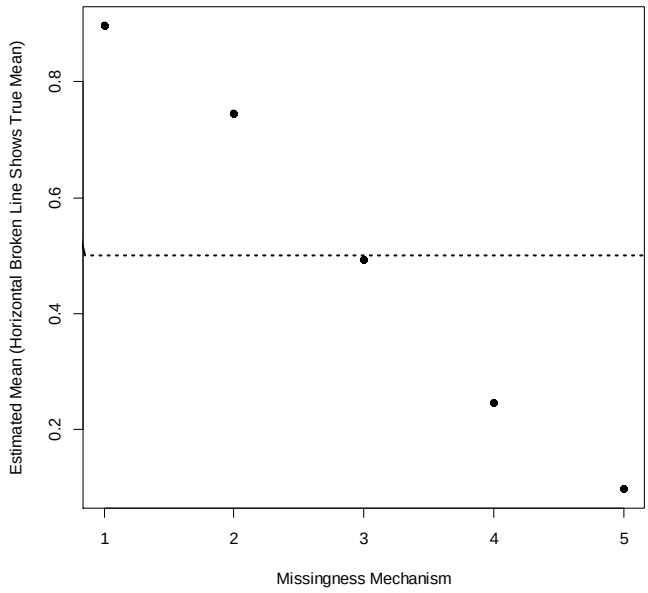

Figure 17 Estimated mean under the different missingness mechanisms, for $\boldsymbol{\pi = 0.500}$.

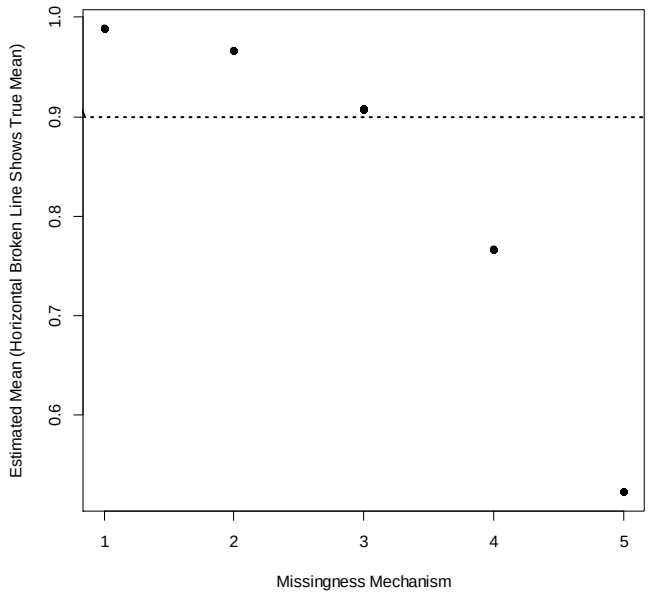

Figure 18: Estimated mean under the different missingness mechanisms, for $\boldsymbol{\pi = 0.900}$.

2.6. Left Censoring

Left censoring is a phenomenon in which a quantity is only known to be below a certain value. This issue has potential to bias analyses, and it was the focus of EFSA, 2010[18], where different approaches for handling this problem were studied.

In that study, left censored data were generated from a variety of distributions, with different percentages of censoring. Various analysis approaches were considered, and the resulting bias studied. One general finding was that the impact of left censoring increases with increase in the censoring percentage.

We hereby briefly illustrate the impact of this problem. In this section, this problem will be explored in general, after which we will focus on the pesticides monitoring data in Section 2.10. Finally, in Section 3, we propose how analyses should proceed in the presence of this problem.

We first introduce the lognormal distribution. A random variable X is said to follow the lognormal distribution with scale and shape parameters μ and σ^2, if its logarithm, Y = log (X), follows the normal distribution, with mean μ and variance σ^2. There is supporting information for the use of this distribution for concentration data[19]. The mean of the lognormal distribution, ϑ, is related to the mean and variance of the normal distribution as follows:

$$\vartheta = \exp\left(\mu + \frac{\sigma^2}{2}\right). \quad (1)$$

Using the log-normal distribution, $B = 1000$ samples, of size $n = 1000$, were first generated. For each of these samples, $20\%, 40\%, 60\%$ and 80% of the values were censored. The mean was then estimated for each of these samples, and for each of these censoring percentages. In computing the mean, in analogy to the replacement of values less than LOQ with the LOQ (as described to have been done for the EFSA 2010 report), all values below the value at which the values were censored, were replaced with that particular "censoring value".

The parameters μ and σ^2 above were first set to 0 and 1 respectively. Given these, the true mean of the values (assumed to follow the log-normal distribution) was

$$\vartheta = \exp\left(\mu + \frac{\sigma^2}{2}\right) = 1.648721$$

For each percentage of censoring, the mean was estimated by

$$\hat{\vartheta} = \frac{1}{B}\sum_{B=1}^{1000} \hat{\vartheta}_B$$

[18] European Food Safety Authority; Management of left-censored data in dietary exposure assessment of chemical substances. EFSA Journal 2010;8(3):.[96pp.]. doi: 10.2903/j.efsa.2010.1557.

[19] European Food Safety Authority; Management of left-censored data in dietary exposure assessment of chemical substances. EFSA Journal 2010;8(3):.[96pp.]. doi: 10.2903/j.efsa.2010.1557.

The bias was then estimated as

$$\text{Bias} = \hat{\vartheta} - \vartheta$$

The percentage bias was then estimated as

$$\text{Percent Bias} = \frac{\hat{\vartheta} - \vartheta}{\vartheta} \times 100$$

The results are shown in Table 28 below. The percent bias increases as the percentage of censoring increases, up to as high as 69%. In Figure 19 and Figure 20 below, we show the overestimation of the mean, and the increase in the percentage bias with increase in censoring percentage, respectively.

Table 28 Bias due to Left Censoring

True Mean	% Censoring	Estimated Mean	Bias	% Bias
1.648721	20	1.6824	0.0337	2.0442
	40	1.7878	0.1391	8.4349
	60	2.0499	0.4012	24.3319
	80	2.7904	1.1417	69.2466

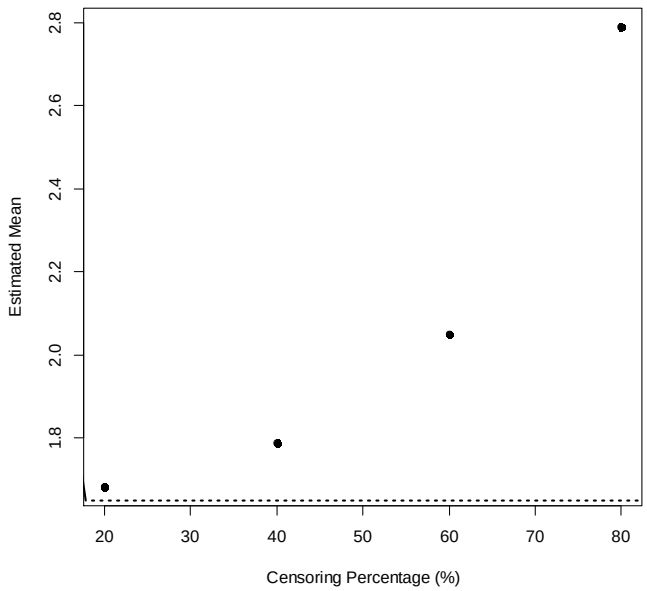

Figure 19 Estimated Mean Under Left Censoring. Broken Line Shows the True Mean.

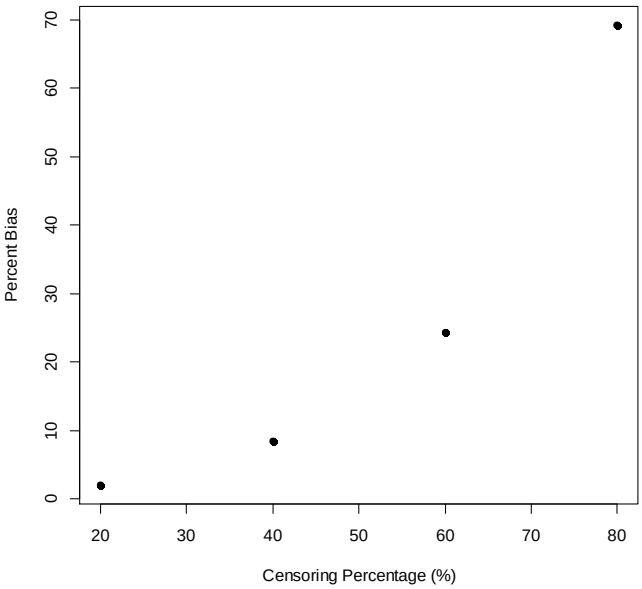

Figure 20: Percent Bias Under Left Censoring

RESULTS: PESTICIDE MONITORING DATA

2.7. Impact of Discrepancies in Sampling Design and Analysis.

To investigate the impact of sampling design accounted for during analysis, the 2010 pesticide monitoring data was re-analysed with different designs. The models considered were

$$\hat{\pi}_f = \mathrm{P}(s_{if} = 1) = \frac{\exp(\beta_{0f})}{1 + \exp(\beta_{0f})}, \text{and } \hat{\pi}_t = \mathrm{P}(s_i = 1) = \frac{\exp(\beta_{0t})}{1 + \exp(\beta_{0t})},$$

where $\hat{\pi}_f$, is the proportion of samples above MRL for food item f and s_{if} is the i^{th} labsample from food item f. A labsample is considered to be above MRL (1) if at least one pesticide has residues above MRL and zero otherwise. Similarly, $\hat{\pi}_t$ gives the overall proportion of samples above MRL for all food items, and s_i is the i^{th} labsample. In practice only the model on the right is fitted, and the model on the left is requested as a domain (sub) analysis.

Analyses considered include; SRS as done in the report, stratified SRS with sampling country as strata, and proportion of their corresponding population sizes to the total EU population as weights. Stratified cluster sampling considers sampling country as strata, and food item as cluster. Results are shown in Table 29. Notable differences for some food items like apples, lettuce, oats, and rye, suggest the need to account for the appropriate design to ensure reliability of the results. For some food items (e.g., apples and rye), the mean is noticeably lower in stratified sampling than SRS, possibly because strata means for those food items vary considerably. Variance under stratified cluster sampling is the largest and those under SRS are the smallest as expected. As observed in simulations, high precision is not always good news; it can lead to inflated type I error.

Table 29 Re-Analysis of 2010 Pesticide Monitoring Data With Different Designs

Food Item	Sampling Design Used for Analysis					
	Simple Random Sampling		Stratified SRS		Stratified Cluster Sampling	
	Proportion	Std. Error	Proportion	Std. Error	Proportion	Std. Error
Apples	0.0131	0.0025	0.0054	0.0016	0.0054	0.0025
Head cabbage	0.0090	0.0030	0.0067	0.0034	0.0067	0.0063
Leek	0.0104	0.0033	0.0097	0.0048	0.0097	0.0038
Lettuce	0.0338	0.0046	0.0464	0.0076	0.0464	0.0131
Oats	0.0528	0.0143	0.0736	0.0204	0.0736	0.0484
Peaches	0.0183	0.0039	0.0147	0.0052	0.0147	0.0014
Pears	0.0129	0.0057	0.0161	0.0072	0.0161	0.0062
Rye	0.0025	0.0025	0.0005	0.0005	0.0005	0.0006
Strawberries	0.0275	0.0046	0.0273	0.0067	0.0273	0.0101
Tomatoes	0.0123	0.0026	0.0184	0.0051	0.0184	0.0043
Overall	0.0162	0.0011	0.0181	0.0018	0.0181	0.0046

If the SRS design was appropriate, we would expect to have similar results for the different analyses. Note that the results above are only for illustration purposes and no conclusions can be made about the preferred analysis since their validity depend on correctness of the assumptions made. For example, we have assumed that the sample allocation was proportional to population size, results from stratified sampling can therefore be given meaningful interpretation only when this was indeed the case. All analyses were implemented using the "Surveylogistic" procedure in SAS. The specific programming codes are given below:

Programming codes for obtaining results in Table 29 using SAS 9.4 (Comments are provided in green)

```
/*SRS Design*/
proc surveylogistic data=ms;/* data specifies the dataset*/
model s(event='1')=/  link=logit; /*s is the response variable*/
domain food_item;  /*Domain asks for individual means for each food item*/
run;



/*Stratified SRS*/
proc surveylogistic data=ms;
strata  sampcountry; /*Specifies sampling country as stratification variable*/
model s(event='1')=/  link=logit;
domain food_item;
run;



/*Stratified Cluster Sampling */
proc surveylogistic data=ms;
strata  sampcountry;
cluster food_item; /* Specifies food item as a clustering variable*/
model s(event='1')=/  link=logit;
domain food_item;
run
```

2.8. Impact of Summarizing Information

Another crucial issue is in summarizing data for analysis. For practical, and efficiency reasons one sample is tested for many pesticide residues, thus each sample has more than one observation. For analysis this is summarized by considering a sample as non-compliant when at least one pesticide residue is above MRL. This might lead to loss of information, and invalid inferences. Several other methods for summarizing data can be explored to come up with a reasonable summary analysis, for example, proportion of non-compliant samples can be computed in each sample and analyze this as a continuous outcome. We investigate the use of summarized and full data. For full data, the model in Section 2.7 was used, and we took into account correlation among pesticide residues coming from the same sample by considering a sample as a cluster. For summary analysis from continuous outcome, a linear regression model was fitted. Proportion of non-compliant samples in each sample served as the outcome and only the intercept was considered as an explanatory variable. Results are shown in Table 30. Differences between using full and summarized data are large for all food items. Different conclusions can be reached from each of the analyses; hence caution needs to be exercised when deciding to summarize information for analysis. Not that the aim of this section is just to illustrate the impact of summarizing information, we do not intend to advocate for a particular methodology. In Section 3.1 we discuss considerations that have to be made when deciding the type of analysis.

Table 30 Impact of Summarizing Information for Analysis.

Food Item	Information Used for Analysis					
	Summarized Sample Level		Not Summarized Determination Level		Summarized-continuous Sample-Determination Level	
	Prop.	Std. Error	Prop.	Std. Error	Prop.	Std.Error
Apples	0.01313	0.00251	0.00013	0.00002	0.00015	0.00003
Head cabbage	0.00901	0.00299	0.00009	0.00003	0.00010	0.00004
Leek	0.01041	0.00327	0.00011	0.00003	0.00015	0.00006
Lettuce	0.03380	0.00456	0.00035	0.00005	0.00035	0.00005
Oats	0.05285	0.01426	0.00052	0.00015	0.00092	0.00025
Peaches	0.01833	0.00387	0.00019	0.00004	0.00020	0.00005
Pears	0.01289	0.00573	0.01289	0.00573	0.01289	0.00573
Rye	0.00246	0.00246	0.00002	0.00002	0.00002	0.00002
Strawberries	0.02752	0.00459	0.00025	0.00004	0.00024	0.00005
Tomatoes	0.01226	0.00260	0.00012	0.00002	0.00011	0.00003
Overall	0.01619	0.00114	0.00018	0.00001	0.00058	0.00018

Estimated proportions and standard errors for analysis at the sample level differed by a hundredth to those estimated at the determination level (except for pears). This is because the estimation of the mean is the same for SRS and cluster sampling under design based analysis. Only standard errors are adjusted for analysis that accounts for clustering. At the sample level the sample size is reduced and this results into bigger proportions than at the determination level. Each of these methods may be valid depending on expert opinion. Consider a sample from a specific food item, if it is known that once residues above MRL have been detected for one pesticide then excess residues are likely to be detected for the other pesticides as well, then summary methods would be preferable. On the other hand, if detecting residues above MRL in one pesticide does not affect the results of the other pesticides then the analysis at the determination level would be preferred. Note that the sample-determination analysis is another form of a summarized method that can be considered as midpoint between sample level and determination level. The sample-determination level analysis keeps more information about the sample than the sample level

analysis. For example, samples with 3 and 10 detects will be given the same value in the sample level analysis but these would be given different values in the sample-determination level analysis. In choosing the analysis to use consideration should be given to meaningful interpretation of the results according to the subject matter.

2.9. Impact of Non-Response

Of preliminary importance while dealing with missing data is to explore both the magnitude and patterns of the missing data. We first explore the pattern related to member states not meeting the required number of samples for the specific food item. In this case, missingness refers to either providing some samples, but failing to meet the required number, or not providing any sample at all. This will be termed "View 1 of Missingness". Other alternative definitions will be explored as well.

In Table 31, we show the missingness pattern for the various food items, for the 2010 pesticides monitoring data. The crosses (X) indicate compliance with the required number of samples, while the dots (.) indicate non-compliance. The final column in the table represents the number of countries in the particular pattern.

We note that only 2 of the member states met the required number for each food item. Failure to meet the required number, for pears and rye and oats, was the most common phenomenon, at 5 member states. In total, there are 21 patterns in which member states provided the samples.

Table 31 Missingness Pattern per Food Item: View 1 of Missingness.

Pattern	Apples	Cabbage	Leek	Lettuce	Milk	Peaches	Pear	RyeOats	Straw b.	Wine	Tomat.	Freq.
1	X	X	X	X	X	X	X	X	X	X	X	2
2	X	X	X	X	X	X	.	X	X	X	X	1
3	X	X	X	X	X	X	.	.	X	X	X	5
4	X	X	X	X	X	X	.	.	X	.	X	2
5	X	X	X	X	X	X	.	.	.	X	X	1
6	X	X	X	X	X	.	.	X	.	X	X	1
7	X	X	X	X	.	X	X	X	X	X	X	1
8	X	X	X	X	.	X	X	.	X	X	X	1
9	X	X	X	X	.	X	.	X	X	.	X	1
10	X	X	X	X	.	X	.	.	X	X	X	2
11	X	X	X	X	.	X	.	.	X	.	X	2
12	X	X	X	X	.	X	.	.	.	X	X	1
13	X	X	X	.	X	.	.	.	X	X	X	1
14	X	X	X	.	.	.	.	X	X	.	.	1
15	X	.	X	X	.	X	X	X	X	.	X	1
16	X	.	.	X	X	X	.	.	X	X	X	1
17	X	.	.	X	.	X	.	.	X	X	X	1
18	X	.	.	X	.	.	.	.	.	.	X	1
19	X	.	.	.	.	.	.	.	.	.	X	1
20	.	.	.	.	.	X	.	X	X	.	X	1
21	.	.	.	.	.	.	.	.	.	.	X	1

An alternative view in this case is to consider missingness as purely failure to provide at least a sample, i.e. providing 0 samples. In Table 32, we provide the results related to this view. In this case, 4 of the member states conformed. The most common missingness patterns were related to pears (6 member states), and pears, ryes and oats (4 member states). In total, there were 14 different patterns.

Table 32 Missingness Pattern per Food Item: View 2 of Missingness.

Pattern	Apples	Cabbage	Leek	Lettuce	Milk	Peaches	Pear	RyeOats	Straw b.	Wine	Tomat.	Freq.
1	X	X	X	X	X	X	X	X	X	X	X	4
2	X	X	X	X	X	X	X	X	X	.	X	1
3	X	X	X	X	X	X	X	.	X	X	X	2
4	X	X	X	X	X	X	.	X	X	X	X	6
5	X	X	X	X	X	X	.	.	X	X	X	4
6	X	X	X	X	X	X	.	.	X	.	X	2
7	X	X	X	X	.	X	X	X	X	X	X	1
8	X	X	X	X	.	X	X	X	X	.	X	2
9	X	X	X	X	.	X	.	X	X	.	X	2
10	X	X	X	X	.	X	.	.	X	X	X	1
11	X	X	X	X	.	X	.	.	X	.	X	1
12	X	.	X	X	X	X	.	.	X	X	X	1
13	X	.	X	X	.	X	X	X	X	X	X	1
14	.	X	.	X	.	X	.	X	X	.	X	1

Another pattern worth exploring is non-response in terms of the pesticide residues. In this case, missingness relates to not providing information related to the required residues. We focus here on the 30 residues which were defined as mandatory for either both commodities of plant and animal origin, or only commodities of animal origin, in the 2010 pesticides monitoring programme. A list of these pesticides is provided in Table 33. Arbitrary labels have been added for convenience in further reference.

Table 33 Mandatory Pesticides for Either Both Animal and Plant Commodities, or Only Animal Commodities, in 2010

Residue name	Label	Residue name	Label
Bifenthrin	B1	Fenvalerate/Esfenvalerate (sum)	F4
Chlordane (sum of cis- and trans-isomers and oxychlordane expressed as chlordane)	C1	Heptachlor (sum of heptachlor and heptachlor epoxide expressed as heptachlor)	H1
Chlorpyrifos	C2	Hexachlorobenzene	H2
Chlorpyrifos-methyl	C3	Hexachlorocyclohexane (HCH), Alpha-isomer	H3
Cyfluthrin (Cyfluthrin including other mixtures of constituent isomers (sum of isomers))	C4	Hexachlorocyclohexane (HCH), Beta-isomer	H4
Cypermethrin (Cypermethrin including other mixtures of constituent isomers (sum of isomers))	C5	Lindane (Gamma-isomer of hexachlorociclohexane (HCH))	L1
DDT (sum of p,p'-DDT, o,p'-DDT, p-p'-DDE and p,p'-TDE (DDD) expressed as DDT)	D1	Methidathion	M1
Deltamethrin (cis-deltamethrin)	D2	Methoxychlor	M2
Diazinon	D3	Parathion	P1
Aldrin and Dieldrin (Aldrin and dieldrin combined expressed as dieldrin)	D4	Parathion-methyl (sum of Parathion-methyl and paraoxon-methyl expressed as Parathion-methyl)	P2

Endosulfan (sum of alpha- and beta-isomers and endosulfan-sulphate expresses as endosulfan)	E1	Permethrin (sum of isomers)	P3
Endrin	E2	Pirimiphos-methyl	P4
Fenthion (fenthion and its oxigen analogue, their sulfoxides and sulfone expressed as parent)	F1	Profenofos	P5
Fenvalerate and Esfenvalerate (Sum of RR and SS isomers)	F2	Pyrazophos	P6
Fenvalerate and Esfenvalerate (Sum of RS and SR isomers)	F3	Triazophos	T1

There were 27 patterns in this case. Only 1 member state provided information on all the pesticides, while the rest had information for at least one pesticide missing.

As mentioned earlier, we will discuss the methods to account for the missing data, in Section 3.6.

Table 34 Missingness Pattern for the 30 Pesticides Mandatory for Either Both Items of Animal and Plant Origin, or Items of Animal Origin. P and F represent pattern and frequency respectively.

P	B1	C1	C2	C3	C4	C5	D1	D2	D3	D4	E1	E2	F1	F2	F3	F4	H1	H2	H3	H4	L1	M1	M2	P1	P2	P3	P4	P5	P6	T1	F
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	1
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	.	X	X	X	X	X	X	X	X	X	X	1
3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	1
4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	X	X	X	X	.	X	X	X	X	X	.	X	1
5	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	.	.	X	X	X	X	X	X	X	X	X	X	2
6	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	.	.	X	X	X	X	X	X	X	X	.	X	1
7	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	.	.	.	X	X	X	X	X	X	X	X	X	1
8	X	X	X	X	X	X	X	X	X	X	X	X	X	X	.	.	X	X	X	X	.	X	X	X	X	X	X	X	X	X	1
9	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	X	X	X	X	X	X	X	.	X	X	X	X	X	.	X	1
10	X	X	X	X	X	X	.	X	X	.	X	X	.	.	.	X	.	X	X	X	X	X	X	X	X	X	X	X	X	X	1
11	X	X	X	X	.	X	X	X	X	X	X	X	X	X	X	.	X	X	.	.	X	X	X	X	X	X	X	X	X	X	1
12	X	X	X	X	.	X	X	X	X	X	X	X	X	.	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	1
13	X	X	X	X	.	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	1
14	X	.	X	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	X	X	X	X	X	X	X	X	X	X	.	X	1
15	X	.	X	X	X	X	X	X	X	X	X	X	X	X	.	X	X	X	.	.	X	X	X	X	X	X	X	X	X	X	1
16	X	.	X	X	X	X	X	X	X	X	X	X	X	.	.	.	X	X	X	.	X	X	X	X	X	X	X	X	X	X	1
17	X	.	X	X	X	X	X	X	X	.	X	X	.	X	.	.	X	X	X	X	X	X	X	X	X	.	X	X	X	X	1
18	X	.	X	X	X	X	.	X	X	.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	1
19	X	.	X	X	X	X	.	X	X	.	X	X	X	X	X	.	.	X	X	X	X	X	X	X	X	X	X	X	X	X	1
20	X	.	X	X	X	X	.	X	X	.	X	X	X	X	X	.	.	X	.	.	X	X	.	X	X	.	X	X	.	X	1
21	X	.	X	X	X	X	.	X	X	.	X	X	X	.	X	X	.	.	X	X	.	X	X	X	X	X	X	X	X	X	1
22	X	.	X	X	X	X	.	X	X	.	X	.	X	X	X	.	.	.	.	.	.	X	.	X	X	.	X	X	.	X	2

P	B1	C1	C2	C3	C4	C5	D1	D2	D3	D4	E1	E2	F1	F2	F3	F4	H1	H2	H3	H4	L1	M1	M2	P1	P2	P3	P4	P5	P6	T1	F
23	X	.	X	X	X	X	.	X	X	.	X	.	X	.	.	X	.	.	.	.	.	X	.	X	.	.	X	X	.	X	1
24	X	.	X	X	X	X	.	X	X	.	X	.	.	.	.	.	.	.	.	.	.	X	.	X	X	.	X	X	.	X	1
25	X	.	X	X	X	X	.	X	X	.	X	.	.	.	.	.	.	.	.	.	.	X	.	X	.	.	X	X	.	X	1
26	X	.	X	X	X	.	X	X	X	.	X	X	.	.	.	.	.	X	X	.	X	X	X	X	X	.	X	X	.	X	1
27	.	.	X	X	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	X	.	.	.	1

2.10. Impact of Left Censoring

We focus once again on the 30 pesticide residues introduced in Section 2.9. These pesticide residues covered a total of 178231 determinations, 99.13% and 0.12% of which were labeled LOQ and LOD, respectively. Only 0.75% of the determinations had residues measured above the LOQ. This information is provided in Table 35 below.

Therefore, the percentage of censoring was quite high for these data. Based on the results in Section 2.6, such a censoring percentage has potential to induce substantial bias in the estimation of the mean, depending on how the censored information is handled.

Table 35 Frequency of Result Type for the 30 Pesticide Residues

Result Type	Frequency	Percentage
LOD	217	0.12
LOQ	176901	99.13
VAL	1330	0.75
Total	178231	

In Table 36, we provide some summary information for the 1330 determinations measured above the LOQ. In Figure 21, we explore these values through a histogram. It is noteworthy that the values exhibit skewness.

Table 36 Summary Information for Values Measured above LOQ for the 30 Residues

Mean	Variance	Median	Mode	Maximum Value	Minimum Value	Skewness
0.0794	0.0434	0.0300	0.0200	4.1000	0.0001	10.4221

The above analysis is consolidated over all the food items covered by the 30 pesticides. We now look at the individual food items covered. In Table 37 below, we notice that for each of the products, only a "small" fraction is measured above the LOQ. In Figure 22, Figure 23 and Figure 24, we explore histograms of the individual food items. The values again exhibit skewness. Summary information for the individual items is presented in Table 38.

The logarithm of the measurements was taken, and summary information computed per food item. This information is provided in Table 39. The average of the mean and of the variance, was -3.8487 and 1.9528, respectively. These values, rounded off (to -4 and 2 respectively), were used as input parameters for μ and σ^2, respectively, to conduct a new investigation on the effect of censoring, as described in Section 2.6, but now using parameters derived from the data.

The results are provided in Table 40, as well as in Figure 25 and Figure 26. The positive bias is noticeable. Once again, it is clear that replacement of values below the "censoring value" with that particular "censoring value" has potential to induce substantial bias in the results. Therefore, replacement of values below the LOQ with the LOQ (as described to have been done for the EFSA 2010 report) can substantially affect the analyses.

As mentioned in Section 2.6, analysis methods to account for left censoring will be discussed in Section 3. It is important to note that substitution methods are still useful in some contexts, and there may be no need for sophisticated analysis methods in such cases. This is generally the case when the

results obtained from various substitution methods do not raise toxicological concerns. Further discussion is provided in EFSA, 2010[20].

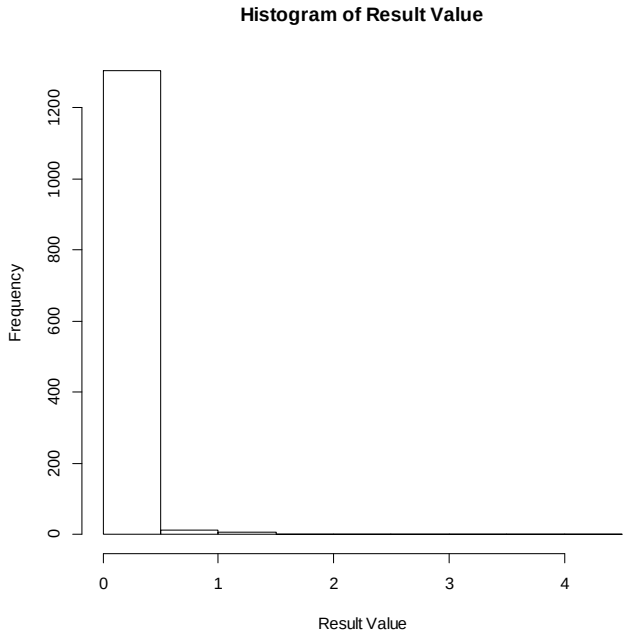

Figure 21 Values Measured Above LOQ, for 28 Pesticide Residues.

[20] European Food Safety Authority; Management of left-censored data in dietary exposure assessment of chemical substances. EFSA Journal 2010;8(3):.[96pp.]. doi: 10.2903/j.efsa.2010.1557.

Table 37 Result Value for Each of the Products Covered by the 30 Pesticides

Product	LOD	LOQ	VAL
Apples	49	29063	379
Peaches	39	17248	303
Straw Berries	49	18099	67
Tomatoes	40	24482	163
Head Cabbage	0	14631	29
Lettuce	20	22327	146
Leek	20	13939	23
Oats	0	3207	62
Rye	0	5372	50
Swine Meat	0	14252	32
Milk	0	14064	76

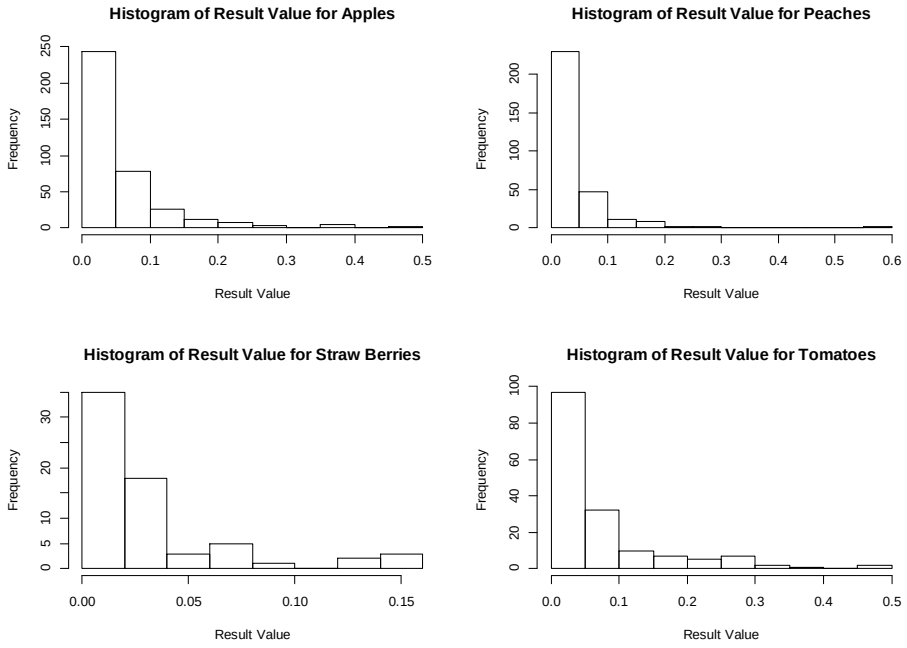

Figure 22 Individual Food Items Covered by the 28 Pesticides

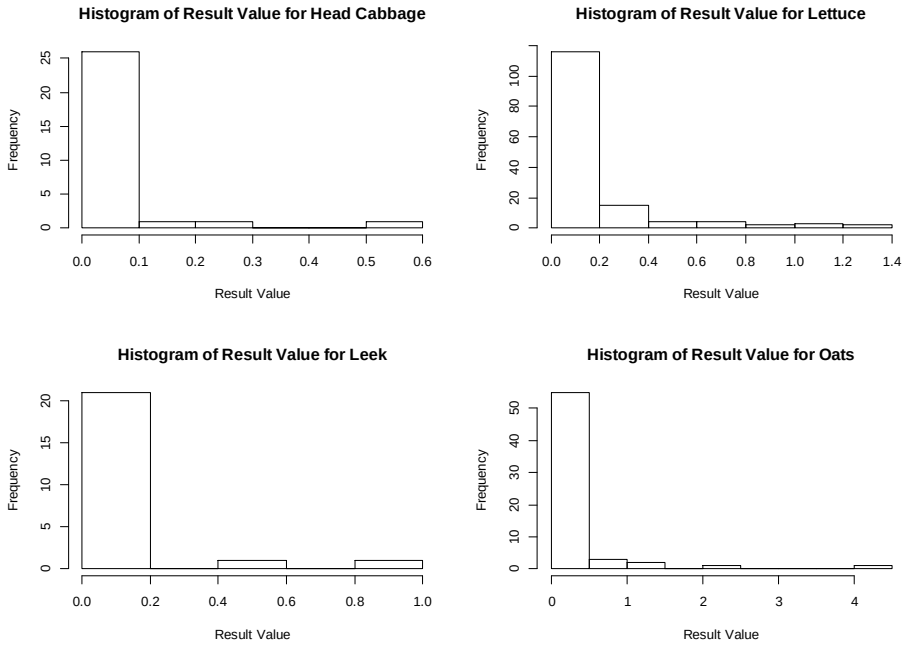

Figure 23 Individual Food Items Covered by the 28 Pesticides

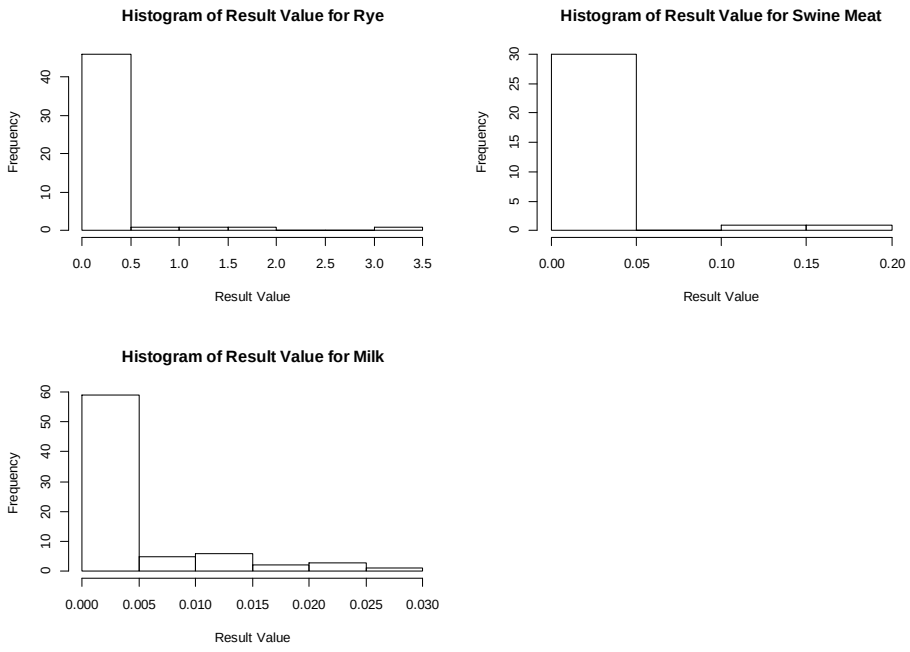

Figure 24 Individual Food Items Covered by the 28 Pesticides

Table 38 Food Item Summary Information for the 30 Pesticides

Product	Mean	Variance	Skewness[(a)]	Minimum	Maximum
Apples	0.0609	0.0052	2.9034	0.0030	0.4800
Peaches	0.0451	0.0045	5.0021	0.0010	0.6000
Straw Berries	0.0346	0.0014	2.0332	0.0020	0.1600
Tomatoes	0.0748	0.0084	2.2297	0.0020	0.5000
Head Cabbage	0.0638	0.0132	3.9274	0.0050	0.5900
Lettuce	0.1610	0.0652	2.8005	0.0030	1.3000
Leek	0.1060	0.0477	3.5763	0.0020	1.0000
Oats	0.2626	0.3733	4.7704	0.0060	4.1000
Rye	0.1886	0.2924	4.3357	0.0020	3.2000
Swine Meat	0.0140	0.0014	3.7253	0.0001	0.1723
Milk	0.0044	0.00004	2.0541	0.0001	0.0287

[(a)] Values close to zero suggest symmetry.

Table 39 Summary Information for the Logarithm of the Values

Product	Mean	Variance
Apples	-3.2793	0.9271
Peaches	-3.6651	1.1375
Straw Berries	-3.8426	1.0230
Tomatoes	-3.1928	1.2305
Head Cabbage	-3.4649	1.2580
Lettuce	-2.7020	1.7346
Leek	-3.4272	2.4288
Oats	-2.6894	2.5397
Rye	-3.4045	2.6728
Swine Meat	-6.0240	3.2506
Milk	-6.6434	3.2780

Table 40 Bias due to Left Censoring

True Mean	% Censoring	Estimated Mean	Bias	% Bias
0.0498	20	0.0503	0.0005	1.0998
	40	0.0526	0.0028	5.5982
	60	0.0595	0.0097	19.4996
	80	0.0841	0.0343	68.9056

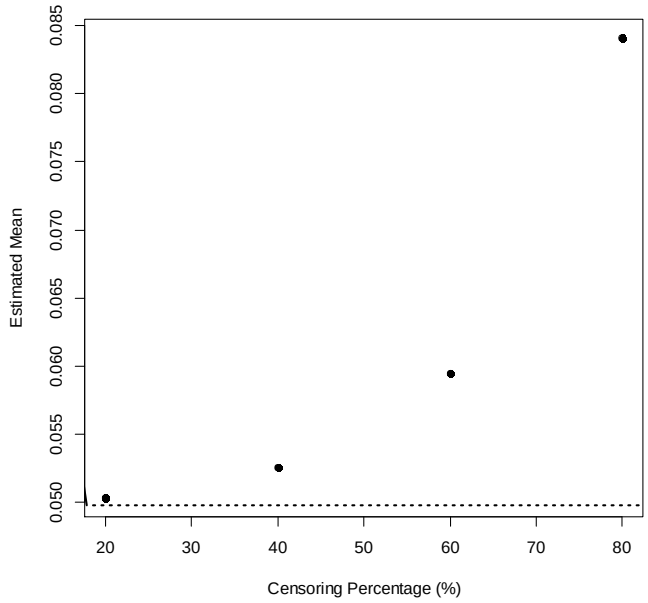

Figure 25: Estimated Mean for Various Censoring Percentages. Broken Line Shows the True Mean.

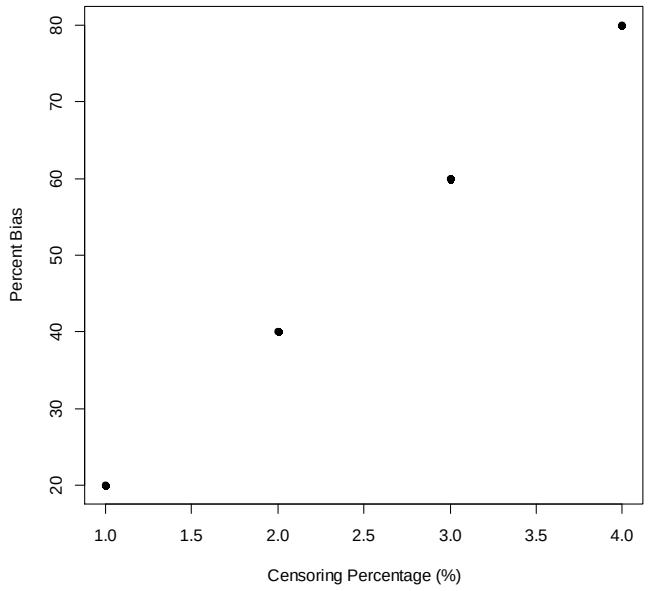

Figure 26 Percent Bias for Various Censoring Percentages.

3. Methods to Deal With the Types of Bias and Issues Identified in the Previous Objective That Could be Used to Propose Potential Corrections to Ensure Reduction or Elimination of Bias From the Inference Process

3.1. Methodology for Analysing Data When Sampling Design Details are Unavailable or Incomplete

In conducting a survey it is recommended to document all steps and decisions of the sampling design so as to have as much information as possible to be used in reflecting the design during analysis. However, for practical reasons this is not always possible, for example, while EFSA can control the selection of food items, the selection of samples of food items is the responsibility of participating member states hence it cannot be guaranteed that all information about the sampling design will be available at all stages.

With incomplete or no information on the sampling design it may be difficult to decide on the method of analysis and use of some methods that require details like selection probabilities may not be possible. In this scenario expert opinion on the possible behaviour of the survey quantities of interest in the population under study is crucial. In the pesticides monitoring study experts can shed light on what would make a meaningful cluster, e.g., samples from the same food item or samples from the same food item and country of origin. This information can then be incorporated into sensitivity analysis to decide whether complex analysis methods are necessary.

Sensitivity analysis basically involves conducting different statistical analysis on the same data under different assumptions to study the stability of the results under the assumed conditions. It is commonly used in situations where some pieces of information are missing and assumptions have to be made (Genelletti *et.al.*, 2011). In this case the missing information is the design hence sensitivity analysis would involve analyzing the same data assuming and incorporating different designs. In general, if all

methods of analysis produce different results it would be recommended to proceed with the most complex method and if they are similar, the simplest method would be preferred.

In the survey context simple and complex data result from SRS and multistage sampling, respectively. A typical sensitivity analysis should therefore at least consider these two designs and other intermediate designs deemed necessary.

To fix some ideas consider the simulation study in Section 2.1 where the sample was selected in a multistage sampling. Assume not all information on the design were available but we were well informed on the variables that could have been used for stratification or clustering. To determine the likely method of analysis the data is analyzed assuming SRS, stratified simple random sampling, cluster sampling and multistage sampling.

In anticipation of complex survey data collected in practice, we resort to model based methods in which tools for integrating complex designs are well developed than in the design based methods used in Section 2. Below is a brief description of the methods used to account for each of the designs.

3.1.1. Generalized Linear Models (GLMs): SRS and Stratified SRS

When the outcome of interest (response variable) in a survey is continuous and can be assumed to follow a normal distribution, linear models like multiple linear regressions are used to estimate parameters of interest. However, response variables can be sampled from distributions other than the normal distribution and examples include dichotomous response variable (sample above MRL or not) which is usually assumed to follow a binary or binomial distribution, and count data, assumed to follow a Poisson distribution. To estimate parameters of interest where such response variables are concerned, generalized linear models (GLMs) are used. In simplest terms generalized linear models can be regarded as extensions of ordinary linear regression models to encompass non-normal response distributions (Agresti, 2000). Specifically, GLMs cover response variables whose distributions are members of the exponential family distributions. Denote Y as the outcome of interest, $y_1, \dots, y_n$ the observed outcomes on n observations sampled from the population of interest, then the distribution of Y is said to be a member of the exponential family distribution if

$$f(y_k|\theta_k) = a(\theta_k)b(y_k)exp[y_k Q(\theta_k)], \qquad k = 1, \dots, n,$$

where θ_k is a function of all parameters of interest and can vary (Agresti, 2002). It can be easily shown that binomial distribution is a member of this family. Precisely, let Y be a binary response and represent the outcomes as 1 for a failure (e.g., residue above MRL) and 0 for a success (e.g., residue equal to or below MRL). Further, let

$$\mathrm{P}(Y = 1) = \bar{Y} \text{ and } \mathrm{P}(Y = 0) = 1 - \bar{Y}.$$

Now, let $y_1, \dots, y_n$ be the observed outcomes on n randomly sampled observations; it follows that

$$f(y_k|\bar{Y}) = (1 - \bar{Y})\exp\left(y_k \log\frac{\bar{Y}}{1 - \bar{Y}}\right).$$

In most practical cases $\bar{Y}$ is assumed to be a function of other variables, i.e., there exist

$$g(\bar{Y}) = \sum_h \beta_h x_{hk} = x'_{hk}\boldsymbol{\beta} = \eta_k \quad \Longrightarrow \bar{Y} = g^{-1}(\eta_k) \qquad k = 1, \ldots, n,$$

where x_{hk} is the value of the h^{th} variable on observation k and β_h is the corresponding estimated coefficient (usually parameter of interest) and $g^{-1}(\eta_k)$ is called the link function. For purposes of illustration, we use the logit link in which

$$g(\bar{Y}) = \log\frac{\bar{Y}}{1-\bar{Y}} \quad \text{hence} \quad \bar{Y} = \frac{\exp(\eta_k)}{1+\exp(\eta_k)}.$$

Estimation of the parameters follows by minimizing the log likelihood given by

$$L(\bar{Y}) = n\log(1-\bar{Y}) + \sum_{i=1}^{n} y_i \log\frac{\bar{Y}}{1-\bar{Y}},$$

and the corresponding estimate for $\bar{Y}$ will be denoted as $\bar{y}$. More details on generalized linear models can be found in Agresti (2005); Aerts, et.al (2004); Molenberghs and Verbeke (2005).

In line with the theory above, the model assuming SRS design was defined as:

$$\bar{Y} = \frac{\exp(\beta_o)}{1+\exp(\beta_0)} \text{ or equivalently } \log\frac{\bar{Y}}{1-\bar{Y}} = \beta_0\,,$$

Model 1

And the mean was estimated as

$$\bar{y} = \frac{\exp(\hat{\beta}_o)}{1+\exp(\hat{\beta}_0)}.$$

(1)

The model assuming stratified SRS was defined as

$$\bar{Y} = \frac{\exp\left(\sum_{h=1}^{H}\beta_h x_{hk}\right)}{1+\exp\left(\sum_{h=1}^{H}\beta_h x_{hk}\right)} \text{ or equivalently } \log\frac{\bar{Y}}{1-\bar{Y}} = \sum_{h=1}^{H}\beta_h x_{hk}\,,$$

Model 2

where H is the total number of strata, $x_{hk} = 1$ if k^{th} element comes from h^{th} stratum and zero otherwise, further

$$\bar{y}_{\mathrm{h}} = \frac{\exp(\hat{\beta}_h)}{1+\exp(\hat{\beta}_h)},$$

is the estimated mean for stratum h. The overall mean follows as;

$$\bar{y}_s = \sum_{h=1}^{H} W_h \bar{y}_h.$$

(2)

where W_h is the allocation weight for the h^{th} stratum.

Variance estimates for (1) and (2) can be obtained through the delta method and they are given by the following expressions:

$$\mathrm{Var}(\bar{y}) = \bar{y}(1-\bar{y})\mathrm{Var}(\hat{\beta}_0),$$

$$\mathrm{Var}(\bar{y}_s) = \sum_{h=1}^{H} W_h^2 \mathrm{Var}(\bar{y}_h) \quad \text{where} \quad \mathrm{Var}(\bar{y}_h) = \bar{y}_h(1-\bar{y}_h)\mathrm{Var}(\hat{\beta}_h),$$

3.1.2. Generalized Linear Mixed Model: Cluster and Multistage Sampling

In a similar spirit to GLMs, generalized linear mixed models (GLMMs) extend linear mixed models to non-normal distribution responses. The main distinguishing feature between GLMs and GLMMs being that the latter recognizes that observations coming from the same cluster can be correlated while the former assumes that all observations are independent and identically distributed. GLMMs account for correlations by introducing cluster specific means in form of random effects. This implicitly assumes that the clusters included in the survey are just a random sample from the population of clusters such that if a different sample is taken, a new set of clusters may be selected. It is further assumed that conditional on the random effects observations within a cluster are independent and follow a distribution belonging to the exponential family. Extending the notation in 3.1.1, denote Y_{ck} as the k^{th} outcome measured for cluster c, $c = 1, \dots, C$ and $k = 1, \dots, n_c$ and $\boldsymbol{Y_c}$ as the n_c-dimensional vector of all measurements available for cluster c. Is assumed that conditionally on random effects $\boldsymbol{b_c}$ drawn independently from $N(\boldsymbol{0}, \boldsymbol{D})$, the outcomes Y_{ck} are independent with densities of the form

$$f(y_{ck}|\boldsymbol{b_i}, \boldsymbol{\theta}) = a(\theta_{ck})b(y_{ck})exp[y_{ck}Q(\theta_{kc})], \qquad k = 1, \dots, n_c.$$

Similarly we can define

$$g(\mu_{ck}) = g[E(Y_{ck}|\boldsymbol{b}_c)] = \boldsymbol{x}'_{ck}\boldsymbol{\beta} + \boldsymbol{z}'_{ck}\boldsymbol{b_c} = \eta(.)$$

Where $g^{-1}[\eta(.)]$ is a known the link function, $\boldsymbol{x_{ck}}$ and $\boldsymbol{z_{ck}}$ are vectors of known covariates and $\boldsymbol{\beta}$ is a vector of unknown fixed regression coefficients.

Further, the likelihood contribution of cluster c is given by

$$f(y_c|\boldsymbol{b_c}, \boldsymbol{\beta}, \boldsymbol{D}) = \int \prod_{k=1}^{n_c} f(y_{ck}|\boldsymbol{b_c}, \boldsymbol{\beta})\, f(\boldsymbol{b_c}|\boldsymbol{D}) d\boldsymbol{b_c},$$

3

With $f(\boldsymbol{b_c}|\boldsymbol{D})$ as the density of the $N(\boldsymbol{0}, \boldsymbol{D})$ distribution for the random effects $\boldsymbol{b_c}$.

Parameters of interest are estimated by maximizing the likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{D}) = \prod_{k=1}^{n} f(y_c|\boldsymbol{b_i}, \boldsymbol{\beta}, \boldsymbol{D}).$$

For elaborate details on GLMMs were refer interested readers to Fitzmaurice *et.al.(*2004) and Molenberghs and Verbeke (2005).

The models considered for both clustering and multistage without stratification are

$$P(y_{ck} = 1|b_c) = \frac{\exp(\beta_0 + b_c)}{1 + \exp(\beta_0 + b_c)}.$$

Model 3

and the stratified multistage model was

$$P(y_{hck} = 1|b_c) = \frac{\exp(\sum_{h=1}^{H} \beta_h x_{hck} + b_c)}{1 + \exp(\sum_{h=1}^{H} \beta_h x_{hck} + b_c)}.$$

Model 4

Where β_h is as defined before and $x_{hck} = 1$ for the k^{th} observation from cluster c and stratum h and zero otherwise. It should be noted that parameters estimated from Model 3 and Model 4 have different interpretation than those estimated from Model 1 and Model 2. The former have conditional interpretation while the latter have marginal interpretation. For example, in the realm of the pesticides monitoring study, $\bar{y}$ from Model 1 estimates the overall mean of the population, we shall refer to such quantities as marginal parameters. On the other hand $P(y_{ck} = 1|b_c)$ estimates the mean of the population conditionally on the value of b_c, and these shall be referred to as conditional parameters. While in most cases marginal parameters are of interest GLMMs offer a lot of insight in survey analysis, for instance in the pesticides monitoring study where food items are considered as clusters, clusters with high b_c would signal food items with high exceedance rate. Further, GLMMs provide the means to account for more than two levels of multistage sampling, this is not the case for models that provide marginal effects. Importantly, they provide an estimate of intra-class correlation by estimating the variance of the random effects. Additionally, for a random intercept model with logit link, approximate marginal parameters can be obtained from conditional parameters as

$$\beta^M \approx \beta^C \sqrt{\frac{1}{1 + 0.3\tau^2}},$$

Rabe-Hesketh and Skrondal (2006) where τ^2 is the variance estimate of the random effects. Marginal parameters can also be derived from GLMMs by integrating out the random effects. Variance estimates will be obtained using delta method as in the previous section. Specifically, the marginal overall and strata specific mean can be obtained as

$$P(y_k = 1) = \bar{y} = \int \frac{\exp(\hat{\beta}_0 + b_c)}{1 + \exp(\hat{\beta}_0 + b_c)} \phi(b_c|0, \hat{\tau}^2)\, db_c, \text{ and}$$

$$P(y_{hk} = 1) = \bar{y}_h = \int \frac{\exp(\hat{\beta}_h + b_c)}{1 + \exp(\hat{\beta}_h + b_c)} \phi(b_c|0, \hat{\tau}^2)\, db_c.$$

(4)

respectively, where $\phi(b_c|0, \hat{\tau}^2)$ is the normal distribution with mean zero and the estimated variance of the random effects $\hat{\tau}^2$. The corresponding variance estimates will be obtained as;

$$Var(\bar{y}) = \int \frac{\left(\exp(\hat{\beta}_0 + b_c)\right)^2}{\left(1 + \exp(\hat{\beta}_0 + b_c)\right)^4} \phi(b_c|0, \hat{\tau}^2)\, db_c, \text{ and}$$

$$Var(\bar{y}_h) = \int \frac{\left(\exp(\hat{\beta}_h + b_c)\right)^2}{\left(1 + \exp(\hat{\beta}_h + b_c)\right)^4} \phi(b_c|0, \hat{\tau}^2)\, db_c.$$

3.1.3. Generalized Estimating Equations: Cluster Sampling

Methods in Section 3.1.1 fail to account for correlated outcomes within a cluster and those in Section 3.1.2 do not directly provide the commonly required marginal parameters. Generalized estimating equations (GEE) (Liang and Zeger, 1986), offer an alternative method to analysing a two-level clustering sample. The main desirable features include; accounting for correlation within clusters, computational simplicity and providing parameters with marginal interpretation. For GEE we are only required to correctly specify the marginal mean which using the notation introduced in Section 3.1.2 can be, $g(\mu_{ck}) = \boldsymbol{x}'_{ck}\boldsymbol{\beta}$ and a working correlation, i.e., the assumed correlation structure for the outcomes of observations within the same cluster. Even with a wrongly specified working correlation estimates for $\boldsymbol{\beta}$ are still consistent and asymptotically normally distributed. Note that correlation parameters do not have meaningful interpretation. Estimation is done through an iterative procedure which together with detailed theory on GEE can be found in Liang and Zeger (1986); Molenberghs and Verbeke (2005) and Agresti (2002).

The models considered for both clustering and multistage are without stratification

$$P(y_{ck} = 1) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}.$$

Model 5

and the stratified multistage model was

$$P(y_{hck} = 1) = \frac{\exp(\sum_{h=1}^{H} \beta_h x_{hck})}{1 + \exp(\sum_{h=1}^{H} \beta_h x_{hck})}.$$

Model 6

Variance estimates follow directly from the variance covariance matrix adjusted for correlation. Exchangeable working correlation which assumes constant correlation among members of the same cluster was adopted for all models.

3.1.4. Sensitivity Analysis: Results and Decision Making

The data used in the sensitivity analysis was generated in a similar manner to that in Section 2.1.1.1, with a slight change in the generation of cluster means. For this exercise cluster means were obtained as

$$\bar{Y}_c = \frac{\exp(c)}{1+\exp(c)} \quad \text{where } c \sim N(0,2.25),$$

thus the variability of the cluster means was increased from 0.25 to 2.25. The increase was necessary to enable fitting of GLMMs which require the variance of random effect to be positive. The small variability implied that the within cluster variance was larger than the between cluster variance and this may result into negative random effects variance estimate. Otherwise all other procedures were similar including the sample selection.

A sensitivity analysis was conducted on the assumption that we did not have all information on the sampling design. Specifically, we do not have all details on how the final sample elements were selected.

Table 41 Results From Non-Stratified Sensitivity Analysis to Determine Aspects of the Design to be Integrated in the Analysis.

Type	Model	$\bar{Y}$	$\bar{y}$	Var($\bar{y}$)	R.bias	$\hat{a}$	$\hat{d}$
Model	GLM	0.030	0.031	0.000005	0.046	0.470	0.004
	GEE	0.030	0.031	0.000046	0.046	0.065	0.013
	GLMM	0.030	0.032	0.000154	0.085	0.000	0.024
Design	Cluster	0.030	0.030	0.000045	0.024	0.065	0.013
	Multistage	0.030	0.030	0.000043	0.024	0.065	0.013
	SRS	0.030	0.030	0.000004	0.024	0.470	0.004
	Stratified SRS	0.030	0.030	0.000004	0.024	0.485	0.004

Table 41 and Table 42 give results for non-stratified and stratified analyses, respectively. The statistical software programming codes for obtaining these results are provided at the end of the section. A comparison of the results from design-based and model-based methods indicate that results from GEE and GLMMs are similar to results from cluster and multistage, and both sets of models account for clustering. Similarly, results from GLM mirrors results from SRS since both methods assume independence of elements in the sample. Further, from Table 42 we note that GLM and GEE results only differ in the Var($\bar{y}$) estimates, while results from are different from the other two models in both mean and variance estimates. In essence GEE accounts for correlation like a design model that takes clustering into account. Thus GEE makes the assumption that at the first stage of sample selection, clusters were sampled and all the elements from the selected clusters were included in the sample. On the other hand GLMM recognizes the hierarchy in sample selection, i.e., after the clusters

are selected at the first stage, the second stage selects elements (less than the number of elements in the cluster) from the selected clusters. In other words, GEE assumes that there is no sampling variability at the second stage while GLMM recognizes the sampling variability. This can be seen in the larger standard error estimates for GLMM than GEE. In this case GLMM has an advantage over design based methods. In general, this shows that model and design based methods lead to similar results.

In practice true values are not available hence decisions would be based on the results from the sensitivity analysis. Results from the design based models were included just for comparison, from now and the rest of the report discussion will only be based on the model based methods.

To determine whether clustering exists at the specified level the variance estimates have to be examined. When the sample observations are independent all models should give similar estimates for variance of the mean. In this case intra-class correlation will approximately be zero and the design effect will approximately be one for designs taking into account clustering. Recall that the design effect determines how much the variance under SRS is inflated when a different sampling design is used.

Considerable differences in variance estimates indicate existence of clustering in the sample and it should be accounted for. Variance of the mean estimates for models accounting for clustering (GEE and GLMM) are at least 9 times larger than the estimates from GLM, a clear indication that clustering should be taken into account. Further, variability for GLMM is larger than that of GEE suggesting that multistage sampling was used. If more than two levels of hierarchy are expected, a model accounting for all the suspected levels should be considered and appropriate statistical tests can be used to determine the levels that need to be taken into account. Again if all stages are important i.e., the variability at each stage is reasonably large, standard error estimates for fixed effects for models taking into account lesser number of levels will be smaller than those taking into account higher number of levels, Van den Noortgate *et.al.*(2004). Generally if variance estimates for all models in the sensitivity analysis are considerable different, then it is recommended to use the most complex model.

Table 42 Results From Stratified Sensitivity Analysis to Determine Aspects of the Design to be Integrated in the Analysis.

Stratum	$\bar{Y}$	GLM		GEE		GLMM	
		$\bar{y}$	Var($\bar{y}$)	$\bar{y}$	Var($\bar{y}$)	$\bar{y}$	Var($\bar{y}$)
1	0.0844	0.0958	0.000473	0.0958	0.004251	0.0764	0.004428
2	0.0004	0.0003	0.000001	0.0003	0.000001	0.0007	0.000006
3	0.0283	0.0185	0.000054	0.0185	0.000182	0.0175	0.000575
4	0.0818	0.0751	0.000362	0.0751	0.001662	0.0678	0.002541
5	0.0731	0.0674	0.000481	0.0674	0.001522	0.0935	0.004901
6	0.0307	0.0381	0.000174	0.0381	0.001401	0.0374	0.001942
7	0.0066	0.0069	0.000038	0.0069	0.000045	0.0090	0.000315
8	0.0003	0.0002	0.000001	0.0002	0.000001	0.0002	0.000001
9	0.0580	0.0483	0.000216	0.0483	0.000614	0.0639	0.002461
10	0.0075	0.0099	0.000032	0.0099	0.000056	0.0107	0.000231
11	0.1222	0.0964	0.000619	0.0964	0.001832	0.0969	0.005158
12	0.0395	0.0381	0.000282	0.0381	0.000386	0.0384	0.001744
13	0.0140	0.0248	0.000037	0.0248	0.000207	0.0347	0.000644
14	0.0002	0.0004	0.000003	0.0004	0.000001	0.0005	0.000004
15	0.0063	0.0074	0.000025	0.0074	0.000038	0.0081	0.000199
16	0.0001	0.0002	0.000001	0.0002	0.000001	0.0005	0.000033

17	0.0165	0.0162	0.000092	0.0162	0.000152	0.0181	0.000731
18	0.0232	0.0256	0.000155	0.0256	0.000240	0.0230	0.000853
19	0.0380	0.0438	0.000163	0.0438	0.000556	0.0464	0.001615
20	0.0001	0.0001	0.000001	0.0001	0.000001	0.0003	0.000008
21	0.1027	0.1083	0.000709	0.1083	0.002874	0.1130	0.004468
22	0.0290	0.0416	0.000181	0.0416	0.001112	0.0407	0.002094
23	0.0104	0.0158	0.000069	0.0158	0.000207	0.0167	0.000515
24	0.0260	0.0381	0.000203	0.0381	0.000488	0.0383	0.001423
25	0.0520	0.0539	0.000188	0.0539	0.001582	0.0522	0.002208
26	0.0143	0.0158	0.000069	0.0158	0.000102	0.0200	0.000661
27	0.0308	0.0352	0.000134	0.0352	0.000323	0.0376	0.001165
28	0.0205	0.0322	0.000183	0.0322	0.000518	0.0381	0.002074
29	0.0697	0.0610	0.000125	0.0610	0.000328	0.0672	0.001490

Following the decision to account for clustering or not is the issue of stratification. In most cases this would not be a big issue because stratification can also be done after selecting a sample even if the sampling design was not stratified, this is referred to as post-stratification. Stratification is usually done to improve precision; hence, examining variance estimates from stratified and non-stratified analysis for an improvement in precision should help decide if stratified analysis is necessary. For some samples stratification can solve bias issues especially when the means for the different strata vary greatly. In such settings mean estimates for stratified and non-stratified analysis can be different. A weighted mean would then give a better estimation than un-weighted mean. However, caution should be exercised in using stratification to correct for bias because it is not always the case that different mean estimates are as a result of lack of stratification; this can be due to other sources of bias. Expert opinion should back up the use of stratification in this case.

The weighted means and their corresponding (variances) were 0.031 (0.00002) and 0.033(0.00002) for both GEE and GLMM, respectively. There is an improvement in precision in both cases, with the GLMM recording the highest improvement. These results suggest that some gains in precision were made due to stratification, especially for GLMM, hence methods accounting for a multistage stratified design would be recommended.

Note that the same sensitivity analysis can be used in deciding the appropriateness of methods that replaces the repeated outcomes with summary statistics (e.g., mean) to create a sample with independent observations. In general, it is recommended to use methods that make use of all information and summary methods often lead loss of information. Once the sensitivity analysis results indicate the need for accounting for clustering, summary methods should be avoided unless it has been shown that the correlation within the cluster is so strong to the extent that information from one element represents information for the whole cluster.

In summary Figure 27 presents the decision making process when deciding aspects of the design that have to be taken into account.

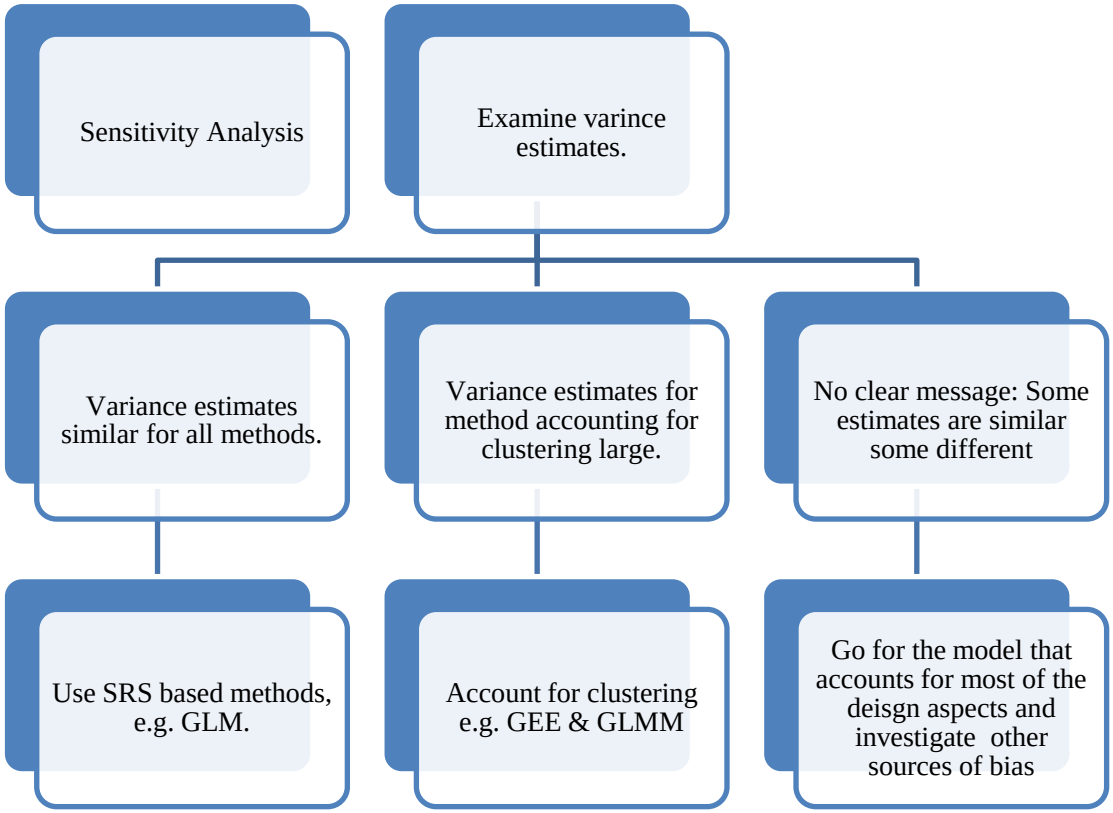

Figure 27 Decision Making Tree in Determining Aspects of Design to be Taken Into Account.

Programming codes for obtaining results in Table 41 using SAS 9.4 (Comments are provided in green)

```
/*GLM Non-Stratified*/

proc glimmix data=final empirical=classical;
/*data specifies the name of a dataset*/
by replicate;
/*by fits the model for each simulation run*/
  model resp (event='1')= / dist=binary link=logit  solution;
/*model specifies the model in this case resp is the response variable and there are no independent
variables apart from the intercept. The link  is logit*/
ods output ParameterEstimates=parms_glm01;
/*Saving the estimates to the dataset parms_glm01*/
run;


/*GEE Non-Stratified analysis*/

proc glimmix data=final empirical=classical;
by replicate;
  class cid;
  model resp (event='1')= / dist=binary link=logit  solution ;
  random residual / subject=cid ;
/*specifies that there is clustering within cid (the clustering variable)*/
ods output ParameterEstimates=parms_gee01;
run;


/*GLMM Non-Stratified*/

proc glimmix data=final  empirical=classical method=quad(qpoints=20);
by replicate;
  class cid;
  NLOPTIONS  tech=trureg lsp=0.0001 maxiter=5000; /*Improving convergence*/
  model resp (event='1')= / dist=binary link=logit  solution ;
  random intercept / subject=cid ;
/*specifies that there is hierarchical clustering within cid (the clustering variable)*/
ods output ParameterEstimates=parms_glimmix CovParms=covest_glimmix;
run;
```

Programming codes for obtaining results in Table 42 using SAS 9.4 (Comments are provided in green)

```
/*GLM Stratified*/

proc glimmix data=final empirical=classical;
/*data specifies the name of a dataset*/
by replicate;
class sid;
/*by fits the model for each simulation run*/
  model resp (event='1')= sid/ dist=binary link=logit  solution;
/*model specifies the model in this case resp is the response variable and sid the strata variable
is specified as the independent variables. The link  is logit*/
ods output ParameterEstimates=parms_glm01;
/*Saving the estimates to the dataset parms_glm01*/
run;


/*GEE Stratified analysis*/

proc glimmix data=final empirical=classical;
by replicate;
  class sid cid;
  model resp (event='1')=sid / dist=binary link=logit  solution ;
  random residual / subject=cid ;
/*specifies that there is clustering within cid (the clustering variable)*/
ods output ParameterEstimates=parms_gee01;
run;


/*GLMM Stratified*/

proc glimmix data=final  empirical=classical method=quad(qpoints=20);
by replicate;
  class sid cid;
  NLOPTIONS  tech=trureg lsp=0.0001 maxiter=5000; /*Improving convergence*/
  model resp (event='1')= sid/ dist=binary link=logit  solution ;
  random intercept / subject=cid ;
/*specifies that there is hierarchical clustering within cid (the clustering variable)*/
ods output ParameterEstimates=parms_glimmix CovParms=covest_glimmix;
run;
```

3.2. Methods for Analysis of a Sample Selected With Unequal Selection Probabilities Independent of the Outcome.

As noted in Section 2.2, when elements are selected into the sample with unequal selection probabilities, biased estimates can result even when the selection probabilities do not depend on the outcome. It was further illustrated that bias can be minimized by incorporating the inverse of selection probabilities in the analysis as weights. Basically the weights create a pseudo-population by replicating each element by its weight to make the sample representative of the population of interest. Obviously, information on selection probabilities has to be available to implement such methods. In this section we explore the use of model based methods which are likely to be used due to the complex nature of practical surveys.

Data simulated under Section 2.2.1 and the sample selected with unequal probabilities that do not depend on the outcome will be used to illustrate the performance of several models with specific interest on bias. Only GLMMs will be considered for this exercise. Theory for GLMMs 3.1.2 is still applicable with some slight changes introduced to accommodate the weights as briefly discussed in the section below.

3.2.1. Weighted Generalized Linear Mixed Model.

In a similar spirit to replication of each element by it inverse weight, the weighted GLMM replicates the likelihood contribution of each cluster by its weight. Specifically Expression (3) becomes

$$f_w(y_c|\boldsymbol{b_c},\boldsymbol{\beta},\boldsymbol{D}) = \int \prod_{k=1}^{n_c} W_{k|c} f(y_{ck}|\boldsymbol{b_c},\boldsymbol{\beta})\, f(\boldsymbol{b_c}|\boldsymbol{D}) d\boldsymbol{b_c},$$

where $W_{k|c}$ is the inverse selection probability of k^{th} element conditional on the c^{th} cluster being selected at the previous stage. The marginal likelihood can further be weighted by W_c , the inverse selection probability for cluster c, such that the marginal pseudo-likelihood becomes;

$$L(\boldsymbol{\beta},\boldsymbol{D}) = \prod_{k=1}^{n} W_c\, f_w(y_c|\boldsymbol{b_i},\boldsymbol{\beta},\boldsymbol{D}).$$

More details on the use and implementation of this method can be found in Rabe-Hesketh and Skrondal (2006). These authors also recommend the use of scaled values of $W_{k|c}$ to minimize bias for the variance components. The two commonly used scaling schemes and used in our analysis include;

$$W_{k|c}^{(1)} = \frac{\sum_{k=1}^{n_c}(W_{k|c})}{\sum_{k=1}^{n_c}(W_{k|c})^2} W_{k|c} \quad \text{and} \quad W_{k|c}^{(2)} = \frac{\mathrm{n_c}}{\sum_{k=1}^{n_c}(W_{k|c})} W_{k|c}.$$

3.2.2. Results: Analysis of Data With Unequal Non-informative Selection probabilities.

In this section we explore model based methods that can be used in analysing data from a sample selected with unequal selection probabilities. Models considered include GEE, weighted GLMM and un-weighted GLMM. In all cases both stratified and non-stratified analyses were implemented. The

performance of each model will be estimated by how well it estimates the marginal mean (for non-stratified analysis) and strata specific means (stratified analysis). For GLMM models, the strata specific means will be estimated as:

$$\bar{y}_h = \int \frac{\exp(\hat{\beta}_h + b_c)}{1 + \exp(\hat{\beta}_h + b_c)} \phi(b_c|0, \hat{\tau}^2)\, db_c,$$

where $\phi(b_c|0, \hat{\tau}^2)$ is the normal distribution with mean zero and the estimated variance of the random effects $\hat{\tau}^2$. Results for non-stratified and stratified models are presented in Table 43 and Table 44 respectively. The programming codes are provided at the end of the section.

Table 43 Results from non-stratified analysis exploring methods for analyzing sample data selected with non-informative unequal selection probabilities.

Scaling Scheme	$\bar{Y}$	$\bar{y}$	Var($\bar{y}$)	R.Bias	$\hat{\alpha}$	$\hat{d}$	τ^2
$W_{k\|c} = 1$	0.047	0.017	0.000010	-0.629	1.000	0.007	0.429
$W_{k\|c}^{(1)}$	0.047	0.015	0.000006	-0.681	1.000	0.005	0.100
$W_{k\|c}^{(2)}$	0.047	0.034	0.000138	-0.278	0.182	0.023	0.394
$W_{k\|c}$	0.047	0.038	0.001034	-0.192	0.000	0.063	10.918

The results for the analysis where the weights were scaled to $W_{k|c}^{(1)}$ and a non-weighted analysis ($W_{k|c} = 1$) are similar and highly biased compared to results where the raw weights ($W_{k|c}$) and scaled weights $W_{k|c}^{(2)}$, were used. Rabe-Hesketh and Skrondal (2006) indicated that the scaled weights $W_{k|c}^{(2)}$ perform well in cases where the selection probabilities are informative, i.e., selection probabilities are related to the outcome. Recall that simulation of population from which the sample used to produce results in Table 43 was selected was such that clusters with bigger sizes have higher means. This might explain why the second scaling scheme, which is a function of cluster size performs better than the first scheme. This also insinuates that despite having non-informative unequal selection probabilities, there might still be some bias in the sample intrinsic to the population simulation. It was however shown in Rabe-Hesketh and Skrondal (2006) even first scaling scheme minimizes bias under purely non-informative selection probabilities. On the other hand, when raw weights are used the bias in the estimate of random effects variance is high as expected. Results from the stratified analysis in Table 44 convey a similar message where the second scaling scheme ($W_{k|c}^{(2)}$) performs better than the rest except for strata with very small mean in which case the GLMM with raw weights performs better.

Table 44 Results From Stratified Analysis Exploring Methods for Analyzing Sample Data Selected With Non-Informative Unequal Selection Probabilities.

Weighting scheme	$\bar{Y}$	$W_{k\|c} = 1$	$W_{k\|c}^{(1)}$	$W_{k\|c}^{(2)}$	$W_{k\|c}$
Stratum			$\bar{y}$		
1	0.0733	0.0373	0.0362	0.0748	0.0955
2	0.0002	0.0000	0.0000	0.0000	0.0003
3	0.0099	0.0053	0.0054	0.0064	0.0271
4	0.1061	0.0518	0.0514	0.0687	0.1192
5	0.0726	0.0375	0.0372	0.0486	0.0940
6	0.0314	0.0165	0.0160	0.0305	0.0528
7	0.0095	0.0039	0.0039	0.0109	0.0204
8	0.0002	0.0004	0.0004	0.0000	0.0004
9	0.0469	0.0221	0.0223	0.0423	0.0686
10	0.0074	0.0041	0.0043	0.0026	0.0134
11	0.0895	0.0444	0.0434	0.0547	0.1151
12	0.0347	0.0164	0.0165	0.0251	0.0495
13	0.0590	0.0282	0.0272	0.0496	0.0761
14	0.0003	0.0002	0.0002	0.0000	0.0007
15	0.0083	0.0038	0.0039	0.0012	0.0160
16	0.0003	0.0002	0.0001	0.0000	0.0007
17	0.0121	0.0048	0.0049	0.0053	0.0198
18	0.0181	0.0071	0.0069	0.0047	0.0279
19	0.0532	0.0227	0.0227	0.0354	0.0611
20	0.0003	0.0002	0.0001	0.0000	0.0004
21	0.1240	0.0521	0.0514	0.1001	0.1161
22	0.0478	0.0197	0.0203	0.0244	0.0584
23	0.0153	0.0058	0.0060	0.0114	0.0270
24	0.0393	0.0163	0.0161	0.0457	0.0546
25	0.0769	0.0300	0.0298	0.0683	0.0816
26	0.0302	0.0097	0.0099	0.0149	0.0345
27	0.0542	0.0178	0.0177	0.0287	0.0582
28	0.0692	0.0195	0.0195	0.0516	0.0586
29	0.1207	0.0213	0.0220	0.0674	0.0622
τ^2	0.25	0.1902	0.1961	0.1789	14.4047

Programming codes for obtaining results in Table 43 using SAS 9.4 (Comments are provided in green)

```
/*GLMM Unweighted*/

proc glimmix data=final  empirical=classical method=quad(qpoints=50);
NLOPTIONS  tech=trureg lsp=0.001 maxiter=5000;
by replicate;
  class sid cid;
  model resp (event='1')= / dist=binary link=logit  solution CL;
  random intercept / subject=cid ;
ods output ParameterEstimates=parms_glimmix_mar CovParms=covest_glimmix_mar;
run;



/*Weighted Glmm*/

proc glimmix data=final  empirical=classical method=quad(qpoints=20);
by replicate;
  class sid cid;
    NLOPTIONS  tech=trureg lsp=0.001 maxiter=5000;
  model resp (event='1')= / dist=binary link=logit  solution  obsweight=wji;
/* obsweight  specifies the inverse probability weights for the elements at the lowest level of hierarchy e.g., the inverse of probability of selecting an element from a cluster given that the cluster was selected at the previous stage. These can be raw weights or scaled according to the various methods*/
    random int / subject=cid weight=samp_wt;
/* weight  specifies the inverse probability weights for sampling units at a higher level of hierarchy e.g., the inverse of probability of selecting a  cluster.*/

ods output ParameterEstimates=parms_wji CovParms=covest_wji;
run;
```

Programming codes for obtaining results in Table 44 using SAS 9.4

To obtain results in Table 44 the programming codes used to obtain results in Table 43 were slightly modified by including a strata variable (sid) as an independent variable in the *model* statement.

3.3. Methods for Analysis of a Sample Selected With Unequal Selection Probabilities Related to the Outcome.

Another likely source of bias occurs when sample selection probabilities are related to the outcome of interest. From Section 2.2.1.2 it is clear that the resulting bias is huge even for a weighted analysis of inverse of selection probabilities. To correct for such kind of bias the precise relationship between the selection probability and the outcome should be known. However in many practical situations this is not always the case hence assumptions have to be made about the existing relationship between the outcome and selection probability. In some studies like, epidemiological studies, bias is induced when

another variable that is not part of the analysis, referred to as a confounding factor is related to both the outcome and the predictor. The bias problem is simply solved by including the confounding variable in the model. We shall explore if the biasness in the sample selected with unequal selection probabilities that depend on the outcome in Section 2.2.1.1 can be solved by considering cluster size as a confounding factor and including it as a variable in the GEE model and GLMMs. The weighted GLMM introduced in Section 3.2.1 will also be used since it can also perform well with informative selection probabilities especially when the second scaling scheme is used.

Methods discussed in the previous paragraph assume that the exact relationship between the outcome and selection probabilities is known. In cases where this is not true we introduce a method which is an extension of the generalized linear models which we shall refer to as the combined model.

3.3.1. The Combined Model

The combined model is basically an extension of generalized linear mixed model which apart from random effects that account for correlation of outcomes within the same cluster, an extra set of random effects is introduced to account for extra sources of bias. In this scenario the extra random effects are supposed to account for extra variation due to selection bias. The method is ideal when the exact relationship between the outcome and selection probabilities is not known.

Specifically, let every element y_{ck} be assigned a latent trait ϑ_{ck} which in this case can represent the selection probability. Further, let $f(y_{ck}, \vartheta_{ck}, b_c)$ represent the joint distribution of the vector $(y_{ck}, \vartheta_{ck}, b_c)^T$, where y_{ck} and b_c are as defined before. We will assume that conditional on $\boldsymbol{\vartheta}_c = (\vartheta_{c1}, \ldots, \vartheta_{cn_c})$ and $\boldsymbol{b_c}$, the components in $\boldsymbol{Y_c}$ (i.e., observations from the same cluster) are independent. Specifically,

$$P(\boldsymbol{Y}_c = \boldsymbol{y_c} | \boldsymbol{\vartheta}_c, \boldsymbol{b_c}) = \prod_{k=1}^{n_c} P(Y_{ck} = y_{ck} | \vartheta_c, b_c), \tag{5}$$

Where $\vartheta_{ck} \sim Beta(\lambda, \omega)$ defines the new set of random effects introduced to account for extra variation. For example, if we consider food item as a cluster in pesticides monitoring study, this would mean that conditional on both sets of random effects, the outcome of a food samples taken from a food item would not be influenced by the outcome of another sample taken from the same food item. Note that (5) also represents the likelihood contribution of cluster $\boldsymbol{c}$. For completeness we will assume that

$$P(Y_{ck} = 1 | \vartheta_c, b_c) = \vartheta_{ck} \frac{\exp(\sum_{h=1}^{H} \beta_h x_{hck} + b_c)}{1 + \exp(\sum_{h=1}^{H} \beta_h x_{hck} + b_c)},$$

Model 7

Hence the estimates will be obtained by maximizing the marginal likelihood

$$L(\boldsymbol{\beta}, \lambda, \omega, \boldsymbol{D}) = \prod_{c=1}^{C} P(\boldsymbol{Y}_c = \boldsymbol{y_c} | \boldsymbol{\vartheta}_c, \boldsymbol{b_c}),$$

under the assumption that distribution functions of the random effects, $f(\vartheta_{ck})$ and $f(b_c)$ are independent. Importantly, since ϑ_{ck} can only take positive values between 0 and 1 we implicitly assume that elements in the sample have high probability of a success than elements not selected. In the specific example of the pesticides monitoring study, we assume that samples selected were more likely to have residues above MRL than those not selected resulting into overestimation of the mean. This can be the case if highly consumed food items which also tend to have high exceedance rates are given high probability of selection. Note that Model 7 is basically a mean weighted. More details on the combined model and its applications can be found in Molenberghs *et.al.* (2010) and Alonso *et. al* (2014).

The marginal probability is obtained as

$$P(Y_{hck} = 1|) = \frac{\lambda}{\lambda + \omega} \int \frac{\exp(\sum_{h=1}^{H} \beta_h x_{hck} + b_c)}{1 + \exp(\sum_{h=1}^{H} \beta_h x_{hck} + b_c)} \phi(b_c|0, \hat{\tau}^2)\, db_c.$$

3.4. Results: Analysis of Data With Unequal Informative Selection Probabilities.

The performance of the methods used to analyse a biased sample due to informative unequal selection bias was investigated by analysing the sample selected from a population where bigger sized cluster have higher means than small sized clusters. Further the selection probability for each element was a function of cluster sizes which resulted into elements from big clusters having higher probability of selection than elements from small sized clusters. As seen in Table 24, just accounting for design aspects of the sampling design in the analysis does not suffice. Results for analysis of the sample with the methods with corrective measures are provided in Table 45. The SAS programming codes are provided at the end of the section.

Results from GLMM and GEE model with cluster size as a variable were similar in that they both overestimated the mean (only GLMM results provided). This is because selection bias differs from confounding bias as illustrated in Hernan et al (2004) in the *"common effect"* scenario. In this case cluster size would be regarded as a common effect i.e., large clusters have high means and high selection probabilities hence the selected sample will mostly be populated by elements from large clusters. If we dichotomize cluster size into large (1) and small (0) clusters then the sample will almost only have one level of cluster size (1) hence conditioning on the cluster size (adding it to the model) will have no effect in reducing bias and still lead to biased estimates.

On the other hand the weighted GLMMs which incorporate some exact information about the relationship between the outcome and selection probabilities (the weights) perform reasonably well in terms of bias of the mean estimates, especially for the second scaling method. Notably, the variance of random effects is heavily biased. The combined model, which simply assumes that the selection probabilities are randomly sampled from a beta distribution performs reasonably well considering that it is based only on the assumption that selected elements are likely to have higher means than the non-selected ones. It tends to overestimate very small probabilities which is a direct result of overestimating the conditional effects($\boldsymbol{\beta}$). The combined model being complicated can have some numerical issues as noted in the estimation of variance, however with a single dataset (unlike 200 simulated datasets) it is usually possible to solve these issues by for example trying several sets of starting values. For GLMM and combined model simulations runs with numerical problems the variance of the random effect was replace by the true value 0.25 when computing the marginal probability.

It follows that when information about the relationship between the selection probabilities and the outcome probability is available, then models that can utilize such information, like weighted GLMM should be used. When such information is not available, models like the combined model that should be used. Importantly, when selection bias is suspected it is recommended to carry out a sensitivity analysis to check the stability of the results under different assumed relationships between the outcome and the selection probabilities. Decision on which method to use when results from a sensitivity analysis are different would have to be based on expert opinion, for example model that produces results close to what is expected in reality should be preferred. In summary, Figure 28 shows the decision tree for choosing methods of analysis in presence of selection bias.

Table 45 Results From Stratified Analysis Exploring Methods for Analyzing Sample Data Selected With Non-Informative Unequal Selection Probabilities.

Model	$\bar{Y}$	**GLMM + Cluster size**	**GLMM +** $W_{k\|c}^{(1)}$	**GLMM +** $W_{k\|c}^{(2)}$	**Combined**
Stratum			$\bar{y}$		
1	0.0733	0.1105	0.0938	0.0714	0.0746
2	0.0002	0.0200	0.0000	0.0000	0.0114
3	0.0099	0.0295	0.0000	0.0125	0.0191
4	0.1061	0.1708	0.1186	0.0975	0.1071
5	0.0726	0.1183	0.0845	0.0689	0.0738
6	0.0314	0.0533	0.0351	0.0411	0.0353
7	0.0095	0.0260	0.0147	0.0090	0.0161
8	0.0002	0.0200	0.0000	0.0002	0.0114
9	0.0469	0.0711	0.0473	0.0471	0.0495
10	0.0074	0.0251	0.0071	0.0071	0.0153
11	0.0895	0.1344	0.0716	0.0793	0.0870
12	0.0347	0.0547	0.0377	0.0332	0.0389
13	0.0590	0.0828	0.0527	0.0529	0.0556
14	0.0003	0.0201	0.0000	0.0001	0.0115
15	0.0083	0.0256	0.0065	0.0082	0.0159
16	0.0003	0.0200	0.0000	0.0013	0.0114
17	0.0121	0.0286	0.0042	0.0140	0.0178
18	0.0181	0.0386	0.0225	0.0242	0.0264
19	0.0532	0.0836	0.0123	0.0578	0.0558
20	0.0003	0.0197	0.0000	0.0003	0.0113
21	0.1240	0.1867	0.1060	0.0994	0.1161
22	0.0478	0.0731	0.0205	0.0467	0.0508
23	0.0153	0.0341	0.0098	0.0183	0.0226
24	0.0393	0.0691	0.0102	0.0441	0.0468
25	0.0769	0.1119	0.0363	0.0639	0.0734
26	0.0302	0.0511	0.0145	0.0322	0.0339
27	0.0542	0.0821	0.0370	0.0507	0.0542
28	0.0692	0.1075	0.0306	0.0619	0.0700
29	0.1207	0.1712	0.0815	0.0900	0.1065
τ^2	0.2500	≈.0000	0.5662	4.744	0.1504

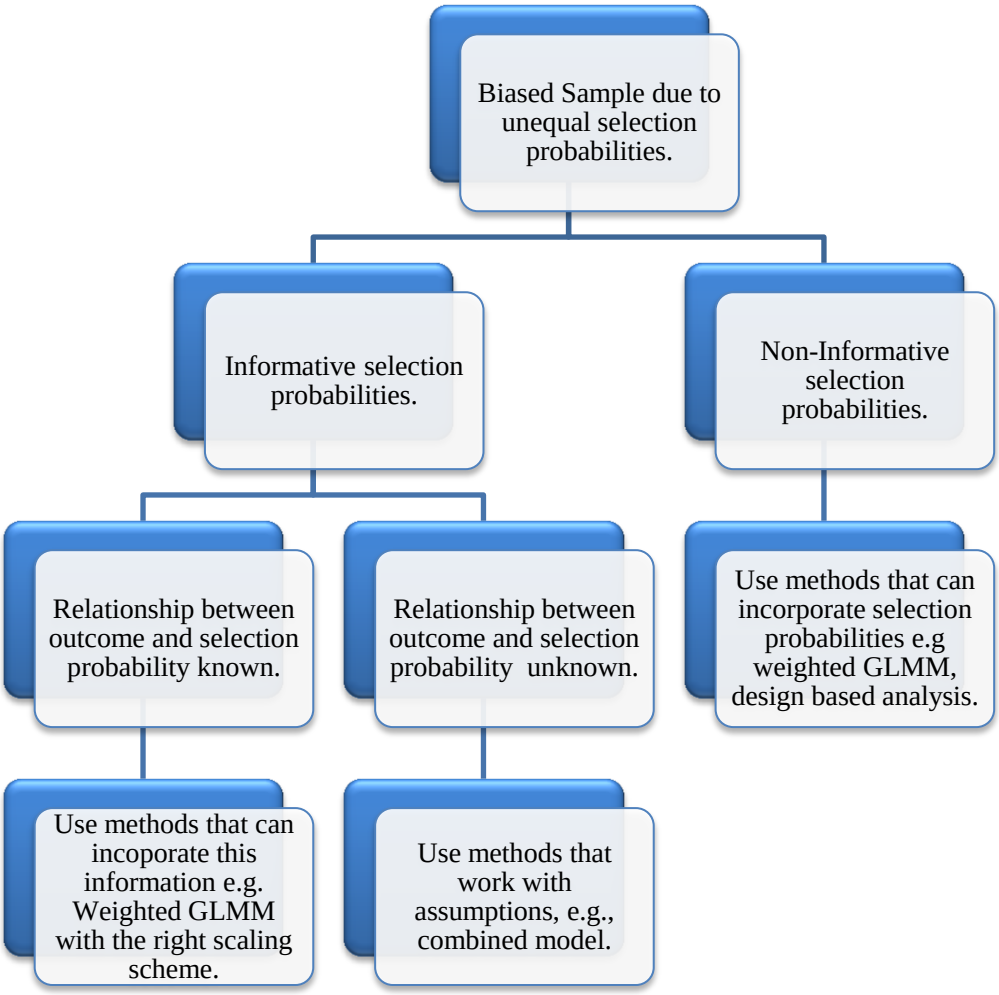

Figure 28 Decision Making Tree in Determining Methods of Analysis For a Biased Sample Due to Unequal Selection Probabilities.

Programming codes for obtaining results in Table 45 using SAS 9.4 (Comments are provided in green)

Codes for obtaining weighted GLMM results are similar to those used to obtain the results for Table 44 and the code for the combined model is as follows:

```
/*The combined model*/

proc nlmixed data=biascomb qpoints=20 empirical ;
/*data specifies the dataset and qpoints the number of quadrature points*/
by replicate;
parms/data=start; /*specifying a dataset containing starting values for the parameters*/
eta = β1S1 + β2S2 + ⋯ + β29S29+ b ; /*The mean structure which in this case is a function of
dummy variables the strata (S1, … , S29) and their corresponding parameter coefficient
(β1, … , β29)*/
expeta=exp(eta);
ll = -log(1+exp(log(const))) + resp*eta - resp*log(1+expeta)+ (1-resp)*log((1-
expeta/(1+expeta)) + exp(log(const)));
/*The likelihood contribution of each cluster after integrating out the beta distributed random
effects, const= ω/λ and resp is the response variable */
model resp ~ general(ll); /*Specifies the likelihood to maximize*/
random b ~ normal(0,exp(2*log(sigma))) subject = cid; /*Specifies the random effects*/
estimate 'variance' sigma**2;

ods output ParameterEstimates=combparms_bias01 AdditionalEstimates=combcov_bias01;
run;
ods select all;
```

3.5. Methods for Analyzing a Non-Representative Sample Due to Non-Coverage.

As noted before, over-coverage mainly affects the economic side of the survey in that more resources than necessary may be used. On the other hand, under-coverage usually results into bias whose direction depends on the nature of the outcome for the left out elements. If elements with high values of the outcome are left out, negative bias will result and if low outcome elements are left out positive bias may result.

The problem of under coverage is mostly dealt with in a similar manner to missing data where the elements left out the sampling frame are considered missing. Due to this similarity, for solutions to

dealing with under-coverage bias, we refer the reader to the following section, which provides methodology for missing data.

3.6. Likelihood Method for Dealing with Non-response

In Section 2.5, we illustrated the bias that results in using simple methods of analysis when the non-response mechanism deviates from the restrictive MCAR assumption. We also highlighted different possible views of non-response: failure to provide the required number of samples, and failure to analyse/report all the prescribed residues. In Section 2.9, we explored the non-response patterns under the different views, for the 2010 pesticides monitoring data.

In this section, we focus on non-response in terms of failure to analyse/report all the defined residues, and provide a likelihood-based analysis method, which is valid under the less-restrictive MAR assumption.

We will deal with results at the determination level. First, we explore the possible hierarchical structures under which results at the determination level can be considered to be clustered.

Under one perspective, all determinations of all samples of a particular commodity, from the same country of origin, can be considered to be clustered. In this case, the "country of origin – food item" combination would be the cluster. The determinations from each of such clusters would be expected to be correlated, as a possible reflection of the agricultural practice in the specific "country of origin" with respect to the commodity of interest. Note that under this perspective, the agricultural practice within each specific "country of origin", for each commodity, is assumed to be uniform across the different pesticides.

Under a second perspective, all determinations for the same residue, for all samples of a particular commodity, from the same country of origin, can be considered to be clustered. In this case, the "country of origin – food item – residue" combination would be the cluster, with the determinations from each of such clusters being expected to be correlated. This perspective extends the one above, by assuming that agricultural practice in the specific country of origin, for the particular commodity, is pesticide-specific.

Now, whether under the first or the second perspective, for each cluster, let $\boldsymbol{Y}_i^o$ represent the observed information (the available determinations), and $\boldsymbol{Y}_i^m$ represent the missing information (the residue determinations which were not conducted/reported). Additionally, let the components of the vector of missing data indicators, $\boldsymbol{R}_i$ take the value 1 if the particular determination is available, and 0 otherwise. This vector represents the missing-data mechanism. For each cluster, the so-called full data (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2000) is represented by $(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, \boldsymbol{R}_i)$.

To base inference on the observed data, we will need to integrate out the missing component, $\boldsymbol{Y}_i^m$:

$$f(\boldsymbol{y}_i^o, \boldsymbol{r}_i) = \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \boldsymbol{\beta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\tau}) d\, \boldsymbol{y}_i^m.$$

The parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are for the process under measurement, and the missingness process, respectively. The above factorization of the full data into the measurement process, and the conditional distribution of the missingness process, given the measurements, is called the selection model factorization (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2000).

Under the MAR assumption, the missing measurements are dropped from the conditional distribution of the missingness process, and the integral becomes:

$$f(\boldsymbol{y}_i^o, \boldsymbol{r}_i) = \int f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \boldsymbol{\beta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i^o, \boldsymbol{\tau}) d\, \boldsymbol{y}_i^m.$$

Note that, as introduced in Section 2.5, this is the assumption that given that the cluster's available information has been taken into account, then the non-response mechanism does not further depend on the unobserved information.

If, in addition, the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are distinct, the integral above becomes:

$$f(\boldsymbol{y}_i^o, \boldsymbol{r}_i) = f(\boldsymbol{y}_i^o | \boldsymbol{\beta}) f(\boldsymbol{r}_i | \boldsymbol{\tau}).$$

The likelihood factorizes into two components: one for the observed measurements, and one for the missingness process. This implies that a direct likelihood analysis based on the observed measurements alone will be valid (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2000).

We will therefore consider the generalized linear mixed model to estimate the probability of MRL exceedance.

Under the "country of origin – food item" cluster perspective, consider the following generalized linear mixed model:

$$\text{logit}\{P(Y_{ijk} = 1 | b_{ij})\} = \beta_0 + b_{ij}.$$

The model states that the logit of the probability of MRL non-compliance for the k^{th} determination, related to commodity j, with country of origin i, depends on an overall parameter, β_0, and a random effect which is specific to all determinations from all samples of a particular commodity coming from a specific country of origin. The random effect has the following distribution: $b_{ij} \sim N(0, \sigma_b^2)$. Note that the model implies the following, for the probability of MRL non-compliance for the k^{th} determination, related to commodity j, with country of origin i:

$$P(Y_{ijk} = 1 | b_{ij}) = \frac{e^{(\beta_0 + b_{ij})}}{1 + e^{(\beta_0 + b_{ij})}}.$$

Under the "country of origin – food item – residue" cluster perspective, consider the following model for the probability of MRL non-compliance for the l^{th} determination related to residue k, for commodity j, with country of origin i:

$$P(Y_{ijkl} = 1 | b_{ijk}) = \frac{e^{(\beta_0 + b_{ijk})}}{1 + e^{(\beta_0 + b_{ijk})}}.$$

Note that here, we consider a random effect $b_{ijk} \sim N(0, \sigma_b^2)$, which is specific to all determinations of a given residue, from all samples of a particular commodity, coming from a specific country of origin.

One thing to notice here is that by using the "country of origin – food item" and "country of origin – food item – residue" as random effects, we are able to generalize our results to the population of all possible "country of origin – food item" and "country of origin – food item – residue" clusters.

In the following, we will focus on the 2010 pesticides monitoring data for the 30 compounds studied in Section 2.9. For these compounds, there were 178231determinations, of which 24 were MRL non-compliant.

Table 46 below provides information on the number of clusters, and summaries of the cluster sizes, for the two clustering perspectives.

Table 46 Number of Clusters, and Cluster Sizes, for the 2 Clustering Perspectives

	Cluster Definition	
	Country of Origin – Food Item	Country of Origin – Food Item – Residue
Number of Clusters:		
Total Number	363	6177
Of At Least Size 10	362	3596
Cluster Sizes:		
Minimum Size	5	1
Average Size	491	29
Maximum Size	6345	426

Once the models are fitted, the estimate of the parameter β_0, and empirical Bayes estimates for the random effects (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2000), can be used to calculate cluster specific exceedance probabilities.

In Table 47, the 20 clusters, with the highest estimated exceedance probabilities, are provided. The table also provides the expected estimates for clusters which are at the mean of the random effects distribution (when the random effect is zero). For the cluster names, the first part of the name is the member state abbreviation, the second part the food item, and the third part the pesticide label (in the case of the "country of origin – food item – residue" clusters). The pesticide labels are as introduced in Section 2.9.

The 20 clusters with the lowest estimated exceedance probabilities are also provided, in Table 48. Under the "country of origin – food item" cluster definition, the exceedance probabilities for all the 363 clusters are provided in Table 68 in Appendix C. Note that the analyses are now valid under the less-restrictive MAR assumption. These analyses can be performed using the SAS procedure "NLMIXED"; the corresponding code is provided below.

In general, the missing data mechanism is unknown, and therefore it is difficult to rule out the possibility of even a MNAR mechanism operating. More complex models, which explicitly define a model for the missingness process, exist (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2000). In general, however, the impact of missingness is usually subjected to a sensitivity analysis, in which a range of assumptions about the mechanism are made. The goal in a sensitivity analysis is to check for the stability, or lack thereof, of the results under the various assumed mechanisms. If the results show a lot of sensitivity to these assumptions, then the results obtained need to be treated with caution.

```
/*************************************************************************************/
/*Code for likelihood analysis in the presence of missing data.                      */
/*Data "nonresp"  contains at least the following variables:                         */
/*        "origcntry_fooditm_resid" = an identifier for the country of origin - food item – residue  */
/*                                    combination,                                   */
/*        "origcntry_fooditm"  = an identifier for the country of origin - food item combination,    */
/*        "response" = response variable; 1 if MRL exceedance, 0 otherwise.          */
/*Data need to be sorted by country of origin, food item, and residue.               */
/*************************************************************************************/

/*"country of origin – food item" cluster "*/

proc nlmixed data=nonresp;
                eta=beta0+b;                                    /*linear predictor for the logit model*/
                p=exp(eta)/(1+exp(eta));                        /*probability of exceedance*/
                loglik=response*log(p)+(1-response)*log(1-p);   /*log likelihood contribution of each
                                                                  response*/
model response ~ general(loglik);
                random b ~ normal(0,sigmab**2) subject=origcntry_fooditm;
                                                        /* random effect of "origcntry_fooditm" */
                estimate 'Prob. at Mean of R.E.s Dist' exp(beta0)/(1+exp(beta0));  /* exceedance
                                                      probability at mean of the random effects distribution*/
                predict exp(beta0+b)/(1+exp(beta0+b)) out=nonresppredclust1; /*construct and output
                                                    the cluster exceedance predictions*/
run;

/*"country of origin – food item – residue" cluster*/

proc nlmixed data=nonresp;
                eta=beta0+b;                                    /*linear predictor for the logit model*/
                p=exp(eta)/(1+exp(eta));                        /*probability of exceedance*/
                loglik=response*log(p)+(1-response)*log(1-p); /*log likelihood contribution of each
                                                                  response*/
model response ~ general(loglik);
                random b ~ normal(0,sigmab**2) subject=origcntry_fooditm_resid;
                                                        /* random effect of "origcntry_fooditm_resid" */
                estimate 'Prob. at Mean of R.E.s Dist' exp(beta0)/(1+exp(beta0));  /* exceedance
                                                      probability at mean of the random effects distribution*/
                predict exp(beta0+b)/(1+exp(beta0+b)) out=nonresppredclust2; /*construct and output
                                                    the cluster exceedance predictions*/
run;
```

Table 47 Cluster-Specific Exceedance Probability Estimates, for the Top 20 Clusters, Based on Generalized Linear Mixed Models

	Country of Origin – Food Item		Country of Origin – Food Item – Residue	
		Probability		Probability
At Average of R.E. Distribution		**0.000001852**		**0.000008817**
	Cluster		Cluster	
	MA,Peaches	.005432881	GR,Lettuce,C2	0.047701
	GR,Lettuce	.004915654	RO,Lettuce,C2	0.011753
	CY,Peaches	.004070267	ES,Tomatoes,B1	0.002701
	CY,Strawberries	.003513229	MA,Peaches,F1	0.002112
	CY,Lettuce	.003513229	CY,Peaches,E1	0.001926
	CY,Apples	.003104947	CY,Strawberries,C5	0.001874
	EG,Strawberries	.002482097	CY,Apples,F1	0.001874
	RO,Lettuce	.002125042	CY,Lettuce,D3	0.001874
	ES,Leek	.001581729	ES,Leek,C5	0.001795
	PT,Lettuce	.001471401	NZ,Apples,D3	0.001584
	NZ,Apples	.001306602	PT,Lettuce,E1	0.001564
	FR,Strawberries	.000996946	EG,Strawberries,C5	0.001544
	ZA,Peaches	.000876351	EG,Strawberries,P5	0.001525
	GR,Peaches	.000625250	ZA,Peaches,E1	0.001407
	ES,Tomatoes	.000530767	FR,Strawberries,E1	0.001377
	CL,Apples	.000459893	GR,Peaches,D2	0.001214
	ES,Peaches	.000134166	CL,Apples,D3	0.001098
	GR,Rye	.000001852	ES,Tomatoes,D2	0.000659
	BG,Oats	.000001851	ES,Peaches,B1	0.000578
	CH,Apples	.000001851	AL,Head cabbage,B1	0.000009

Table 48 Cluster-Specific Exceedance Probability Estimates, for the Lowest 20 Clusters, Based on Generalized Linear Mixed Models

Country of Origin – Food Item		Country of Origin – Food Item – Residue	
	Probability		Probability
Cluster		Cluster	
BE,Leek	.000001686	ES,Peaches,P2	0.000009
DE,Lettuce	.000001685	ES,Peaches,C5	0.000009
RO,Tomatoes	.000001677	ES,Tomatoes,P5	0.000009
DE,Apples	.000001676	ES,Tomatoes,E1	0.000009
GB,Swine meat	.000001673	ES,Tomatoes,P1	0.000009
DE,Milk and milk products	.000001666	ES,Tomatoes,T1	0.000009
DE,Swine meat	.000001656	ES,Tomatoes,D3	0.000009
NL,Leek	.000001652	ES,Tomatoes,P4	0.000009
DK,Swine meat	.000001647	ES,Tomatoes,C2	0.000009
DE,Strawberries	.000001646	ES,Tomatoes,C3	0.000009
ES,Strawberries	.000001631	ES,Peaches,P5	0.000009
IT,Apples	.000001624	ES,Peaches,P4	0.000009
DE,Head cabbage	.000001602	ES,Peaches,D2	0.000009
RO,Apples	.000001601	ES,Peaches,P1	0.000009
FR,Apples	.000001565	ES,Peaches,T1	0.000009
FR,Lettuce	.000001556	ES,Peaches,M1	0.000009
NL,Tomatoes	.000001551	ES,Peaches,E1	0.000009
ES,Lettuce	.000001544	ES,Peaches,D3	0.000009
IT,Peaches	.000001536	ES,Peaches,C3	0.000009
GB,Milk and milk products	.000001485	ES,Peaches,C2	0.000009

3.7. Likelihood Method for Dealing with Left Censoring

In this section, we focus on maximum likelihood estimation of the mean residue concentration, taking left censoring into account. The objective will be to estimate the mean residue concentration for a food item-residue combination. We will also account for the clustering of food items within the countries of origin.

Censored data are usually represented by the pair $(Y_i, \delta_i), i = 1 \ldots n$, where Y_i is the outcome for the i^{th} subject, and δ_i indicates whether the outcome is observed ($\delta_i = 1$), or censored ($\delta_i = 0$) (Duchateau and Janssen, 2008; Klein and Moeschberger, 1991; Rizopoulos, 2012). There are various censoring types, but here we focus on left censoring.

Given a distributional assumption $f(y)$ for Y, the left censored outcomes are often assumed to contribute "partial" information to the likelihood. In particular, while the observed outcomes are assumed to contribute the probabilities $P(Y_i = y_i)$, the left censored cases are assumed to contribute the probabilities $P(Y_i < y_i)$. Thus, the censored outcomes are taken to provide the information that the outcome is less than the censoring value. This assumption is referred to as the non-informativeness assumption.

For each food item-residue combination separately, let Y_{ij} denote the residue concentration result for the j^{th} determination, related to a sample which has country i as its origin.

Assuming the lognormal distribution, then Z_{ij} =log (Y_{ij}) can be assumed to follow the normal distribution. We therefore consider a mixed model (Verbeke and Molenberghs, 2000) for the estimation of the mean residue concentration. We consider the following model:

$$Z_{ij} = \beta_0 + b_i + \varepsilon_{ij}$$

The model contains an overall mean, β_0, a random effect of the country of origin, b_i, and the residual error of the j^{th} determination, related to a sample with country of origin i. The effects b_i and ε_{ij} are independent, with distributions: $b_i \sim N(0, \sigma_b^2)$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$.

This implies that $Z_{ijk}|b_i \sim N(\beta_0 + b_i, \sigma^2)$, the so-called conditional distribution of the outcome, given the random effect. The likelihood contribution of determinations related to samples with the same country of origin is therefore

$$L_i = \int \left[\prod_j \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\{z_{ij}-(\beta_0+b_i)\}^2} \right\}^{\delta_{ij}} \left\{ \int_{-\infty}^{z_{ij}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\{z_{ij}-(\beta_0+b_i)\}^2} dz_{ij} \right\}^{1-\delta_{ij}} \right] db_i.$$

The indicator δ_{ij} distinguishes between results which are measured above the LOQ ("observed results"; $\delta_{ij} = 1$), and those which are measured below the LOQ (censored; $\delta_{ij} = 0$). The values measured below the LOQ contribute the information that the response is below the LOQ.

The model above can be implemented by defining the likelihood contribution in software like the SAS procedure NLMIXED. Sample SAS code is provided at the end of this section.

Note that $Z_{ij} \sim N(\beta_0, \sigma^2 + \sigma_b^2)$, the so-called marginal distribution of the outcome. From this, the mean residue concentration for the food item-residue combination can be estimated from the mixed model as

$$\theta = \exp\left\{\beta_0 + \frac{(\sigma^2+\sigma_b^2)}{2}\right\}.$$

We focus on the subset of the residue concentration data, introduced in Section 2.10. One point to note is that when estimating models in a censored data context, there is a recommended minimum number of non-censored cases, for each parameter/effect to be included in the model (Allison, 2010; Harrell, 1996; Peduzzi, 1995; Rizopoulos, 2012); the recommendations range from 5 up to 15 non-censored observations for each parameter.

We focus once again on concentration data for the 30 residues, explored in Section 2.10. For these data, there were 222 food item-pesticide combinations, and 77 of these combinations had at least one sample with concentration measured above the limit of quantification. Out of the 77 combinations, 53 had at least 2 samples measured above the LOQ, with 26 combinations having at least 15 samples measured above the LOQ, and 15 combinations with at least 30 samples measured above the LOQ.

For our 3-parameter model above, we will consider for illustration purposes the 15 "food item – residue" combinations having at least 30 non-censored cases. For these 15 combinations, we summarize the total number of determinations in each, the number (percentage) censored, and the number observed (measured above LOQ), in Table 49. In Table 50, a summary of the cluster sizes within the "food item – residue" groups is presented.

The mean residue concentrations for the 15 food item – residue combinations, estimated using the likelihood method, taking both censoring and clustering into account, are provided in Table 51.

The results obtained by only taking censoring into account, ignoring the possible clustering, are also provided in the same table. Note that these results are obtained by omitting the random effect in the model above, i.e. assuming the model

$$Z_{ij} = \beta_0 + \varepsilon_{ij}.$$

In that case, the mean residue concentration is estimated using

$$\theta = \exp(\beta_0 + \sigma^2).$$

Finally, the results obtained by substituting the results below LOQ by either the LOQ, or a very small value, are also provided. These results are obtained by first making the corresponding substitution, computing the sample average and the sample variance from the log-transformed values of the substituted data, and then estimating the mean residue concentration as $\hat{\theta}_0 = \exp(\bar{y} + s^2/2)$. Note that no model is fitted in this case.

Table 49 Number of Determinations (Percentages) Observed and Censored, for 15 Food Item – Residue Combinations

Food Item – Residue	Censored Determinations	Observed Determinations	Total Determinations
Apples,B1	1934 (98.02)	39 (1.98)	1973
Apples,C2	1737 (86.76)	265 (13.24)	2002
Apples,C3	1935 (98.22)	35 (1.78)	1970
Lettuce,B1	1424 (96.22)	56 (3.78)	1480
Lettuce,C5	1276 (97.40)	34 (2.60)	1310
Milk and milk products,D1	308 (89.53)	36 (10.47)	344
Milk and milk products,H2	600 (94.34)	36 (5.66)	636
Oats,P4	210 (87.14)	31 (12.86)	241
Peaches,B1	1102 (95.49)	52 (4.51)	1154
Peaches,C2	1009 (86.17)	162 (13.83)	1171
Peaches,C5	963 (95.92)	41 (4.08)	1004
Rye,P4	322 (90.20)	35 (9.80)	357
Tomatoes,B1	1613 (97.52)	41 (2.48)	1654
Tomatoes,C2	1678 (97.96)	35 (2.04)	1713
Tomatoes,C3	1667 (98.23)	30 (1.77)	1697
Total	17778 (95.04)	928 (4.96)	18706

Table 50 Summary of Cluster Sizes within the 15 Food Item – Residue Combinations

Total number of clusters	530
Number of clusters of at least size 10	287
Number of clusters of at least size 5	371
Minimum Cluster Size	1
Average Cluster Size	35
Maximum Cluster Size	426

Table 51 Mean Residue Concentration Estimated Using Different Methods. ML=Maximum Likelihood

Food Item – Residue Combination	Substitution by LOQ	Substitution by Small Value	ML – Clustering Not Taken Into Account	ML – Clustering Taken into Account
Apples,B1	0.0178	0.000000007	0.0033	0.0260
Apples,C2	0.0199	8.1178	0.0177	0.0223
Apples,C3	0.0184	0.000000003	0.0011	0.0010
Lettuce,B1	0.0144	0.0000005	0.7174	0.1242
Lettuce,C5	0.0276	0.00000005	0.8607	0.6722
Milk and milk products,D1	0.0107	0.0011	0.0023	0.0590
Milk and milk products,H2	0.0040	0.0000001	0.0003	0.0003
Oats,P4	0.0210	11.3322	0.0684	0.0375
Peaches,B1	0.0150	0.00000085	0.0036	0.0054
Peaches,C2	0.0159	5.3550	0.0083	0.0071
Peaches,C5	0.0283	0.00000073	0.0140	0.0335
Rye,P4	0.0174	0.0305	0.0657	0.0840
Tomatoes,B1	0.0150	0.00000002	0.0023	0.0007
Tomatoes,C2	0.0144	0.00000001	0.0082	0.0084
Tomatoes,C3	0.0145	0.00000001	0.0582	0.0141

The SAS code for estimating the mean residue concentration, taking both censoring and clustering into account, is provided below.

Note that though we have used the lognormal distribution here, there should be said that goodness of fit test should be performed in order to be able to use the results obtained and make an informed comparison, but in any case it is possible to consider other distributions, and even compare them, to select the distribution which provides the best fit for the data. The Weibull distribution is an example of alternative distributions that could be considered. Further treatment of this issue can be found in EFSA, 2010[21].

[21] European Food Safety Authority; Management of left-censored data in dietary exposure assessment of chemical substances. EFSA Journal 2010;8(3):.[96pp.]. doi: 10.2903/j.efsa.2010.1557.

```
/*************************************************************************************/
/*Code for likelihood analysis, taking both censoring and clustering into account.    */
/*Data "leftcens"  contains at least the following variables:                         */
/*        "fooditm_resid" = an identifier for the food item – residue combination,    */
/*        "noncensbyfooditmresid" = the number of non-censored cases by "fooditm_resid",*/
/*        "delta" = indicator for censoring; 1 if not censored, 0 otherwise,         */
/*        "logresval" = the logarithm of the residue values; for censored, logarithm of censoring*/
/*                      value,                                                       */
/*        "origcountry" = the sample's country of origin.                            */
/*Data need to be sorted by "fooditm_resid" and "origcountry".                        */
/*************************************************************************************/

proc nlmixed data=leftcens;
        where noncensbyfooditmresid ge 30; /*analysis for food item - residue combinations with at
                                             least 30 non-censored cases*/
        by fooditm_resid;                  /*analysis by food item – residue combination*/
                mu=beta0+b;                /*conditional mean of the mixed model for the logarithm
                                             of the values*/
                if delta=1 then do;
                loglik=(-0.5*log(2*constant('pi')))-log(sigmaError)-(0.5*(1/(sigmaError**2))*(logresval-
                        mu)**2);           /*if not censored, contribution to the (log) likelihood is
                                             the  (log) probability density function. "pi"=π*/
                end;
                else if delta=0 then do;
                lik=cdf('normal',logresval,mu,sigmaError); /*if censored, contribution to the likelihood
                                                             is the cumulative distribution function*/
                loglik=log(lik);
                end;
                model logresval ~ general(loglik);
                random b ~ normal(0,sigmab**2) subject=origcountry; /*random effect of country of
                                                                      origin*/
                estimate 'Log-Normal Mean' exp(beta0+(sigmaError**2+sigmab**2)/2); /*estimate
                                                                                     lognormal mean: the mean residue concentration*/
                ods output AdditionalEstimates=meanresidconc; /*output the estimated mean residue
                                                                concentration*/
run;
```

3.8. Handling "Not Evaluated" Results

In the 2010 pesticides monitoring data, there was a notable presence of results referenced as "result not evaluated", coded as "J029A". Out of the 1226916 determinations, 345405 results, representing

28.15% of all the determinations, were "not evaluated". Results are usually coded as such for a variety of reasons[22].

From an analysis perspective, there is need to define whether such results should be treated as missing data, or otherwise. For example, in the 2010 report, the percentages of the MRL exceedances provided did not separately report the percentage of "not evaluated" results; as such, since the percentages reflected the number of samples exceeding MRL out of all the other results, "not evaluated" results included, then the "not evaluated" results played the role of results which did not exceed the MRL.

How the "not evaluated" results are treated will have an impact on the estimates obtained. In the case of MRL non-compliance estimation, whether or not these results are treated as missing data, will have an impact on the percentage of non-response.

In the case of mean residue concentration, whether these results are treated as missing data or not, will generally define the number of issues to be dealt with: either left censoring only, or both left censoring and missing data.

Consider the maximum likelihood analysis conducted in Section 3.7 to estimate the mean residue concentration, accounting for both left censoring, and clustering. For the data considered in that section, out of the 178231 results, 48354 of them, which represented 27.13% of the results, were "not evaluated" cases. Table 52 below contrasts what happens when these cases are considered as actual data, versus missing data. Already, some differences emerge; for instance, 0.0080 versus 0.0260, and 0.0046 versus 0.0590. The message is that results could differ, from mildly, to drastically, hence the need to define the role these cases should play in data analyses.

Table 52 Mean Residue Concentration Estimated Using Maximum Likelihood, Taking Censoring and Clustering into Account, but Considering J029A as either Actual Data or Missing Data

"Not Evaluated" Results as Missing Data	"Not Evaluated" Results as Actual Data
0.0080	0.0260
0.0199	0.0223
0.0014	0.0010
0.1536	0.1242
0.5225	0.6722
0.0046	0.0590
0.0004	0.0003
0.0380	0.0375
0.0030	0.0054
0.0075	0.0071
0.0226	0.0335
0.0891	0.0840
0.0009	0.0007
0.0079	0.0084
0.0122	0.0141

[22] European Food Safety Authority; Use of the EFSA Standard Sample Description for the reporting of data on the control of pesticide residues in food and feed according to Regulation (EC) No 396/2005 (Revision 2). EFSA Journal 2013;11(1):3076.[54 pp.] doi:10.2903/j.efsa.2013.3076.

CONCLUSIONS

In this report, we have combined survey sampling methodology, with statistical modelling, to assess the design and analysis of the multiannual control programmes. We have illustrated the problems that may arise if the design deviates from good survey sampling practices, and proposed how such deviations could be avoided. However, in the conduct of surveys, certain problems may be unavoidable, for instance, missing data. We have illustrated how statistical methodology could be used to mitigate the effect of such problems, on the study results. It is therefore essential that the two ingredients, namely, survey sampling methodology, and statistical modelling, take a prominent role right from the design stage, to the analysis stage, of the programmes.

RECOMMENDATIONS

Targeted Population:

The targeted population should be clearly defined. This determines the extent to which results can be generalized. The structure of the population usually determines the sampling design to be used.

Sampling Frame:

This is critical in determining representativeness of the data, and should be clearly defined. If the sampling frame excludes some elements of the targeted population, representativeness of the data may be questionable. For example, assuming that the targeted population in the pesticide monitoring program is "all food items available to European consumers", a good sampling frame would need to ensure that every food item available to the European consumer has a non-zero selection probability. However, such a sampling frame may be unrealistic, since a list of all food products available to European consumers may be hard to obtain. Reasonable exclusion criteria could be defined as long as the resulting sampling frame, after exclusion, remains representative enough of the targeted population. The impact of using a non-exhaustive sampling frame depends on how much has been excluded. Bias results if the exclusion criterion is related to the outcome; e.g., excluding elements that are known to have high/low values of the outcome. On the other hand, the impact will be less pronounced if the exclusion criterion does not depend on the outcome.

Objectives:

It is important that these are stated clearly and are self-explanatory. Objectives should provide information on the targeted population, the outcome of interest, and the population statistic to be estimated and/or investigated.

Sampling Design:

The sampling design needs to be available. It is crucial for calculating the sample size, and making valid inferences. Practical consideration may require modifications to the well-known sampling designs. This is acceptable as long as such modifications are motivated, documented, and taken into account when drawing inferences. Without information regarding the sampling design, it is difficult to assess the representativeness of the data, and the validity of inferences drawn thereof.

Sample Size Calculations:

Sample size calculation follows from the carefully selected sampling design. It should be geared towards achieving the objectives, and desired quality aspects of the data like precision and accuracy.

Different objectives require different sample sizes (e.g., estimation versus hypothesis testing), hence it is not recommended to use the sample size calculated on a specific objective to achieve another objective. Allocation of the sample to the various countries should be done in the framework of stratification.

It is worth noting that one aspect connects to another, hence it is important to clearly define each of these.

Data Analysis:

Analyses of the data should reflect the design; for instance, the stratified mean should be employed if a stratified design was used.

Summarizing repeated measures into summary statistics or dichotomized values leads to loss of information, and may lead to biased estimates and invalid inferences. Such summaries can only be considered if there is enough evidence that there will not be much information loss, or if this is the best approach for the scenario at hand.

It is important that the various possible sources of bias are accorded due attention, to mitigate their effect. However, one of the limitations when addressing problems such as selection bias, and missing data bias, is the lack of information on the underlying mechanisms. It is therefore recommended to always conduct a sensitivity analysis so as to assess the stability of the results under different assumptions regarding the underlying mechanisms.

REFERENCES

Agresti A, 2002. Categorical Data Analysis (2nd ed.). New York: John Wiley & Sons.

Alonso A, Milanzi E, Molenberghs G, Buyck C, and Bijnens L, 2014. A new modeling approach for quantifying expert opinion in the drug discovery process. Submitted for publication.

Allison PD, 2010. "Survival analysis." Pp. 413-425 in The Reviewer's Guide to Quantitative Methods in the Social Sciences, edited by Gregory R. Hancock and Ralph O. Mueller. Routledge, New York.

Barnett V, 1991. Sample Survey: Principles and Methods. Arnold, London.

Brick JM and Kalton G, 1996. Handling missing data in survey research. Statistical Methods in Medical Research, 5, 215-238.

Burton A, Altman DG, Royston P and Holder RL, 2006. The design of simulation studies in medical statistics, 25, 4279-4292.

Chambers RL and Skinner CJ, 2003. Analysis of survey data. John Wiley & Sons, New York.

Cochran WG, 1977. Sampling techniques. John Wiley & Sons, New York.

Codex Alimentarius Commission ALINORM 08/31/31 September 2007. Joint FAO/WHO Food Standards Programme. Codex Alimentarius Commission. Report of the 17th Session of the Codex Committee on Residues of Veterinary Drugs in Foods. www.codexalimentarius.org/input/download/report/.../al31_31e.pdf

Concise Database summary statistics - Consumers only; http://www.efsa.europa.eu/en/datex/docs/datexfooddbstatistics2.xls (Download: 27-04-2014 19:39:14)

Concise Database summary statistics - Total population; http://www.efsa.europa.eu/en/datex/docs/datexfooddbstatistics1.xls (Download: 27-04-2014 19:38:54) 27

Crow EL and Shimizu K, 1988. Lognormal Distributions: Theory and Applications. Dekker, New York.

Duchateau L and Janssen P, 2008. The Frailty Model. Springer, New York.

European Food Safety Authority; Management of left-censored data in dietary exposure assessment of chemical substances. EFSA Journal 2010;8(3):.[96pp.]. doi: 10.2903/j.efsa.2010.1557.

European Food Safety Authority, 2013. Sample Size Considerations for Hierarchical Population. EFSA Journal 2013;11(7):3292, 47 pp. doi:10.2903/j.efsa.2013.3292

European Food Safety Authority; Standard sample description for food and feed. EFSA Journal 2010;8(1):1457 [54 pp.]. doi:10.2903/j.efsa.2010.1457.

European Food Safety Authority; The 2010 European Union Report on Pesticide Residues in Food. EFSA Journal 2013;11(3):3130. [808 pp.] doi:10.2903/j.efsa.2013.3130.

European Food Safety Authority; Use of the EFSA Standard Sample Description for the reporting of data on the control of pesticide residues in food and feed according to Regulation (EC) No 396/2005 (Revision 2). EFSA Journal 2013;11(1):3076.[54 pp.] doi:10.2903/j.efsa.2013.3076.

Eurostat: Population on 1 January; http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&init=1&plugin=1&language=en&pcode=tps00001 (Download: 17-12-2013 18:20:01)

Euostat, 2003. Methodological documents - definition of quality in statistics. Brussels.

Eurostat, 2009a. Handbook for quality reports. Brussels.

Faes C, Molenberghs G, Aerts M, Verbeke G and Kenward MG, 2009. The effective sample size and an alternative small-sample degrees-of-freedom method. The American Statistician, 63, 389-399.

Fitzmaurice GM, Laird NM, and Ware JH, 2004. Applied longitudinal analysis. Wiley-Interscience, Hoboken, NJ.

Fleiss JL, 1981. Statistical Methods for Rates and Proportions. Wiley, New York.

Foreman EK, 1991. Survey Sampling Principles. Dekker, New York.

Fowler FJ Jr, 1988. Survey research methods. Sage, Newbury Park, CA.

Genelletti S, Mason A, Best N, 2011. Adjusting for selection effects in epidemiological studies;

Why sensitivity analysis is the only solution". Commentary in Epidemiology, 22,36-39.

Ghosh S and Pahwa P. SSC Annual Meeting, May 2006. Design-based versus model-based methods: a comparative study using longitudinal survey data. Proceedings of the Survey Methods Section. www.ssc.ca/survey/documents/SSC2006_Sunita_Ghosh.pdf (Download 06-04-2014 21:33:53)

Groves RM, Fowler FJ Jr, Couper MP, Lepkowski JM, Singer E and Tourangeau R, 2004. Survey Methodology. John Wiley & Sons, New York.

Guidance Document for the use of the Concise European Food Consumption Database in Exposure Assessment; http://www.efsa.europa.eu/en/datexfoodcdb/datexfooddb.htm (Download: 27-04-2014 19:39:32)

Harrell FE Jr, Lee KL and Mark DB, 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 15, 361–387.

Harkness JA, Braun M, Edwards B, Johnson TP, Lyberg L, Mohler PPh, Pennell B-E and Smith TW, 2010. Survey Methods in Multinational, Multiregional, and Multicultural Contexts. Wiley, New Jersey.

Health Interview Survey 2013, Research Protocol. https://his.wiv-isp.be/Shared%20Documents/Protocol2013.pdf (Download: 03-01-2015 22:35:00)

Heeringa SG, West BT, and Berglund PA, 2010. Applied survey data analysis. Chapman & Hall/CRC, Boca Raton.

Kalton G, 1983. Introduction to Survey Sampling. Sage, California.

Klein JP and Moeschberger ML, 1997. Survival Analysis: Techniques for Censored and Truncated Data. Springer, New York.

Knottnerus P, 2003. Sample Survey Theory: Some Pythagorean Perspectives. Springer, New York.

Liang KY and Zeger SL, 1986. Longitudinal Data Analysis Using Generalized Linear Models. Biometrika, 73, 13-22.

Little RJA and Rubin DB, 2002. Statistical Analysis with Missing Data. Wiley, New York.

Lohr SL, 2009. Sampling: Design and Analysis. Brookes/Cole Cengage learning, Massachusetts.

Lunn AD and Davies SJ, 1998. A note on generating correlated binary variables, 82(2),487-490.

Molenberghs G and Verbeke G, 2005. Models for Discrete Longitudinal Data. Springer, New York.

Molenberghs G, Verbeke G, Demetrio C, Vieira A, 2010. A family of generalized linear models

for repeated measures with normal and conjugate random effects. Statistical Science,25, 325-347.

Metzger MJ, Brus DJ, Bunce RGH, Carey PD, Gonçalves J, Honrado JP, Jongman RHG, Trubacco A and Zomer R, 2012. Environmental stratifications as the basis for national, European and global ecological monitoring. Ecological Indicators. http://dx.doi.org/10.1016/j.ecolind.2012.11.009

Olsson U, 2005. Confidence intervals for the mean of a log-normal distribution. Journal of Statistics Education, 13; http://www.amstat.org/publications/jse/v13n1/olsson.html (Download: 01-12-2014 20:14:58)

Peduzzi P, Concato J, Feistein AR and Holford TR, 1995. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. Journal of Clinical Epidemiology 48, 1503—1510.

Population on 1 January by age and sex. http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_r_d2jan&lang=en (Download: 08-09-2014 14:49:00)

Population on 1 January by five years age groups and sex; http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_r_pjangroup&lang=en (Download: 08-09-2014 14:49:00)

Population on 1 January by broad age groups and sex; http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_r_pjanaggr3&lang=en (Download: 08-09-2014 14:49:00)

Rabe-Hesketh S, and Skrondal A, 2006. Multilevel modelling of complex survey data. Journal of Royal Statistical Society A, 169, 805-827.

Ramsey CA and Hewitt AD, 2005. A methodology for assessing sample representativeness. Environmental Forensics, 6, 71-75.

Reasoned Opinion on the Potential Chronic and Acute Risk to Consumers' Health Arising from Proposed Temporary EU MRLS. According to Regulation (EC) NO 396/2005 on Maximum Residue Levels of Pesticides in Food and Feed of Plant and Animal Origin. 15 March 2007.

Recommended methods of sampling for the determination of pesticide residues for compliance with MRLs. CAC/GL 33-1999. www.codexalimentarius.net/input/download/.../361/CXG_033e.pdf (Download: 10-10-2013 21:49:17)

Rubin DB (1976). Inference and missing data. Biometrika 63, 581–592.

Rizopoulos D, 2012. Joint Models for Longitudinal and Time-to-Event Data. Chapman and Hall/CRC, Boca Raton.

Schmidtke J and Schmidt K. Standardised statistical programming practices for R and SAS. Supporting Publications 2013:EN-390. [39 pp.]. Available online: www.efsa.europa.eu/publications

Shen H, Brown LD and Zhi H, 2006. Efficient estimation of log-normal means with application to pharmacokinetic data. Statistics in Medicine, 25, 3023-3038.

Som RK, 1996. Practical Sampling Techniques, 2nd Edition, Revised and Expanded. Dekker, New York.

Steel D and McLaren C. Design and Analysis of Repeated Surveys, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 11-08, 2008, 13p. http://ro.uow.edu.au/cssmwp/10

The SURVEYSELECT Procedure; http://www.math.wpi.edu/saspdf/stat/chap63.pdf (Download: 13-09-2014 11:41:00)

Van den Noortgate W, Opdenakker M and Onghena P, 2004. The Effects of Ignoring a Level in Multilevel Analysis. School Effectiveness and School Improvement, 16, 281-303.

Verbeke G and Molenberghs G, 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.

Working Group "Assessment of quality in statistics". Sixth meeting. Luxembourg, 2-3 October 2003 at 9 h 30. epp.eurostat.ec.europa.eu/portal/.../ess%20quality%20definition.pdf (Download: 18-12-2013 16:02:03)

Zhou XH, 1998. Estimation of the log-normal mean. Statistics in Medicine, 17, 2251-2264.

Zhou XH and Gao S, 1997. Confidence intervals for the log-normal mean. Statistics in Medicine, 16, 783-790.

APPENDICES

Appendix A More Results From the Case Study

Table 53 Estimated Variances for Commodities per Country: Apples

Country	Number of Samples	Proportion Exceeding	Variance
Austria	15	0	0
Belgium	15	0	0
Bulgaria	35	0	0
Cyprus	28	0.0714	0.0663
Czech	53	0.0377	0.0363
Germany	204	0	0
Denmark	72	0	0
Estonia	17	0	0
Spain	88	0.0227	0.0222
Finland	102	0	0
France	135	0.0074	0.0074
United Kingdom	143	0.0070	0.0069
Greece	90	0.0333	0.0322
Ireland	89	0	0
Iceland	16	0	0
Italy	56	0	0
Lithuan	20	0	0
Luxembourg	20	0.0500	0.0475
Latvia	29	0	0
Malta	15	0	0
Netherlands	132	0.0076	0.0075
Norway	18	0	0
Poland	61	0	0
Portugal	63	0.0794	0.0731
Romania	296	0.0203	0.0199
Sweden	149	0.0201	0.0197
Slovenia	76	0	0
Slovakia	20	0	0

Table 54 Estimated Variances for Commodities per Country: Pears

Country	Number of Samples	Proportion Exceeding	Variance
Czech	10	0	0
Spain	7	0	0
Finland	6	0	0
France	120	0.0083	0.0083
United Kingdom	149	0.0268	0.0261
Greece	26	0	0
Italy	1	0	0
Luxembourg	9	0	0
Norway	15	0	0
Slovenia	31	0	0
Slovakia	14	0	0

Table 55 Estimated Variances for Commodities per Country: Peaches

Country	Number of Samples	Proportion Exceeding	Variance
Austria	17	0	0
Belgium	15	0	0
Bulgaria	36	0	0
Cyprus	27	0.0370	0.0357
Czech	28	0.0714	0.0663
Germany	188	0.0160	0.0157
Denmark	53	0	0
Estonia	12	0	0
Spain	35	0.0286	0.0278
Finland	16	0	0
France	88	0.0114	0.0112
United Kingdom	148	0.0135	0.0133
Greece	61	0.0492	0.0468
Hungary	16	0	0
Ireland	20	0	0
Iceland	9	0	0
Italy	27	0	0
Lithuania	14	0	0
Luxembourg	15	0.0667	0.0622
Latvia	24	0	0
Malta	15	0.2000	0.1600
Netherlands	70	0	0
Norway	22	0	0
Poland	50	0.0200	0.0196
Portugal	33	0	0
Romania	56	0	0
Sweden	31	0.0645	0.0604
Slovenia	60	0.0167	0.0164
Slovakia	14	0.0714	0.0663

Table 56 Estimated Variances for Commodities per Country: Strawberries

Country	Number of Samples	Proportion Exceeding	Variance
Austria	15	0.0667	0.0622
Belgium	14	0	0
Bulgaria	31	0.0323	0.0312
Cyprus	27	0.1111	0.0988
Czech	18	0	0
Germany	199	0.0201	0.0197
Denmark	60	0.0167	0.0164
Estonia	24	0	0
Spain	32	0	0
Finland	50	0.0200	0.0196
France	97	0.0928	0.0842
United Kingdom	96	0	0
Greece	53	0.0377	0.0363
Hungary	15	0	0
Ireland	17	0	0
Iceland	5	0	0
Italy	30	0	0
Lithuania	19	0	0
Luxembourg	15	0	0
Latvia	22	0	0
Malta	14	0.0714	0.0663
Netherlands	97	0.0722	0.0670
Norway	19	0	0
Poland	49	0.0204	0.0200
Portugal	53	0	0
Romania	94	0	0
Sweden	34	0	0
Slovenia	60	0.0667	0.0622
Slovakia	13	0	0

Table 57 Estimated Variances for Commodities per Country: Tomatoes

Country	Number of Samples	Proportion Exceeding	Variance
Austria	16	0.1250	0.1094
Belgium	15	0	0
Bulgaria	37	0	0
Cyprus	29	0	0
Czech	51	0.0784	0.0723
Germany	193	0.0259	0.0252
Denmark	64	0	0
Estonia	17	0	0
Spain	106	0.0189	0.0185
Finland	47	0	0
France	122	0.0246	0.0240
United Kingdom	108	0	0
Greece	163	0	0
Hungary	17	0	0
Ireland	18	0	0
Iceland	15	0	0
Italy	67	0	0
Lithuania	14	0	0
Luxembourg	16	0	0
Latvia	27	0	0
Malta	18	0.0556	0.0525
Netherlands	130	0.0231	0.0225
Norway	24	0	0
Poland	50	0.0400	0.0384
Portugal	69	0	0
Romania	237	0	0
Sweden	47	0	0
Slovenia	60	0	0
Slovakia	17	0	0

Table 58 Estimated Variances for Commodities per Country: Head Cabbage

Country	Number of Samples	Proportion Exceeding	Variance
Austria	15	0.0667	0.0622
Belgium	15	0	0
Bulgaria	32	0	0
Czech	39	0.0256	0.0250
Germany	184	0	0
Denmark	24	0	0
Estonia	19	0	0
Spain	5	0	0
Finland	16	0	0
France	64	0.0469	0.0447
United Kingdom	72	0	0
Greece	27	0	0
Hungary	10	0	0
Ireland	16	0	0
Iceland	10	0	0
Lithuania	17	0.0588	0.0554
Luxembourg	14	0.0714	0.0663
Latvia	30	0	0
Malta	15	0	0
Netherlands	71	0	0
Norway	19	0	0
Poland	60	0	0
Portugal	63	0.0159	0.0156
Romania	99	0	0
Sweden	18	0	0
Slovenia	30	0	0
Slovakia	15	0.0667	0.0622

Table 59 Estimated Variances for Commodities per Country: Lettuce

Country	Number of Samples	Proportion Exceeding	Variance
Austria	15	0	0
Belgium	15	0	0
Bulgaria	29	0	0
Cyprus	27	0.1111	0.0988
Czech	40	0	0
Germany	175	0.0343	0.0331
Denmark	57	0	0
Estonia	13	0	0
Spain	46	0.0435	0.0416
Finland	47	0	0
France	312	0.0769	0.0710
United Kingdom	96	0	0
Greece	78	0.0641	0.0600
Hungary	14	0	0
Ireland	38	0.0263	0.0256
Iceland	8	0	0
Italy	17	0.0588	0.0554
Lithuania	14	0	0
Luxembourg	18	0	0
Latvia	27	0	0
Malta	15	0.0667	0.0622
Netherlands	156	0.0064	0.0064
Norway	21	0	0
Poland	50	0.0400	0.0384
Portugal	41	0.0244	0.0238
Romania	74	0.0270	0.0263
Sweden	35	0.0286	0.0278
Slovenia	75	0.0400	0.0384
Slovakia	15	0	0

Table 60 Estimated Variances for Commodities per Country: Leek

Country	Number of Samples	Proportion Exceeding	Variance
Austria	15	0.0667	0.0622
Belgium	15	0	0
Bulgaria	37	0	0
Cyprus	14	0	0
Czech	26	0.0385	0.0370
Germany	191	0.0157	0.0155
Denmark	22	0	0
Estonia	15	0	0
Spain	24	0	0
Finland	17	0	0
France	79	0	0
United Kingdom	96	0	0
Greece	28	0	0
Ireland	15	0	0
Iceland	7	0	0
Italy	13	0	0
Lithuania	15	0	0
Luxembourg	9	0	0
Latvia	25	0	0
Malta	15	0	0
Netherlands	56	0.0179	0.0175
Norway	22	0.0455	0.0434
Poland	50	0	0
Portugal	65	0.0462	0.0440
Romania	25	0	0
Sweden	25	0	0
Slovenia	25	0	0
Slovakia	15	0	0

Table 61 Estimated Variances for Commodities per Country: Oats

Country	Number of Samples	Proportion Exceeding	Variance
Austria	4	0	0
Belgium	5	0	0
Bulgaria	1	0	0
Czech	15	0	0
Denmark	11	0.0909	0.0826
Estonia	6	0	0
Spain	4	0	0
France	52	0	0
United Kingdom	80	0.1500	0.1275
Greece	3	0	0
Hungary	8	0	0
Ireland	22	0	0
Italy	3	0	0
Lithuania	4	0	0
Netherlands	1	0	0
Norway	9	0	0
Portugal	3	0	0
Slovenia	11	0	0
Slovakia	4	0	0

Table 62 Estimated Variances for Commodities per Country: Rye

Country	Number of Samples	Proportion Exceeding	Variance
Austria	9	0	0
Belgium	3	0	0
Bulgaria	5	0	0
Czech	36	0	0
Germany	92	0	0
Denmark	26	0	0
Estonia	7	0	0
Spain	5	0	0
Finland	29	0	0
France	31	0	0
United Kingdom	3	0	0
Greece	2	0	0
Hungary	7	0	0
Italy	1	0	0
Lithuania	12	0	0
Latvia	9	0	0
Netherlands	8	0	0
Norway	7	0	0
Poland	50	0	0
Portugal	4	0	0
Romania	11	0	0
Sweden	28	0	0
Slovenia	9	0	0
Slovakia	12	0.0833	0.0764

Table 63 Estimated Variances for Commodities per Country: Swine meat

Country	Number of Samples	Proportion Exceeding	Variance
Austria	16	0	0
Belgium	15	0	0
Cyprus	36	0	0
Germany	98	0	0
Denmark	120	0	0
Estonia	15	0	0
Finland	16	0	0
United Kingdom	108	0	0
Greece	15	0	0
Italy	2	0	0
Lithuania	8	0	0
Luxembourg	15	0	0
Latvia	16	0	0
Malta	15	0	0
Netherlands	20	0	0
Norway	15	0	0
Poland	47	0	0
Sweden	16	0	0
Slovenia	15	0	0
Slovakia	15	0	0

Table 64 Estimated Variances for Commodities per Country: Milk

Country	Number of Samples	Proportion Exceeding	Variance
Austria	17	0	0
Belgium	15	0	0
Cyprus	5	0	0
Germany	94	0	0
Denmark	15	0	0
Estonia	15	0	0
Spain	16	0	0
Finland	16	0	0
United Kingdom	235	0	0
Ireland	68	0	0
Lithuania	10	0	0
Luxembourg	18	0	0
Latvia	8	0	0
Netherlands	22	0	0
Norway	15	0	0
Poland	1	0	0
Romania	38	0	0
Sweden	30	0	0
Slovenia	1	0	0
Slovakia	15	0	0

Appendix B Allocation for Multistage Sampling

Table 65 Multistage Allocation for Design Effect=1.19, Margin of error=0.005.

Country	Oats	Rye	Head Cabbage	Leek	Lettuce	Tomatoes	Strawberries	Apples	Peaches	Pears	Swine meat	Milk	Total
Austria	6	6	2	2	2	2	2	2	2	2	6	6	40
Belgium	5	5	3	3	3	3	2	2	2	2	5	9	44
Bulgaria	4	4	2	2	2	2	1	1	1	1	4	5	29
Cyprus	1	1	1	1	1	1	1	1	1	1	1	1	12
Czech republic	6	6	2	2	2	2	2	2	2	2	8	8	44
Germany	44	44	19	19	19	19	15	15	15	15	51	95	370
Denmark	3	3	1	1	1	1	1	1	1	1	3	8	25
Estonia	1	1	1	1	1	1	1	1	1	1	1	2	13
Spain	14	14	6	6	6	6	4	4	4	4	19	16	103
Finland	2	2	1	1	1	1	1	1	1	1	3	9	24
France	38	38	13	13	13	13	8	8	8	8	49	63	272
United Kingdom	29	29	10	10	10	10	6	6	6	6	37	58	217
Greece	4	4	2	2	2	2	1	1	1	1	5	4	29
Hungary	5	5	2	2	2	2	2	2	2	2	7	10	43
Ireland	2	2	2	2	2	2	1	1	1	1	3	6	25
Iceland	1	1	1	1	1	1	1	1	1	1	1	1	12
Italy	31	31	14	14	14	14	12	12	12	12	31	48	245
Lithuania	1	1	1	1	1	1	1	1	1	1	2	2	14
Luxembourg	1	1	1	1	1	1	1	1	1	1	1	1	12
Latvia	1	1	1	1	1	1	1	1	1	1	1	1	12
Malta	1	1	1	1	1	1	1	1	1	1	1	1	12
Netherlands	7	7	3	3	3	3	2	2	2	2	10	24	68
Norway	2	2	1	1	1	1	1	1	1	1	2	10	24
Poland	22	22	11	11	11	11	10	10	10	10	37	26	191
Portugal	3	3	2	2	2	2	1	1	1	1	5	4	27
Romania	6	6	3	3	3	3	2	2	2	2	9	7	48
Sweden	5	5	2	2	2	2	2	2	2	2	6	14	46
Slovenia	1	1	1	1	1	1	1	1	1	1	1	1	12
Slovakia	4	4	1	1	1	1	1	1	1	1	4	2	22
Total													2035

Table 66 Multistage Allocation for Design Effect=1.14 and margin of error=0.0077.

Country	Oats	Rye	Head Cabbage	Leek	Lettuce	Tomatoes	Strawberries	Apples	Peaches	Pears	Swine Meat	Milk	total
Austria	3	3	1	1	1	1	1	1	1	1	3	3	20
Belgium	2	2	1	1	1	1	1	1	1	1	2	4	18
Bulgaria	2	2	1	1	1	1	1	1	1	1	2	2	16
Cyprus	1	1	1	1	1	1	1	1	1	1	1	1	12
Czech republic	3	3	1	1	1	1	1	1	1	1	3	3	20
Germany	18	18	8	8	8	8	6	6	6	6	21	39	152
Denmark	1	1	1	1	1	1	1	1	1	1	2	4	16
Estonia	1	1	1	1	1	1	1	1	1	1	1	1	12
Spain	6	6	3	3	3	3	2	2	2	2	8	7	47
Finland	1	1	1	1	1	1	1	1	1	1	1	4	15
France	16	16	6	6	6	6	4	4	4	4	20	26	118
United Kingdom	12	12	4	4	4	4	3	3	3	3	15	24	91
Greece	2	2	1	1	1	1	1	1	1	1	2	2	16
Hungary	2	2	1	1	1	1	1	1	1	1	3	4	19
Ireland	1	1	1	1	1	1	1	1	1	1	1	3	14
Iceland	1	1	1	1	1	1	1	1	1	1	1	1	12
Italy	13	13	6	6	6	6	5	5	5	5	13	19	102
Lithuania	1	1	1	1	1	1	1	1	1	1	1	1	12
Luxembourg	1	1	1	1	1	1	1	1	1	1	1	1	12
Latvia	1	1	1	1	1	1	1	1	1	1	1	1	12
Malta	1	1	1	1	1	1	1	1	1	1	1	1	12
Netherlands	3	3	2	2	2	2	1	1	1	1	4	10	32
Norway	1	1	1	1	1	1	1	1	1	1	1	4	15
Poland	9	9	5	5	5	5	4	4	4	4	15	11	80
Portugal	2	2	1	1	1	1	1	1	1	1	2	2	16
Romania	3	3	1	1	1	1	1	1	1	1	4	3	21
Sweden	3	3	1	1	1	1	1	1	1	1	3	6	23
Slovenia	1	1	1	1	1	1	1	1	1	1	1	1	12
Slovakia	2	2	1	1	1	1	1	1	1	1	2	1	15
Total													962

Table 67 Multistage Allocation for Design Effect=1.7 and margin of error=0.0077.

Country	Oats	Rye	Head Cabbage	Leek	Lettuce	Tomatoes	Strawberries	Apples	Peaches	Pears	Swine meat	Milk	Total
Austria	4	4	1	1	1	1	1	1	1	1	4	4	24
Belgium	3	3	2	2	2	2	1	1	1	1	3	5	26
Bulgaria	3	3	1	1	1	1	1	1	1	1	2	3	19
Cyprus	1	1	1	1	1	1	1	1	1	1	1	1	12
Czech republic	4	4	1	1	1	1	1	1	1	1	5	5	26
Germany	26	26	12	12	12	12	9	9	9	9	31	57	224
Denmark	2	2	1	1	1	1	1	1	1	1	2	5	19
Estonia	1	1	1	1	1	1	1	1	1	1	1	1	12
Spain	8	8	4	4	4	4	2	2	2	2	12	10	62
Finland	1	1	1	1	1	1	1	1	1	1	2	6	18
France	23	23	8	8	8	8	5	5	5	5	29	38	165
United Kingdom	18	18	6	6	6	6	4	4	4	4	23	35	134
Greece	2	2	1	1	1	1	1	1	1	1	3	3	18
Hungary	3	3	2	2	2	2	1	1	1	1	5	6	29
Ireland	2	2	1	1	1	1	1	1	1	1	2	4	18
Iceland	1	1	1	1	1	1	1	1	1	1	1	1	12
Italy	19	19	9	9	9	9	7	7	7	7	19	29	150
Lithuania	1	1	1	1	1	1	1	1	1	1	1	1	12
Luxembourg	1	1	1	1	1	1	1	1	1	1	1	1	12
Latvia	1	1	1	1	1	1	1	1	1	1	1	1	12
Malta	1	1	1	1	1	1	1	1	1	1	1	1	12
Netherlands	5	5	2	2	2	2	1	1	1	1	6	15	43
Norway	2	2	1	1	1	1	1	1	1	1	2	6	20
Poland	14	14	7	7	7	7	6	6	6	6	22	16	118
Portugal	2	2	1	1	1	1	1	1	1	1	3	3	18
Romania	4	4	2	2	2	2	1	1	1	1	5	5	30
Sweden	4	4	1	1	1	1	1	1	1	1	4	8	28
Slovenia	1	1	1	1	1	1	1	1	1	1	1	1	12
Slovakia	3	3	1	1	1	1	1	1	1	1	3	2	19
Total													1304

Appendix C Dealing with Missing Data in the Analysis of the MACPS

Table 68 MRL Non-Compliance by "country of origin – food item" cluster

Cluster	MRL Non-Compliance Probability
MA,Peaches	.005432881
GR,Lettuce	.004915654
CY,Peaches	.004070267
CY,Strawberries	.003513229
CY,Lettuce	.003513229
CY,Apples	.003104947
EG,Strawberries	.002482097
RO,Lettuce	.002125042
ES,Leek	.001581729
PT,Lettuce	.001471401
NZ,Apples	.001306602
FR,Strawberries	.000996946
ZA,Peaches	.000876351
GR,Peaches	.000625250
ES,Tomatoes	.000530767
CL,Apples	.000459893
ES,Peaches	.000134166
GR,Rye	.000001852
BG,Oats	.000001851
CH,Apples	.000001851
IS,Lettuce	.000001851
RS,Head cabbage	.000001851
RS,Tomatoes	.000001851
XD,Apples	.000001851
CL,Tomatoes	.000001851
CN,Tomatoes	.000001851
EC,Apples	.000001851
EC,Strawberries	.000001851
EU,Tomatoes	.000001851
IT,Rye	.000001851
ZA,Tomatoes	.000001851
EU,Peaches	.000001851

Cluster	MRL Non-Compliance Probability
GF,Tomatoes	.000001851
GP,Lettuce	.000001851
PL,Milk and milk products	.000001851
TR,Apples	.000001851
TN,Peaches	.000001851
CH,Tomatoes	.000001851
HU,Swine meat	.000001851
IT,Swine meat	.000001851
RE,Apples	.000001851
RS,Peaches	.000001851
TH,Head cabbage	.000001851
XD,Tomatoes	.000001851
EG,Leek	.000001851
CR,Tomatoes	.000001851
EU,Head cabbage	.000001851
EU,Lettuce	.000001851
FI,Oats	.000001851
LU,Tomatoes	.000001851
MK,Lettuce	.000001851
MK,Peaches	.000001851
MT,Apples	.000001851
PL,Oats	.000001851
TR,Lettuce	.000001851
AL,Head cabbage	.000001851
AR,Lettuce	.000001851
CZ,Peaches	.000001851
ET,Strawberries	.000001851
EU,Oats	.000001851
EU,Rye	.000001851
HK,Strawberries	.000001851
HR,Strawberries	.000001851
MA,Head cabbage	.000001851
NL,Peaches	.000001851
PS,Tomatoes	.000001851
SN,Lettuce	.000001851

Cluster	MRL Non-Compliance Probability
TR,Leek	.000001851
US,Strawberries	.000001851
ZA,Lettuce	.000001851
IS,Head cabbage	.000001850
JO,Tomatoes	.000001850
MD,Tomatoes	.000001850
ES,Swine meat	.000001850
EU,Swine meat	.000001850
IE,Swine meat	.000001850
SI,Milk and milk products	.000001850
TR,Head cabbage	.000001850
BE,Oats	.000001850
EU,Milk and milk products	.000001850
CY,Leek	.000001850
BE,Rye	.000001850
AT,Strawberries	.000001849
MD,Apples	.000001849
EE,Tomatoes	.000001849
SK,Lettuce	.000001849
CZ,Leek	.000001849
GB,Rye	.000001849
AL,Lettuce	.000001849
DE,Peaches	.000001849
HR,Apples	.000001849
NI,Leek	.000001849
NO,Rye	.000001849
SY,Tomatoes	.000001849
IT,Oats	.000001849
MK,Apples	.000001849
CY,Milk and milk products	.000001848
IS,Tomatoes	.000001848
AL,Tomatoes	.000001848
DO,Tomatoes	.000001848
ES,Milk and milk products	.000001848
LT,Leek	.000001848

Cluster	MRL Non-Compliance Probability
LU,Head cabbage	.000001848
LU,Leek	.000001848
MQ,Lettuce	.000001848
NL,Rye	.000001848
SK,Oats	.000001848
SK,Strawberries	.000001848
AT,Oats	.000001847
AR,Peaches	.000001847
IE,Tomatoes	.000001847
RS,Strawberries	.000001847
ES,Oats	.000001847
XX,Oats	.000001847
IL,Peaches	.000001847
BG,Rye	.000001846
EE,Apples	.000001846
CN,Strawberries	.000001846
HU,Apples	.000001846
DK,Oats	.000001846
EE,Oats	.000001846
LT,Lettuce	.000001846
CZ,Tomatoes	.000001846
IL,Strawberries	.000001846
SI,Oats	.000001846
RE,Tomatoes	.000001846
AT,Tomatoes	.000001845
FI,Leek	.000001845
EE,Rye	.000001845
LT,Oats	.000001845
LV,Strawberries	.000001845
SK,Tomatoes	.000001845
CZ,Strawberries	.000001844
ES,Rye	.000001844
HU,Tomatoes	.000001844
PL,Peaches	.000001843
UY,Apples	.000001843

Cluster	MRL Non-Compliance Probability
IT,Head cabbage	.000001843
LT,Tomatoes	.000001843
LV,Apples	.000001843
SE,Oats	.000001843
EG,Tomatoes	.000001842
HU,Peaches	.000001842
IE,Apples	.000001842
FI,Head cabbage	.000001842
DE,Oats	.000001842
SN,Tomatoes	.000001842
EE,Leek	.000001841
LT,Strawberries	.000001841
LV,Leek	.000001841
LV,Lettuce	.000001841
SK,Head cabbage	.000001841
TN,Tomatoes	.000001841
TR,Peaches	.000001841
HU,Rye	.000001841
IE,Leek	.000001841
SI,Rye	.000001841
SI,Tomatoes	.000001841
FR,Swine meat	.000001841
AT,Leek	.000001840
MA,Strawberries	.000001840
LT,Apples	.000001840
LU,Apples	.000001840
LV,Tomatoes	.000001840
LV,Rye	.000001839
HU,Oats	.000001839
NO,Tomatoes	.000001839
SE,Head cabbage	.000001839
PT,Peaches	.000001839
LU,Lettuce	.000001838
AT,Rye	.000001838
NO,Oats	.000001837

Cluster	MRL Non-Compliance Probability
RO,Rye	.000001837
FI,Rye	.000001837
LV,Head cabbage	.000001837
DK,Leek	.000001836
HU,Strawberries	.000001836
SE,Leek	.000001836
SI,Leek	.000001836
AT,Lettuce	.000001835
IT,Leek	.000001835
SK,Rye	.000001834
NO,Apples	.000001833
FI,Lettuce	.000001832
EE,Lettuce	.000001832
LT,Rye	.000001832
MT,Peaches	.000001832
MT,Strawberries	.000001832
AT,Head cabbage	.000001832
LV,Milk and milk products	.000001831
MK,Tomatoes	.000001831
XX,Head cabbage	.000001831
IE,Strawberries	.000001831
XX,Swine meat	.000001831
MT,Leek	.000001831
XX,Rye	.000001830
IE,Head cabbage	.000001829
NO,Leek	.000001829
NO,Lettuce	.000001829
NO,Strawberries	.000001829
CY,Tomatoes	.000001829
MT,Head cabbage	.000001829
MT,Lettuce	.000001829
XX,Lettuce	.000001829
MT,Swine meat	.000001829
BG,Strawberries	.000001829
FI,Apples	.000001828

Cluster	MRL Non-Compliance Probability
HU,Head cabbage	.000001828
CZ,Oats	.000001828
DK,Head cabbage	.000001828
SE,Tomatoes	.000001828
EE,Head cabbage	.000001828
LT,Head cabbage	.000001828
RO,Milk and milk products	.000001828
SK,Apples	.000001828
LT,Milk and milk products	.000001827
AT,Swine meat	.000001827
RO,Leek	.000001827
GB,Tomatoes	.000001826
CZ,Rye	.000001826
AT,Apples	.000001825
FI,Tomatoes	.000001825
SE,Lettuce	.000001825
XX,Strawberries	.000001825
EG,Peaches	.000001825
MT,Tomatoes	.000001825
LT,Swine meat	.000001824
NO,Head cabbage	.000001824
IL,Tomatoes	.000001824
HU,Lettuce	.000001824
GR,Swine meat	.000001823
TR,Strawberries	.000001823
BG,Apples	.000001822
GR,Leek	.000001822
AT,Milk and milk products	.000001821
CN,Apples	.000001821
GR,Head cabbage	.000001821
LU,Swine meat	.000001821
BG,Head cabbage	.000001821
EE,Strawberries	.000001820
FI,Swine meat	.000001820
SI,Swine meat	.000001820

Cluster	MRL Non-Compliance Probability
CZ,Lettuce	.000001820
SI,Head cabbage	.000001820
SI,Peaches	.000001820
XX,Peaches	.000001819
BG,Lettuce	.000001819
LV,Swine meat	.000001819
BG,Tomatoes	.000001818
EE,Milk and milk products	.000001818
EE,Swine meat	.000001818
BE,Milk and milk products	.000001818
BE,Head cabbage	.000001817
RO,Peaches	.000001816
MK,Head cabbage	.000001816
PL,Lettuce	.000001816
NO,Milk and milk products	.000001815
NO,Swine meat	.000001815
SK,Milk and milk products	.000001815
SK,Swine meat	.000001815
DK,Lettuce	.000001815
SE,Strawberries	.000001815
LU,Milk and milk products	.000001814
BG,Peaches	.000001814
ES,Head cabbage	.000001814
IE,Oats	.000001814
DK,Milk and milk products	.000001814
DK,Rye	.000001813
FI,Strawberries	.000001813
FI,Milk and milk products	.000001813
FR,Rye	.000001813
BE,Swine meat	.000001812
BG,Leek	.000001812
SE,Swine meat	.000001811
DK,Strawberries	.000001810
US,Apples	.000001810
CZ,Apples	.000001809

Cluster	MRL Non-Compliance Probability
IE,Lettuce	.000001807
CZ,Head cabbage	.000001807
SE,Rye	.000001807
CY,Swine meat	.000001806
XX,Tomatoes	.000001805
DK,Tomatoes	.000001804
DE,Tomatoes	.000001803
DK,Apples	.000001802
BE,Apples	.000001802
SI,Strawberries	.000001801
PL,Rye	.000001801
PT,Strawberries	.000001797
SE,Apples	.000001796
XX,Leek	.000001796
NL,Milk and milk products	.000001794
GR,Strawberries	.000001794
IT,Strawberries	.000001793
PT,Apples	.000001793
FR,Oats	.000001792
PL,Leek	.000001791
ZA,Apples	.000001790
GB,Apples	.000001789
PL,Tomatoes	.000001787
PT,Tomatoes	.000001786
BE,Tomatoes	.000001785
XX,Apples	.000001785
ES,Apples	.000001785
SE,Milk and milk products	.000001784
SI,Lettuce	.000001784
FR,Head cabbage	.000001784
BE,Strawberries	.000001782
NL,Swine meat	.000001782
PT,Leek	.000001779
PL,Strawberries	.000001778
GB,Lettuce	.000001777

Cluster	MRL Non-Compliance Probability
FR,Tomatoes	.000001775
CL,Peaches	.000001774
PT,Head cabbage	.000001772
GB,Oats	.000001771
RO,Strawberries	.000001768
GB,Head cabbage	.000001768
RO,Head cabbage	.000001766
GB,Strawberries	.000001765
PL,Head cabbage	.000001765
AR,Apples	.000001764
BR,Apples	.000001763
FR,Peaches	.000001757
TR,Tomatoes	.000001757
MA,Tomatoes	.000001757
NL,Apples	.000001749
GR,Apples	.000001749
PL,Apples	.000001748
FR,Leek	.000001744
PL,Swine meat	.000001742
GB,Leek	.000001739
SI,Apples	.000001737
BE,Lettuce	.000001731
IT,Lettuce	.000001726
IT,Tomatoes	.000001725
DE,Rye	.000001724
NL,Head cabbage	.000001722
IE,Milk and milk products	.000001720
GR,Tomatoes	.000001719
NL,Lettuce	.000001717
NL,Strawberries	.000001717
DE,Leek	.000001703
BE,Leek	.000001686
DE,Lettuce	.000001685
RO,Tomatoes	.000001677
DE,Apples	.000001676

Cluster	MRL Non-Compliance Probability
GB,Swine meat	.000001673
DE,Milk and milk products	.000001666
DE,Swine meat	.000001656
NL,Leek	.000001652
DK,Swine meat	.000001647
DE,Strawberries	.000001646
ES,Strawberries	.000001631
IT,Apples	.000001624
DE,Head cabbage	.000001602
RO,Apples	.000001601
FR,Apples	.000001565
FR,Lettuce	.000001556
NL,Tomatoes	.000001551
ES,Lettuce	.000001544
IT,Peaches	.000001536
GB,Milk and milk products	.000001485