



Food and Agriculture
Organization of the
United Nations



World Health
Organization

Viale delle Terme di Caracalla, 00153 Rome, Italy - Tel: (+39) 06 57051 - E-mail: codex@fao.org - www.codexalimentarius.org

Agenda Item 14

CX/CF 25/18/15
June 2025

JOINT FAO/WHO FOOD STANDARDS PROGRAMME

CODEx COMMITTEE ON CONTAMINANTS IN FOODS

Eighteenth Session
23-27 June 2025
Bangkok, Thailand

GUIDANCE ON DATA ANALYSIS FOR THE DEVELOPMENT OF MAXIMUM LEVELS AND FOR IMPROVED DATA COLLECTION

(Prepared by the Electronic Working Group chaired by the European Union
and co-chaired by Japan, The Netherlands, and The United States of America)

BACKGROUND

1. At the 12th session¹, CCCF considered the proposal of the JECFA Secretariat to develop a general guidance on data analysis for maximum level (ML) development as it was observed that different approaches were taken by the EWGs. These differences concerned for example the handling of occurrence data without information on limit of quantification (LOQ). A general guidance would help future EWGs to take consistent approaches for data analysis. CCCF agreed to establish an EWG chaired by EU, co-chaired by the United States of America, the Netherlands and Japan, working in English, to prepare a discussion paper.
2. At its 13th session², the EU as chair of the EWG, informed the CCCF that it has not been possible to prepare in time a discussion paper for consideration by the established EWG. Therefore, a paper prepared by the EU as Chair of the EWG containing a non-exhaustive list of topics that could be considered to be covered by the general guidance on data analysis for ML development was presented and CCCF agreed to extend the scope of the work to address improved data collection.
3. At the 14th session³, CCCF agreed that the work should be focused on data collection, data analysis and data presentation as a priority and that discussion on elements for consideration such as appropriate rejection rates would not be taken up and CL 2021/78 CF⁴, with the Annex to CX/CF 21/14/15 in annex, was circulated in October 2021 with the request for comments on the guidance on data analysis for development of maximum levels and for improved data collection.
4. CCCF15⁵ agreed on the status, goals/objectives and target user to be outlined in the Preamble and on the structure and content of the guidance document, with the understanding that further fine-tuning might be needed following the discussion in the EWG.
5. CCCF16⁶, agreed on the changes in the GEMS/Food database template as recommendations to the GEMS/Food administrator for review. After receiving feedback from the GEMS/Food database administrator on which of the recommendations can be effectively implemented and on the timeframe of their implementation, the section “data collection and submission and data extraction” would need to be updated.
6. However, the GEMS/Food administrator was too late consulted by the Chair of the EWG and therefore the section “Data collection and submission” could not be updated based on the feedback, circulated for comments to the EWG prior to CCCF17 and submitted to CCCF17 for finalisation.

¹ REP18/CF, paras 155-156

² REP19/CF, paras 156-165

³ REP21/CF, paras 186-210

⁴ <https://www.fao.org/fao-who-codexalimentarius/resources/circular-letters/en/>

⁵ REP22/CF15, paras 202-208

⁶ REP23/CF16 para 98 (i) and (ii)

7. CCCF17⁷ agreed on the feedback to be provided to the GEMS/Food database administrator based on the preliminary comments on the changes proposed by CCCF16 on the GEMS/Food database template and agreed on the proposed new work procedure for finalizing the Guidance on data analysis for development of MLs and for improved data collection.
8. As the guidance on data analysis for development of MLs and for improved data collection is being developed for internal CCCF working procedure, this work could be undertaken in a pre-session working group (which could operate in a physical or virtual mode) or an in-session working group (hereafter referred to as “the WG”). Consideration would be given to convening VWG meetings to discuss and advance the work on certain sections of the guidance document. The outcome of the VWGs would be circulated for comments via a CL and the comments would be discussed in a pre-session WG (virtual or physical) or an in-session WG. The Codex EWG Platform would also be used to facilitate this work.
9. At CCCF17⁸, it was discussed and agreed that the work prior to CCCF18 would focus on:
 - (i) finalizing the modifications to the GEMS/Food database template and related guidance. It was foreseen to organize one Virtual Working Group (VWG) meeting to discuss on data collection and submission and data extraction; and
 - (ii) the discussion on the structure and the content of the main document for the section data selection/clean up and statistical data analysis with a decision on which of the more complex issues were to be addressed in the future in separate Annex(es) to the main document. It was foreseen to organize one other VWG to discuss this part.
10. Following the discussions at the VWG, it could be considered if one round of comments within the Codex EWG Platform would be appropriate before the main document on these parts would be finalized for circulation as a CL to gather comments.
11. CCCF17⁹ agreed with the intention to produce a document that would provide practical guidance to the EWG performing data analysis for the development of MLs, for consideration by CCCF18. CCCF17 also agreed that in the future, the more complex issues identified would be addressed in separate annexes that would be developed and discussed post-CCCF18, following completion of the main guidance, together with the issues identified for future discussion.
12. A first VWG meeting took place on the part “data collection, data submission and data extraction” of the guidance on data analysis for development of maximum levels and for improved data collection on 31 October 2024, with the United States as VWG chair. The comments made at the VWG meeting by the participants and the additional comments posted on the Codex Forum Platform by Japan were taken into account in the version of the guidance available in Appendix I.
13. A second VWG meeting took place on the combined parts “data selection/clean-up of data” and “statistical analysis” of the guidance on data analysis for development of maximum levels and for improved data collection on 26 February 2025, with The Netherlands and Japan as VWG Chairs. The comments made at the VWG meeting by the participants and the additional comments posted on the Codex Forum Platform by Thailand, USA, Singapore, Brazil and International Council of Beverages Associations (ICBA) were taken into account in the version of the guidance available in Appendix I.
14. The current fields for individual analysis results in the GEMS/Food database template are listed in Guidance Table 1 of the guidance document in Appendix I. The GEMS/Food Database Administrator informed the Chair of the WG that the agreed changes to the GEMS/Food database template could not be implemented prior to CCCF18. Therefore, the bracketed text (whether single or double brackets) indicates changes that have been agreed on, but not yet implemented. Blue text in single brackets indicates edits based on changes that could be implemented in a first phase (timing to be made available at CCCF18). Purple text in double brackets indicates edits that could potentially be made at a later stage.

POINTS FOR DISCUSSION

15. Discussion on comments received prior and during CCCF18 on the main part of the Guidance on data analysis for development of MLs and for improved data collection as provided in Appendix I.
16. Discussion on the roles of JECFA and CCCF as risk assessor and risk manager, respectively, in the calculation of dietary exposure reduction rates when considering MLs and to which extent preliminary exposure assessment from the target commodity can also be conducted as reference information in EWGs tasked with recommending MLs if resources are available (paragraphs 156 and 157 and Annex IV of the Guidance in Appendix I)
17. Discussion on further work on the Annexes to the Guidance as provided in Appendix I and on the issues identified for future discussion provided in Appendix II. In this discussion the merging of Annex III ‘Drawing charts/graphs and plots on distribution of occurrence data’ with Annex V ‘Presentation of data analysis/statistical analysis’ should be considered, given that the visualisation of the distribution of occurrence data is an important aspect of presentation of occurrence data.

⁷ REP24/CF17, paras 140, 141 and 145

⁸ REP24/CF17, para 142

⁹ REP24/CF17, para 144

CONCLUSION

18. As the guidance is developed for internal CCCF working procedure, the main part of the guidance on data analysis for development of MLs and for improved data collection could be endorsed at CCCF18 after having taken into account comments received prior and during CCCF18, with the understanding that the document will be updated following the implementation of the agreed changes to the GEMS/Food database template and experience gained in EWGs in applying the guidance.

RECOMMENDATIONS

19. Based on the topics raised in the points for discussion (paragraphs 15 - 17) and the conclusion (paragraph 18), CCCF18 is invited to:
- (i) endorse the main part of the Guidance on data analysis for development of maximum levels and for improved data collection provided in Appendix I;
 - (ii) clarify the roles of JECFA and CCCF as risk assessor and risk manager, respectively, in the calculation of dietary exposure reduction rates when considering MLs and to which extent preliminary exposure assessment from the target commodity can also be conducted as reference information in EWGs tasked with recommending MLs if resources are available (see also 3rd indent of Appendix II);
 - (iii) consider further work on the Annexes to the main part of the Guidance provided in Appendix I;
 - Whether/which Annexes are needed.
 - Whether/which Annexes could be combined.
 - Determination of the timeline for further work on the Annexes.
 - (iv) consider further work on the issues identified for future discussion provided in Appendix II.
 - Determination of the issues on which further discussion is appropriate.
 - Determination of the order of priority of the issues for further work.
 - Determination of the timeline for further work on these issues

LIST OF APPENDICES

- APPENDIX I: Guidance on data analysis for development of maximum levels and for improved data collection
- APPENDIX II: Issues identified for possible future discussion
- APPENDIX III: List of Members of the Codex Forum Platform Group “CCCF18 - Guidance on data analysis for development of maximum levels and for improved data collection”

APPENDIX I

GUIDANCE ON DATA ANALYSIS FOR THE DEVELOPMENT OF MAXIMUM LEVELS AND FOR IMPROVED DATA COLLECTION

TABLE OF CONTENTS

	Paragraphs
Preamble	paras 1-7
Data collection and submission	paras 8-16
Filling out the GEMS/Food template	paras 17-52
Data extraction	paras 53-57
Data Selection /clean-up of data	
General considerations	para 58
Origin of data	paras 59-60
Clean-up of data	paras 61-65
Handling of data	
Missing information	paras 66-73
Incorrect information	paras 74-77
Data for which information on the portion analysed is not clear	para 78
Data originating from suspected fraudulent/ economically adulterated samples	para 79
Data from targeted sampling	paras 80-81
Determination of outliers/extreme values and handling them	paras 82-92
Limit of Quantification (LOQ) and Limit of Detection (LOD) considerations	
Exclusion based on LOQ	paras 93-96
Sum of components and LOQ	para 97
Data analysis: generating overview of data	
Overview which countries, number of data points, period of data coverage	para 98
Geographical coverage of the provided occurrence data	paras 99-104
Time period coverage of the provided occurrence data	paras 105-110
Statistical analysis of occurrence data and handling of datasets for ML development	
General considerations	paras 111-113
Minimum number of data points for estimating high percentile values	para 114
Handling datasets	
Handling datasets with low number of data points	paras 115-119
Handling datasets where available on individual commodity(ies)/food(s) are insufficient, but data for the food group are sufficient	paras 120-123
Handling of datasets (including use of substitution methods) with a large proportion of left-censored data	paras 124-126
Substitution methods for datasets with a large proportion of left-censored data	paras 127-130

Handling of multiple datasets decision

Analysis of individual and combined datasets and making decision on the necessity of comparing individual datasets before combining, especially when distribution patterns are significantly different. paras 131-136

Cases where the analysis of individual datasets is recommended paras 137-140

Conducting data analysis by visualization para 141**Data aggregation and calculation of descriptive statistics** paras 142-144**Calculation of rejection rates at hypothetical MLs**

Estimation of hypothetical MLs paras 145-148

Calculation of rejection rates at the hypothetical MLs paras 149-150

Assessment of impact of an ML on rejection rate paras 151-152

Improvement of calculation of rejection rates paras 153-155

Preliminary calculation of effects of MLs on the reduction of dietary exposure to the contaminant from the target commodity(ies) at hypothetical MLs

Calculation of dietary exposure and reduction at hypothetical MLs paras 156

Improvement of calculation of exposure reduction rates para 157

Preparing final recommendations to CCCF paras 158-164**Data presentation in EWG reports to CCCF**

Presentation of data analysis: statistical analysis paras 165-166

ANNEXES**ANNEX I: Determination of outliers/extreme values and handling them****ANNEX II: Minimum number of data points for estimating high percentile values****ANNEX III: Drawing charts/graphs and plots on distribution of occurrence data****[ANNEX IV: Preliminary calculation of dietary exposure and its reduction rate to the contaminants from target commodity(ies) at the hypothetical MLs]****ANNEX V: Presentation of data analysis/statistical analysis****ANNEX VI: Glossary of terms**

PREAMBLE

1. The steps in development of a maximum level (ML) can include:
 - Identification of a new health or trade issue relevant to a contaminant – commodity/food combination
 - Development of a discussion paper that explores preliminary occurrence data, exposure data, and global significance of the contaminant-commodity/food combination.
 - Agreement by CCCF to begin a new work, including discussion of Terms of Reference (TOR) of a working group, and submission of a proposal for new work to the CAC
 - Development of a document recommending MLs in the Codex step process and a more in-depth analysis of occurrence data, exposure data, global significance of the contaminant-commodity/food combination, and impact of proposed MLs.
 - Recommendation to send MLs to the CAC for adoption.
2. The primary data source for CCCF is GEMS/Food, an international database run by the World Health Organization, containing data on contaminant levels in different foods. Member countries are encouraged to submit data from their national monitoring programs either on a routine basis or in response to calls for data from CCCF; the data must meet certain criteria for submission (such as including a limit of quantification (LOQ) or limit of detection (LOD) for non-quantified data). External data, such as data from scientific literature, may be referenced, but are typically not used in setting MLs.
3. Prior to starting work on a discussion paper or ML document, CCCF shall establish TOR for the working group and may request Codex secretariat to issue a Call for Data. As outlined in the CAC Procedural Manual, 30th Ed., the TOR shall clearly state the objective(s) to be achieved by the establishment of the working group, the language(s) to be used, and the time frame by which the work is expected to be completed. The Call for Data typically identifies the contaminant and food/commodities of interest and the date range of requested data. Previous Calls for Data have also asked for information such as the LOQ and LOD of the analytical method and specific sample names; they also have identified fields in the GEMS/Food database where information should be entered and identified the appropriate basis of results.
4. Establishing TOR and planning the scope of a Call for Data is an important step in the data collection process. Careful attention to the TOR and Call for Data will result in better quality data for use in establishing MLs.
5. The management of data play a key role in the work of elaborating maximum levels, and it is of common interest to have data of good quality (such as data that reflect an accurate picture of the contamination of food presented with statistical analysis). Occurrence data should ideally be obtained through statistically based sampling (Ref. CXG 50-2004 General Guidelines on Sampling), and analysis should be conducted using validated methods with appropriate LOQ and LOD in laboratories that have quality assurance systems. Ideally, data submitted by member countries should be nationally representative.
6. The aim of this guidance document is to provide the elements for ensuring good quality data and to ensure a harmonised use and analysis of the available occurrence data by different EWGs for the development/elaboration of Codex MLs.
7. This guidance is for internal use in CCCF but national/regional authorities may use the relevant information contained in this guidance document for the development/elaboration of national/regional MLs.

DATA COLLECTION AND SUBMISSION

8. The introductory page for the WHO GEMS/Food database is [Global Environment Monitoring System \(GEMS\) / Food Contamination Monitoring and Assessment Programme](#). The data submission (upload) and data extraction (download) process begins at the website, [GEMS/Food Contaminants Database](#).
9. The database page opens to a Welcome page with two tabs, a Home Page tab and a Search tab. For full functionality, members must register and log in to their accounts. After logging in, the data submitter will have access to an Upload tab, in addition to the Home Page tab and Search tab. The submitter will also be able to access regular and bulk templates for uploading data, the GEMS/Food e-learning tool, and useful links such as Frequently Asked Questions.

10. Prior to submitting data, submitters should review materials on the GEMS/Food internet site ([Nutrition and Food Safety \(who.int\)](http://Nutrition and Food Safety (who.int))) or linked GEMS/Food pages. Detailed instructions are found in the document, INSTRUCTIONS FOR ELECTRONIC SUBMISSION OF DATA ON CHEMICALS IN FOOD AND THE DIET (hereafter referred to as INSTRUCTIONS) on the GEMS/Food internet site. This document provides instructions on registering an account, logging into the GEMS/Food database, inserting data into the Excel template, and uploading the Excel template. Familiarity with Excel is very helpful.
11. Data can be submitted to the GEMS/Food database on any food at any time, not just in response to a Call for Data specifying specific foods or time periods of interest. If data are submitted in response to a specific Call for Data, consider noting this information in the Remarks field. Data that fall outside the date frame referenced in a Call for Data can also be submitted. These data may be informative for study of contaminant levels over time.
12. If questions arise about technical aspects of data submissions, the submitter should contact the GEMS/Food database administrator. Questions could include error messages on upload, registration problems, how to name samples, what fields are mandatory, the definition of fields, problems with mapping, etc.
13. If questions arise about whether data align with a specific Call for Data, the submitter should consult the EWG Chair and, if needed, the Codex or JECFA Secretariats. Questions could include whether the samples correspond to the definitions provided in the Call for Data or the TOR of the EWG.
14. Data submitters should develop and retain metadata associated with data submissions. The metadata will help answer questions that might arise from the EWG. Metadata could include the year of production, the overall LOD and LOQ range associated with a data set, information on product labels, information on location of collection (e.g., import or retail), names of staff who submitted the data and when the data were submitted, the batch ID associated with the submitted dataset, etc.
15. Data quality should be assessed by the submitter before data are uploaded to GEMS/Food. If serious questions arise about data quality (missing information, suspect analyses), do not submit the data until these questions can be addressed.
16. If the submitter identifies a problem with a dataset after submission, consult with the GEMS/Food administrator on withdrawing or correcting the dataset, which should be identifiable by batch ID.

Filling out the GEMS/Food template

17. The template worksheet for regular (non-bulk) submissions¹ contains five tabs, which include (1) a checklist for submitting institutions, (2) Food Mapping of the sample, (3) a template for Individual Analysis results, (4) the WHO and FoodEx2 classification system, and (5) chemicals currently listed as options for submission in a drop-down menu.
18. The first step when submitting data is to fill out Tab "1. Start", which contains a checklist for the Institution preparing a dataset for submission, including identification of the chemical of interest. (Note that an option is outlined in the INSTRUCTIONS for chemicals that are not available in the drop-down menu.)
19. The second step is to review the food/feed/product names in the dataset and map the national food classification with the WHO and FoodEx2 classification. Tab "2. Food Mapping" contains the mapping tool: the Local Food Identifier (column A, free text) and two levels of classification in drop-down menus, i.e., Level 1: WHO Food Group (Column B) and Level 2: WHO Food Identifier (Column C). After the Local Food Identifier, WHO Food Group, and WHO Food Identifier fields are filled in, the WHO Food Code, FoodEx2 code, and the FoodEx2 name are generated automatically in columns E, F and G of Tab 2.
20. One source of confusion in data submissions is how often each food needs to be mapped on the food mapping template. For example, if the submitter is uploading three foods with the following "Local Food Identifiers" -- Ginger, crystallized; Ginger powder, dried; and Ginger slices, dried -- all three would be entered separately on the food mapping template, Tab 2, and mapped to WHO "Herbs, spices, and condiments" (Column B) and "Ginger, root" (Column C). However, if the submitter is uploading 100 additional data points for "Ginger, crystallized," the mapping only needs to be done once for all the "Ginger, crystallized" samples.
21. The INSTRUCTIONS also state that mapping should be done only once if the national classification is stable. While some countries or regions may have centralized data submission, in other countries, different institutions or parts of institutions may have accounts and submit data separately. If this is the case, institutions should attempt to coordinate how they are mapping food in order to have consistency across submissions.

¹ See INSTRUCTIONS FOR ELECTRONIC SUBMISSION OF DATA ON CHEMICALS IN FOOD AND THE DIET for discussion of bulk template submissions.

22. The third step in filling out the GEMS/Food template is to enter Individual Analysis results in Tab “3. Individual Analysis Results.” Fields include the Local Food Identifier (previously mapped to codes in Tab 2), chemical concentration, units of measurement, LOD, LOQ, etc. Because the Local Food Identifiers have been mapped in Tab 2, columns B, C and D on Tab 3 will be filled automatically with the information from the mapping exercise. Column A will automatically indicate an error if any of the fields on this Tab are incorrectly filled out. The remaining columns should be filled following the detailed instructions in INSTRUCTIONS.
23. Note that columns with blue headings in the GEMS/Food template are mandatory. Columns with white headings are optional (can be left blank) if the information is not available.
24. The current fields for Individual Analysis Results in the GEMS/Food database are listed in Guidance Table 1. Bracketed text (whether single or double brackets) indicates changes that have been agreed on, but not yet implemented. Blue text in single brackets indicates edits based on changes that could be implemented in a first phase (timing to be made available at CCCF18). Purple text in double brackets indicates edits that could potentially be made at a later stage.
25. Paragraphs 26 to 52, below Guidance Table 1, provide additional guidance on fields to data submitters.

Guidance Table 1: Current fields in GEMS/Food template

Column	Field	Field type /Drop-down menu	Mandatory or Optional	Flag language (proposed new or revised)
E	Local Food Identifier	Free text	Mandatory	[Provide a <i>brief but descriptive</i> name of the food, such as “Orange roughy” (versus “Fish”) or “Polished/white rice” (versus “rice.”)]
F	Serial no of the Record	Free text	Mandatory	[One serial number is used for each sample. Data on different contaminants in the same sample should have the same serial number. National institutions should coordinate serial number selection to ensure numbers are informative and non-duplicative.]
G	[Submitting] Country/Region/Observer	Drop-down menu (List of countries, regions, observers) - Unspecified	[Mandatory]	[Identifies country, region, or observer (region unspecified) submitting the data; this is not the country of production. If observer is not listed in dropdown menu, choose Unspecified and note name of Observer in Remarks.]
H	Contaminant	Drop-down menu (List of contaminants)	Optional	[Please select a contaminant from the list. A contaminant is required, but manual entry in “Column H: Contaminant” is optional if a contaminant has been added on Worksheet 1: Start. If “multiple” is selected in Worksheet 1: Start, manual entry of contaminants in Field H is required.]
I	Food Origin	Drop-down menu - Domestic - Imported - Mixed origin - Unknown	Optional	
J	Sampling Date	Free text (YYYY)	Mandatory	
K	Sample representative-ness [reliability]	Drop-down menu - Random [(routine)] sampling - Targeted sampling - Unknown	Mandatory	[Targeted sampling refers to targeted follow-up of specific findings of contamination. Random (routine) sampling refers to sampling that is not targeted and can include routine surveillance or sampling specific food types or importing countries.]
L	Laboratory Identification	Free text	Optional	[Laboratory that completed the analysis.]

Column	Field	Field type /Drop-down menu	Mandatory or Optional	Flag language (proposed new or revised)
M	Analytical Quality Assurance	Drop-down menu - Internal quality assurance and reference standards only. - Successful participation in relevant proficiency tests/[interlaboratory comparisons] during the sampling and analysis period. - Official accreditation for the relevant methods during the sampling and analysis period. - Unknown quality assurance of the lab.	Optional	
N	Measurement units for Contaminant Levels	Drop-down menu [• mg/kg • µg/kg • ng/kg • pg/kg • Bq/kg]	Mandatory	[Check units carefully. Make sure units chosen from dropdown menu align with sample results.]
O	LOD	Free text	Mandatory for results not quantified [(i.e., non-detect)] if LOQ is not provided. [[Optional]]	Enter a numeric value greater than 0 and less than LOQ. This field contains the limit of detection reported by the laboratory. [LOD and or LOQ are mandatory if Result is not provided-non-detect is entered in Results (T).]
P	LOQ	Free text	Mandatory for results not quantified if LOD is not provided. [[Mandatory]]	Enter a numeric value greater than 0 and LOD. This field contains the limit of quantification reported by [the] laboratory. [LOD and or LOQ are mandatory if Result is not provided-non-detect is entered in Results (T).]

Column	Field	Field type /Drop-down menu	Mandatory or Optional	Flag language (proposed new or revised)
Q	Results based on	Drop-down menu - Fat content - Dry weight - As is (raw, fresh [, as sold]) - As consumed	Mandatory	
R	Portion analyzed	Drop-down menu - Edible only [Whole food (edible + inedible)]	Mandatory	[Example: shelled nut (edible) versus unshelled nut (whole food)]
S	State of food analyzed [Cooked/Raw]	Drop-down menu - Cooked - Raw - Unknown	Optional	[Example: raw fish versus cooked fish]
T	Results	Free text	Mandatory	[Entering a result is mandatory: either a zero, non-detect, or a numeric result. Zero or non-detect can be entered only if LOQ or LOD are provided.]
U	Individual vs Aggregated data	Drop-down menu - Individual - Aggregated	[[Mandatory]]	

Column	Field	Field type /Drop-down menu	Mandatory or Optional	Flag language (proposed new or revised)
V	Confidentiality of Data	Drop-down menu • Yes • Blank	Optional	[All data for which “blank” is chosen or no option is selected will be considered as non-confidential in data handling and analysis.]
W	Remarks/References	Free text	Optional	
X	[[Year of production/harvest]]	[[Free text (YYYY)]]	[[Optional]]	
Y	[[Compositional information]]	[[Free text]]	[[Optional]]	[[Information from labels or as determined analytically such as major ingredients, fat content, water content, or percent total cocoa solids.]]
Z-1	[[Country/Region of Production of Finished Product]]	[[Menu • Unknown • Countries/Regions (A-Z)]]	[[Optional]]	

Column	Field	Field type /Drop-down menu	Mandatory or Optional	Flag language (proposed new or revised)
Z-2	Country/Region of Origin of Raw Materials	[[Menu • Unknown Countries/Regions (A-Z)]]	[[Optional]]	
AA	[[Product type]]	[[Menu: • Destined for further processing • Ready to eat • Not applicable • Unknown]]	[[Optional]]	[[“Destined for further processing” and “ready to eat” are defined for certain contaminants and commodities in the Codex General Standard for Contaminants and Toxins in Food and Feed, CXS-193. See paragraph XX below.]]
BB	[[Sampling location in production chain]]	[[Menu: • Unknown • Production site • Bulk lot transport • Border (import/export) • Market/Retail • Other]]	[[Mandatory]]	[[Provides information on where the sample was obtained in the production chain.]]
CC	[[Principle of method of analysis]]	[[Menu • Method A • Method B • Method Z • Other • Unknown]]	[[Optional]]	

26. **E. Local food identifier.** Mandatory field. When possible, the data submitter should provide names in English. Adding details to the name can help the data analyst with sample classification (e.g., “pineapple-orange juice” versus “juice.”) On the other hand, an overly long sample name (e.g., listing all ingredients in a multi- ingredient food) can complicate the work of analysts. Supplemental name information can also be added to the Remarks column.
27. **F. Serial No.** Mandatory field. One serial number (sample ID) should be used for each sample. If information on multiple contaminants is submitted for one sample, the same serial number should be used. National institutions are responsible for coordinating use of the same serial number for all submissions of analyte data for the same food sample. [National institutions should coordinate serial number selection to ensure numbers are informative and non-duplicative.](#) The numbering conventions should provide information that is traceable to the submitter/time of submission/monitoring program. Avoid numbering with uninformative serial numbers like 1, 2, 3, etc. If submissions are made in response to two different data calls, the submitter of the second data set should ensure that the new data has not replaced the previous data, checking with the GEMS/Food administrator as needed.
28. **G. [Submitting] Country/Region/Observer.** [\[Mandatory field. Identifies country, region, or observer \(region unspecified\) submitting the data; this is not the country of production. If observer is not listed in dropdown menu, choose Unspecified and note name of Observer in Remarks.\]](#)
29. **H. Contaminant.** [\[Optional field. A contaminant can be added on “Worksheet 1: Start” or by manual entry in “Column H: Contaminant. A contaminant must be identified, but entering the contaminant name in Column H is optional if a contaminant has been added on Worksheet 1. However, if “multiple” is selected in Worksheet 1: Start, manual entry of contaminants in Field H is required.\]](#)
30. **I. Food origin.** [\[Optional field. Identifies samples as domestic, imported, mixed origin, or unknown.\]](#)
31. **J. Sampling date.** [\[Mandatory field. Identifies the sampling date.\]](#)
32. **K. Sample representativeness.** [\[Mandatory field. The term “random \(routine\) sampling” refers to sampling that is not targeted and can include routine surveillance or sampling specific food types or specific importing countries. For example, testing a wide range of imported samples of a certain food category for the presence of a certain contaminant is “random.” The term “targeted sampling” should be chosen for follow-up sampling after specific findings of contamination. For example, if a country identifies a sample from a particular manufacturer as having high levels of a contaminant, additional sampling of the same lot or lots produced at the same time by the same manufacturer would be “targeted.”\]](#)
33. **L. [Laboratory Identification.** [Optional field. Identifies laboratory that completed the analysis.\]](#)
34. **M. [Analytical Quality Assurance.** [Optional field. Provides analytical quality assurance information on analyzing laboratory.\]](#)
35. **N. Measurement Units for Contaminant Levels.** [\[Mandatory field. Provides units for contaminant results. Check units carefully. Make sure units chosen from dropdown menu align with sample results. Ensure that the reporting unit is the same for results, LOD, and LOQ. Ideally, the data submitter should provide both the LOQ and LOD, even though these fields are currently only mandatory for non- quantified results.\]](#)
36. **O. LOD.** [\[LOD field contains the limit of detection reported by the laboratory. Enter a numeric value greater than 0 and less than LOQ. Note that LOD or LOQ are mandatory if non-detect is entered in Results \(T\).\] \[\\[\\[Reporting LOD is optional but encouraged.\\]\\]\]\(#\) \(Comment: Fields O. LOD and P. LOQ to be switched so that the mandatory filed P. LOQ comes first\).](#)
37. **P. LOQ.** [\[LOQ field contains the limit of quantification reported by the laboratory. Enter a numeric value greater than 0 and the LOD. Note that LOD or LOQ are mandatory if non-detect is entered in Results \(T\).\] \[\\[\\[LOQ is mandatory for samples submitted after XX/XX/XXXX. Prior to XX/XX/XXXX, LOQ was only mandatory for results not quantified if LOD was not provided.\\]\\]\]\(#\) \(Comment: Fields O. LOD and P. LOQ to be switched so that the mandatory filed P. LOQ comes first\).](#)
38. **Q. Results based on.** Mandatory field. Provides information on whether results are based on analysis “as is (raw, fresh, as sold)”; “as consumed”; or based on fat content or dry weight. Examples include (1) “fat content”: results for lipid (fat)-soluble contaminants in animal meat tissue, based on fat content, (2) “dry weight”: results for contaminants in animal tissue, when sample is dried in laboratory until moisture is removed before analysis, (3) “as is (raw, fresh, as sold)”: dried vegetables and fruits sold at retail; also, results for powdered infant formula, unreconstituted with water and (4) “as consumed”: results for diluted powdered infant formula.

39. **R. Portion analysed.** [Mandatory field.](#) Provides information on whether the whole food is analysed or only the edible portion (e.g., shelled nut (edible) versus unshelled nut (whole food)).
40. **S. State of food analysed.** [\[Optional field. Provides information on whether the food sample is cooked or raw \(e.g., raw fish versus cooked fish\).\]](#)
41. **T. Results.** [\[Entering a result is mandatory: either a zero, non-detect, or a numeric result. Zero or non-detect can be entered only if LOQ or LOD are provided.\]](#)
42. **U. Aggregated sample.** [Optional \[\\[\\[Mandatory.\\]\\]\]\(#\)](#) Aggregated data refers to results based on pooled samples, such as samples from Total Diet Studies. Aggregated data are often excluded from violation rate analyses conducted to determine appropriate maximum levels (MLs), which are based on observing the distribution of the data and upper percentiles exceeding proposed MLs. However, aggregated data can be included in the GEMS/Food database and limited data have been included in CCCF analyses in the past. The GEMS/Food database administrator or a statistician should be consulted before including aggregated data. If aggregated data are included in an ML analysis, this fact should be noted in the EWG paper.
43. **V. Confidential data.** [\[Optional field. Note that all data for which “blank” is chosen or no option is selected will be considered as non-confidential in data handling and analysis.\]](#) Countries can submit data as “Confidential” if they wish to limit access to use by FAO, WHO and related technical bodies, such as Codex. The GEMS/Food Administrator can provide records marked “Confidential” to EWG Chairs; therefore, EWG Chairs should always consult with the GEMS/Food Administrator on data extraction before downloading data. If a country submitted data as “Confidential” in response to a Call for Data, the submitting country also should make the EWG Chair aware of this fact during the data extraction/analysis phase.
44. **W. Remarks/References.** [\[Optional field that can be used to add relevant information for which there is not a defined field.\]](#) Examples of information that has been included in this column are information on product labels (such as ingredients or detailed product names) or detailed information on method of analysis. Other information that may be entered in this column is a reference to a specific Call for Data.]
45. [\[\[X. Year of production/harvest. Optional field. Provides information on year of production, e.g., for food of animal origin, or year of harvest, e.g., for food of plant origin, if information is available.\]\]](#)
46. [\[\[Y. Compositional information. Optional. Information from labels such as major ingredients, fat content, water content, or percent total cocoa solids for chocolate.\]\]](#)
47. [\[\[Z. Country/Region of Origin/production. Optional. Provides information on the Country or Region of Origin/Production of the \[food or raw material or sampled lot or consignment\]. If this information is not known, Unknown can be selected.\]\]](#)
48. [\[\[AA. Product type. Optional. Provides information on whether the sample is “Destined for Further Processing” or “Ready to Eat,” or whether this information is “Not applicable” or “Unknown.” “Destined for Further Processing \(FFP\)” and “Ready to Eat” are defined in several places in the Codex \[General Standard for Contaminants and Toxins in Food and Feed\]\(#\) \(CXS-193-1995\). For aflatoxin in maize, sorghum, tree nuts, peanuts, and dried figs or deoxynivalenol in cereal grains, “destined for further processing” means intended to undergo an additional processing/treatment that has proven to reduce levels of aflatoxin or deoxynivalenol before being used as an ingredient in foodstuffs, otherwise processed or offered for human consumption. For aflatoxin in tree nuts and dried figs, “Ready to Eat” means not intended to undergo an additional processing/treatment that has proven to reduce levels of aflatoxins before being used as ingredient in foodstuffs, otherwise processed or offered for human consumption. For the commodities mentioned above, “For Further Processing” and “Ready to Eat” or “Unknown” can be selected; for other commodities, “Not Applicable” should be selected.](#)
49. CXS-193-1995 also refers in several places to “ready to eat” infant meals and liquid milk “for further processing.” Field AA does not apply to these uses of “ready to eat” and “for further processing.”]]
50. [\[\[BB. Sampling location in production chain. Mandatory. Provides information on where the sample was obtained in the production chain, i.e., Production site \(e.g., firm, food plant\), Bulk lot transport, Border \(e.g., testing at import or export\), Market/Retail, or Other. If this information is not known, Unknown can be selected. If Other is selected, information can be added to Remarks.\]\]](#)
51. [\[CC. Principle of method of analysis. Optional. The method of analysis, such as ICPMS or GC-MS, can be selected from a pull-down menu.\]\]](#)

52. **Errors.** Prior to upload, the data submitter should review the file carefully for errors. During upload, the data file is scanned to identify problems before writing data into the database. The data submitter is responsible for correcting errors and re-submitting the template. Datasets can be rejected for a variety of reasons, some of which are listed below. The GEMS/Food database administrator can be contacted for assistance.
- Reported result < LOD, -missing LOQ or LOD when result is non-detect **[-missing LOQ-]**, reported LOD > LOQ
 - Dates entered in the wrong format
 - Mandatory fields incomplete
 - Duplicate entries in the current worksheet or in the database

DATA EXTRACTION

53. The data extraction process begins at the database website: [GEMS/Food Contaminants Database](#). As noted above, for full functionality, analysts must register and log in to their accounts. After logging in, analysts will see a Welcome page with two tabs, a Home Page tab and a Search tab. The Home Page tab contains a limited number of prepared extracted datasets by region and contaminant. For specific searches, the analyst selects the Search tab. The Search function allows the analyst to filter data by WHO Region, Contaminant, Food Category, and Food Name, and Sampling Period. These filters will allow the analyst to identify data responsive to a particular Call for Data or TOR.
54. To identify the most accurate dataset for extraction for development of ML proposals, it is best to consult with the GEMS/Food database administrator. Data submitters may make choices when submitting data that could result in data being missed during extraction. For example, data uploaded as “food for infants and children” may be missed in a search limited to “fruit and vegetable juices.” Another example is that juice data may be mistakenly mapped as “fruit and fruit products” although the Local Food Identifier or Remarks field clearly identifies the samples as juice. Consultation with the GEMS/Food database administrator before extraction may help the EWG ensure they have extracted all the relevant data for the ML analysis from GEMS/Food.
55. Confidential data is another reason EWG Chairs should always consult with the GEMS/Food Administrator on data extraction before downloading data. The GEMS/Food Administrator can provide records marked “Confidential” to EWG Chairs. These records will not show up in a routine search as described above. EWG members who are interested in more detailed analysis of confidential data can consult with the EWG Chair.
56. It is important to maintain a record of all filters and search terms for the EWG report.
57. **[[Changes were made in the GEMS/Food database in 2025/6 to make LOQ mandatory and to introduce new fields such as X, Y, Z1, Z2, AA, BB, CC. Legacy data submitted before these changes should be considered valid and be used in ML development.]]**

DATA SELECTION /CLEAN-UP OF DATA

General considerations

58. This section provides guidance on the selection and clean-up of data for submission and analysis in the context of CCCF work. It is intended to support consistent and transparent handling of data across different sources, while acknowledging that key decisions—e.g. if MLs need to be developed— fall within the remit of the CCCF. The procedures outlined here aim to facilitate high-quality data analysis while promoting harmonization among contributors.

Origin of data

59. For development of MLs based on occurrence data, only data in the GEMS/Food database should be used. Non-GEMS/Food data can only be used to support a complementary analysis to inform understanding of the data, e.g., when only limited data are available in the GEMS/Food database for certain time periods or regions, particularly limited data from primary producing countries.
60. Non-GEMS/Food data, such as data directly submitted to the EWG by Codex Member Country(ies)/Organization or Observer(s) or obtained through a literature search, are also subject to clean-up procedures, as necessary.

Clean-up of data

61. The purpose of data clean-up is to remove non-relevant or inappropriate samples from the dataset before analysing the dataset to recommend MLs.

62. Examples of samples that should be removed from a dataset include:
- Samples that are clearly outside the TOR of the work e.g., ketchup samples for work on tomato sauce.
 - Samples that are outside the date range of the TOR of the work, e.g., samples that are 20 years old and the TOR refers to data from 10 years previously. This is particularly important if a Code of Practice (COP) for prevention and reduction of contamination of relevant contaminants in foods has been adopted since older data were generated.
 - Samples missing crucial information (see paragraphs 66-73 below).
 - Samples with unacceptably high LOQs (see paragraphs 93-96 below).
63. Clean-up of data refers only to the extracted dataset. The original data in the GEMS/Food database will not be modified and will remain unaffected by the steps indicated below, unless the data submitter requests corrections or other changes from the GEMS/Food administrator.
64. For the clean-up of data, it is recommended to involve an expert on the specific contaminant who may have insight in which patterns in data are irregular or not.
65. All steps taken in the clean-up of data should be recorded and described in the final Codex working document presented by the EWG to the Plenary of CCCF, e.g., detailed information on the reason for data exclusions (e.g. LOD/LOQ specified is higher than hypothetical MLs being considered, questionable outliers, etc.), number of data points excluded during the clean-up process, if possible also including a breakdown of how many were excluded at each step, etc.

Handling of data

Missing information

66. Once all non-relevant or inappropriate data are removed from the data set, then data should not be excluded if all mandatory fields are completed (see section data collection and submission) and the data meet the criteria for uploading in the GEMS/Food database. It should be noted that even if all mandatory fields are filled in and the data meet the criteria, the data may still be incomplete for deriving MLs.
67. In cases where sample information in the GEMS/Food database is needed but incomplete, (e.g., missing or unclear), the first step should be to reach out to the contact point for the data-submitting country/organization or observer to allow for missing information to be obtained. This is particularly important if the missing information was requested in a Call for Data. The GEMS/Food database administrator can also be asked to conduct this outreach.
68. If missing information is available, a corrected data file should be provided by the submitting country/organization or observer to the EWG and the GEMS/Food administrator. Analysis using the corrected samples may continue. The request and corrections should be noted in the EWG working document.
69. If missing information cannot be obtained and the EWG chair concludes that the data should be excluded in the analysis due to missing information, the EWG should note the exclusion of the data and possible impacts of exclusion in their working document.
70. Examples of missing information indicating that data should possibly be excluded from further data analysis:
- the state of the food analysed is missing (e.g., cooked versus raw, and the analysis is intended to be based on raw food only)
 - inadequate product description in the local food identifier field (e.g., the analysis is being performed on "mackerel", but the product is described as "fish," and the analysis depends on identifying fish species)
 - and others
71. Examples of missing information that would not prevent further data analysis (depending on case-by-case review):
- sampling information: type of sampling, location of sampling in production chain
 - state of the product, for example, raw or cooked is not identified in Field 'S' but information can be deduced from other information provided, e.g., the sample is described as cooked fish.
 - principle of method of analysis used
 - when ML is based on a sum-of-components and data are not reported for all the components but for those that contribute significantly to the sum or in case the occurrence is reported as sum.

72. The EWG chair should not exclude all data that is missing any optional field but should consider what details are needed for the development/elaboration of MLs. For example, some commodities may require additional information to develop MLs, e.g., for raw grains, information on processing stage may be important to propose MLs, but this information would not be needed for finished product beverages.
73. As noted above, non-GEMS data considered as part of a complementary analysis may be missing crucial information. The EWG Chair should apply the same criteria and questions (e.g., is the missing information critical to inclusion) to the non-GEMS data.

Incorrect information

74. In cases where there are clear indications that the unit of the data or the basis on which the data are reported are incorrect, the first step should be to reach out to the contact point for the country/organization or observer that submitted the data and request that they review the entries for possible corrections. The GEMS/Food database administrator can also be asked to conduct this outreach.
75. If the submitting country/organization or observer agree that a correction is needed, a corrected datafile should be provided to the EWG and the GEMS/Food administrator by the submitting country or observer. Analysis using these samples may continue. The request and corrections should be noted in the EWG working document.
76. If the accuracy of the data cannot be confirmed and corrections cannot be made, these data should be excluded from further data analysis.
77. Examples of apparent errors that should lead to contacting submitters for possible correction and resubmission:
 - All data in a 200-sample dataset are expressed as µg/kg, except 5 quantified data points expressed as mg/kg. When plotting these data in a frequency distribution curve, after having converted them to the same unit, the five data points in mg/kg would be identified as possible outliers (see paragraphs 82-92 and Annex I).
 - 195 results in a 200-sample dataset of samples of food with a typical fat content of 5 % fall in the range of 0-20 mg/kg; however, 5 data points fall within in the range of 100 – 400 mg/kg, suggesting they were reported on a fat basis rather than the designated whole weight basis. When plotting these data in a frequency distribution curve they would be identified as possible outliers (see paragraphs 82-92 and Annex I).

Data for which the information on the portion analysed is not clear

78. For some foods (e.g., fruits, rice), if the portion analysed is not clear (e.g., peeled vs whole fruit, or husked rice vs polished rice), the point of contact for the country/organization or observer that submitted the data can be contacted for clarification. In case no clarification is provided, it should be reflected whether the unclear information is important for the contaminant in question and the final concentration found in the product. In addition, for some foods it may be assumed that the portion was analysed in the state that it is usually sold/consumed, e.g., citrus fruit is usually fresh unless it is clearly identified as canned. Any such assumptions should be recorded and presented in the final document presented by the EWG to the plenary of CCCF. If no reasonable assumption can be made, these data should be excluded from further data analysis unless the necessary information was obtained.

Data originating from suspected fraudulent/economically adulterated samples

79. In assessing whether contaminated samples are the result of fraudulent/economic adulteration, the nature of the contaminant must be taken into account first (e.g., lead versus mycotoxin). Wide differences from year-to-year may be the result of natural variability (e.g., high level of mycotoxins due to specific climate conditions in a certain region/production year). In other instances, wide differences may be the result of poor practices (e.g., lead). Possible signs of fraudulent/economically adulterated samples are:
 - certain samples are orders of magnitude higher than others, e.g., 0.1 mg/kg versus 100 mg/kg, or
 - temporal variability in data (depending on contaminant), e.g., data are much higher in one year of the dataset.

Data that are clearly related to fraudulent/economically adulterated samples should be excluded from the analysis and the exclusion must be documented.

Data from targeted sampling and bias

80. Targeted sampling differs from random sampling in that targeted sampling refers to targeted follow-up of specific findings of contamination. In principle, these data should not be used in the derivation of MLs as they are not representative of the general population and may not reflect achievable levels in regular situations.
81. It should also be noted that some bias could be introduced in random sampling as there might be reasons for more extensive sampling in specific regions or types of products. Such data could include higher or lower levels than the normal range and should not be excluded without further consideration as these reflect natural variation in the occurrence data.

Determination and handling of outliers/extreme values

82. The EWG Chair should review the dataset to determine if there are outliers/extreme values that should be removed. This section provides guidance on different approaches to identifying outliers/extreme values and protocols for removing them, if appropriate.
83. As a general rule, it is important that outliers/extreme values not be discarded unless there is a valid reason to do so.
84. Extreme values can be valid values, e.g., due to the heterogeneity of contaminant distribution (such as hotspots for mycotoxins) or due to natural variation of measured contaminants (e.g., resulting from weather conditions, soil condition, etc.). Other extreme values can be erroneous, e.g., errors in measuring and processing data, including incorrect calculation or using the incorrect unit of measurement, or result from fraudulent behavior (economic adulteration).
85. Erroneous values or values resulting from economic adulteration should always be removed from the final dataset before determining MLs. Valid extreme values will have to be reviewed on a case-by-case basis, using visual inspection of the data (see Annex III for examples) first, followed by statistical methods (Annex I).
86. The presence of outliers in datasets has a significant impact on the arithmetic mean and extreme values, but not on the median. However, consideration should be given to the percentage of any potential outliers present in the whole dataset. Since the high percentile values, rather than maximum values, are used as a basis for the development of MLs based on rejection rates, the impact of outliers on derived MLs will usually be small. However, in cases where a notable percentage of data points (e.g. 2–5%) are excluded from the dataset as outliers, this could affect the high percentile values (see also paragraph 92). In these cases, it is appropriate to provide a comment regarding the effect of the exclusion of the outliers on the achievability of MLs (i.e. rejection rate) under consideration.
87. Annex I to this guidance document provides more details on statistical approaches to identify outliers.
88. “Outlier” defined in the Guideline on Analytical Terminology (CXG 72-2009) assumes a normal distribution in a dataset comprised of results from repeated analysis of the same sample. It states, “*the statistical outliers are discarded unless the statistician for good reason decides to retain them*”. In contrast, the datasets addressed in this guidance document are analytical results from a variety of samples and from different analytical methods. Because the distribution of these data is unknown (in many cases, it will not follow a normal distribution), and the data may be combined from multiple sources, it is difficult to predict the range of variation within a dataset. Therefore, this Guidance recommends that the statistical outliers not be discarded unless a good reason to exclude them is identified and scientifically explained.
89. Extreme values that are identified as errors should be addressed as described in paragraphs 74-77. However, extreme values that lack a clear explanation should be retained and evaluated to determine if they should be handled as outliers in the final dataset. A sensitivity analysis can be performed to understand the impact that inclusion of outliers has on the overall assessment.
90. As there can be many causes for extreme values and some of these values may not be regarded as extreme if combined with data from other sources (countries/regions, different years, etc.), the possible exclusion of an extreme value as an outlier should be determined based on the combined, cleaned-up dataset. If individual datasets are analysed separately, more careful consideration should be given to the exclusion of extreme values as outliers.
91. There may be cases where extreme values are scientifically valid depending on production and weather conditions and/or other potential factors such as volcanic eruptions, etc. Considering the characteristics of the contaminant distribution of occurrence data in food, it is not recommended to simply exclude extreme values based on the results of statistical outlier tests or other methods such as visual inspection. Since the range of the concentration distribution that can be empirically or theoretically assumed varies significantly depending on the type of contaminant (heavy metals, mycotoxins, etc.), the handling of extreme values must be determined on a case-by-case basis. For example, special consideration should be given to mycotoxins whose concentrations can vary significantly depending on the sampling methods utilized due to the heterogeneous distribution in a lot, as well as very large annual variations.

92. If outliers are excluded, it is recommended that the reason for exclusion be clearly reported in the final document presented by the EWG to the plenary of CCCF. Sensitivity analyses can be performed to show how the exclusion or non-exclusion of outliers may or may not affect the calculation of high percentile values. It should be reiterated that a few extreme values remaining in the dataset will have little effect on the calculation of the high percentile values, provided that the total number of data points in the dataset is sufficiently larger than the minimum number of data points required to calculate high percentile values.

Limit of Quantification (LOQ) and Limit of Detection (LOD)

Exclusion based on LOQ

93. Different methods of analysis provide different LODs and LOQs. A high LOQ does not automatically mean that the data should be excluded.
94. Guidance for data inclusion/exclusion in different LOQ/LOD scenarios
- In the case where no LOQ/LOD is provided for a specific dataset
 - The submitting country/organization or observer should be contacted as a first step to obtain such information (i.e. LOD and/or LOQ).
 - In the case where the dataset contains (nearly) only quantified results: the data set could be used.
 - In the case where the dataset contains a significant amount of left-censored data (individual data without quantified (finite) values, generally referred to as data below the reported LOQs/LODs): data set should not be used.
 - In the case where an LOQ is provided:
 - Identify a cut-off level for the LOQ in the analysis depending on the MLs being considered (examples: $LOQ < ML$ under discussion, $LOQ < 1/2 ML$ under discussion).
 - In the case where the dataset contains a significant amount of left-censored data: further contextual information should be considered (e.g., are data from an importing country or producing country).
95. If almost all data in the dataset are below the LOQ or reported as non-detects (NDs, <LOD), it is not possible to estimate high percentile values to establish proposed MLs. When there are only a small number of quantitative values, the dataset should be handled on a case-by-case basis following the guidance provided in the section on "Handling of datasets with a large proportion of left-censored data". In this case, when proposing MLs, it is not appropriate to calculate high percentile values using only quantitative values as this may result in unnecessarily high MLs.
96. The EWG working document should clearly outline the criteria by which certain data were excluded from the dataset due to high LOD/LOQs (e.g., the reported LOD/LOQ for some samples is higher than the proposed ML, or the reported LOD/LOQ for some samples is 'x' orders of magnitude greater than the lowest LOD/LOQ of the vast majority of other samples in the dataset suggesting an error) or if the whole dataset should be excluded from the analysis, as removing individual data can introduce bias.

Sum of components and LOQ

97. In the case of levels of contaminants which are a sum of components (e.g., total aflatoxins), the following should be considered.
- The general rule is that levels of contaminants that are a sum of components are reported as the lower bound, i.e., for non-quantified levels of components, the values are set to 0.
 - The LODs or LOQs for the individual components are required to be provided for the non-quantified components below the LOD or LOQ.
 - When only data on individual components are reported, the individual data can be summed into a total result.
 - In specific cases, it may be appropriate to report levels of contaminants that are a sum of components using a middle bound approach (i.e. non-quantified levels of components are set equal to $\frac{1}{2}$ of LOQ (or the LOD in case the LOQ is not provided) or upper bound approach (i.e. non-quantified results are set equal to LOQ (or the LOD in case the LOQ is not provided); however, these cases should be clearly identified in advance before data submission in the instructions for the Call for data.

DATA ANALYSIS: GENERATING OVERVIEW OF DATA

Overview of countries, number of data points, period of data coverage

98. After clean-up of the dataset, the remaining data are considered to be of sufficient quality for the analysis. An overview of these remaining data with details (e.g., country of production of finished product, production/harvest year, amount of data included and excluded) should be provided in a table and described in the working document. All steps taken in the data clean-up and the rationale and assumptions made should be provided with the overview. In addition, it could be useful to provide information (e.g., from FAO) on the major production regions for the commodity under discussion. Based on this overview, the EWG can also present more focused analysis of specific geographical areas and time periods.

Geographical coverage of the provided occurrence data

99. When evaluating an ML for a particular commodity, the dataset should include representation of production regions that are important to international trade. Therefore, it is helpful (and may be required in the future) that the county/region that produces the finished products is reported in the GEMS/Food Database (see Section Data Collection and Submission). Also, major production regions for certain commodities may be geographically limited; if this is the case, the EWG Chair can report this information in the working document. In that context, data from producing regions should be considered in relation to data from countries importing the food, as the data might be biased if the food has to comply with the requirements of the importing country such as an ML already established in that country. If a region has very restrictive MLs such that contamination levels of imported foods are skewed to the left, this may not be an accurate representation of the variability in contamination from producing regions. However, data from importing countries also reflect foods (ingredients) traded internationally and as consumed and should be considered to some degree. Indeed, additional contamination could take place during transport from the producing country (e.g., mycotoxin production).
100. In some cases, it could be appropriate to give priority to datasets from producing countries over datasets from importing countries. However, in that case, guarantees should be provided that the datasets from producing countries do reflect the implementation of good agricultural and manufacturing practices as provided in Codex CoPs and are representative of products that would be traded internationally.
101. Another option is to provide separate analyses for producing and importing country datasets. If possible, a sensitivity analysis on using data from producing versus importing countries should be performed to guide the selection of data used to set MLs.
102. Only if there are enough data that show an indication of large differences in reported levels between regions or between countries in a region, analysis could be performed by region or country. It should be noted that for a country approach, this should be done for major producing countries in the region and sufficient data should be available. (See Section on Statistical analysis of occurrence data/Handling of datasets for ML development).
103. In summary, although there are different approaches possible to look at data from importing countries versus producing countries or regional data, it is important to keep in mind that Codex MLs are global standards, so the default approach for data processing should be to analyse data globally.
104. Guidance for datasets that lack geographic coverage:
 - If the region(s) for which data are lacking is/are important production region(s) and on the condition of a clear commitment from the region(s) to provide additional data, some additional years (e.g. 2-3 years) typically are allowed for data collection before continuing the discussion on ML proposals, provided that there is no urgency and CCCF agrees. After expiry of the granted additional years, the discussion on MLs would be continued based on available data, regardless of whether geographic coverage has been reached or not.
 - If there is no commitment from the important producing region(s) to provide the additional data or if collecting additional data within agreed timeframe (e.g. 2-3 years) is not feasible, the consideration on MLs will be continued based on available data or be discontinued.
 - If the region(s) for which data are lacking is/are not important production region(s): the consideration on MLs will be continued based on available data.
 - If there is an urgent need to establish an ML for consumer health protection, a compromise needs to be reached in CCCF to set an ML based on available data. In these cases, the ML can be reviewed within 3–5 years to assess whether adjustments are necessary when more data are available.

Time period coverage of the provided occurrence data

105. It is appropriate that the provided occurrence data relate to multiple production years for ML development. Requirements may be different for different types of contaminants (e.g., mycotoxins, plant toxins, marine biotoxins, processing and environmental contaminants) and is a function of the assumed year-to-year variation or evolution of contamination in time.
106. For contaminants such as mycotoxins which are known to have year-to-year variation, data from the last 10 years may provide a very good representation of the year-to-year variation; however, there may be cases where more than 10 years of data should be considered (e.g., sampling effort reduced in recent years or fewer higher quality datasets are available). For other contaminants, year-to-year variation is less relevant and possibly more recent data (or a smaller time range) can be selected. In any case it should be discussed whether data older than 10 years are relevant for the analysis.
107. Further, it could be relevant to investigate/include older data to learn whether certain species/subgroups from a food group/category tend to have higher levels.
108. It may be appropriate in certain cases to perform a time trend analysis. In these cases, data from more than 10 years are to be considered to determine if concentrations have changed/are changing with time and this could be used to determine whether a certain number of years of data should be used for ML elaboration to represent current concentrations.
109. If a Codex CoP has been established and implemented, the data under consideration should be from the years after the implementation of that CoP to reflect good agricultural and manufacturing practices, unless indicated by a country that equivalent or stricter (national or regional) practices had already been implemented before the establishment of the CoP, in which case such data can be included.

Elements that could indicate whether a CoP was implemented could be:

- A consistent drop in levels after a certain year, and
 - Differences in levels from neighbouring countries within one region which cannot be explained by geographical factors.
 - Information from submitting country about implementation practices
110. It is often difficult to judge from datasets whether or not a CoP has been implemented. Preferably, information on implementation of a CoP and whether the country submitting that data has any of their own already established MLs in place is requested in the call for data. If the EWG excludes data on the basis of failure to implement a CoP, the exclusions and rationale should be clearly documented in the final document presented by the EWG to the Plenary of CCCF.

STATISTICAL ANALYSIS OF OCCURRENCE DATA /HANDLING OF DATASETS FOR ML DEVELOPMENT

General considerations

111. The following sections explain considerations before conducting a statistical analysis of the extracted data/cleaned-up data and how the results of statistical analysis should be presented in the EWGs for developing globally applicable MLs.
112. Statistical analysis should be conducted on the extracted/cleaned-up data. When assessing the data, the first step should be analysing the distribution of the dataset. In general, the distribution of contaminant data in food tends to be right skewed, e.g., a log-normal distribution (See Figure 1). For such non-normal distributions, use of parametric statistical methods, which are based on the normal distribution are not appropriate.
113. The GSCTFF states in Annex I, *"MLs should be set at a level which is (slightly) higher than the normal range of variation in levels in food and feed."* This means that to develop an ML, there is a need to estimate high percentile values (generally 95th percentile values) with a significantly high confidence level. In food safety, a confidence level of 95% is usually used. The figure below (Figure 1) explains, using a modelled distribution, the relationship among a high percentile value, hypothetical ML (usually a rounded value of the percentile value) and percentage of data points that exceed the proposed ML when the ML is the same as the 98th percentile value.

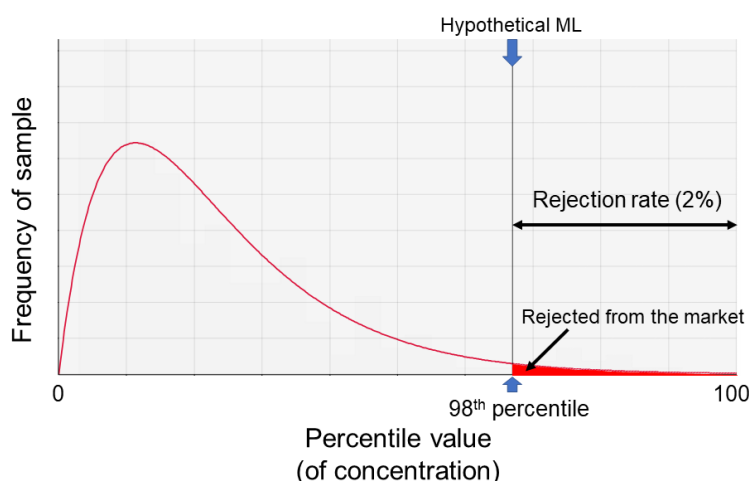


Figure 1. Simplified depiction of the relationship among a high percentile value, hypothetical ML, rejection rate and percentage of samples exceeding the proposed ML.

Note: In the above, it is assumed that the hypothetical ML is the same as 98th percentile value. The choice of the 98th percentile for hypothetical ML in this figure is for example only.

Minimum number of data points for estimating high percentile values

114. For development of an ML, it is necessary to estimate high percentile values (generally 95th percentile values) of a dataset. A minimum number of 59 data points is required for a 95th percentile estimation with 95% confidence (see Option 1, Annex II). Annex II provides additional details on alternative options for calculating the minimum number of data points required in relation to estimating high percentile values.

Handling datasets

Handling datasets with low number of data points

115. When the JECFA evaluation suggests that a health risk from exposure to a contaminant is significant, a smaller number of data points than the minimum number of data points (i.e. 59 data points, refer paragraph 114 and Annex II) would be considered adequate for developing an ML, provided the confidence level of the estimated high percentile values is only slightly lower than the expected high confidence level, such as 95%. (See Annex II for an example of how to calculate the confidence level.) For example, when an ML is urgently needed for consumer health protection, the EWG tasked with recommending MLs should consider recommending to the CCCF the development of MLs even if only a small number of data points are available. If sufficient data become available in the future, revision of the previously established ML can be considered.
116. If there is not an immediate health risk, and the number of data points is insufficient for developing an ML, additional data calls could be requested. However, if after repeated data calls, the available number of data points is still much lower than the required minimum number of data points, a decision should be recommended to the CCCF on a case-by-case basis about whether to develop an ML using the limited dataset or to discontinue the work. Another option may be to expand the ML to a larger food group, if an ML for the larger group is justified.
117. For commodities that are not routinely consumed and/or not traded internationally, availability of occurrence data may be insufficient. In such cases, the EWG tasked with recommending MLs should consider recommending to the CCCF that the ML requested for the commodity/contaminant combination may not meet the criteria described in the GSCTFF and Procedural Manual (Section 4.5 “Policy of the Codex Committee on Contaminants in Foods for Exposure Assessment of Contaminants and Toxins in Foods or Feed Groups”) for developing MLs.
118. If the number of data points is significantly less than the required minimum number of data points (i.e. 59, refer paragraph 114 and Annex II), and there is no strong reason for developing an ML immediately, there is no need to perform further statistical analyses. If additional data are needed to establish a statistically robust ML, further data calls may be necessary.
119. In reviewing existing MLs, even if only a small number of data points from limited regions are available and/or no new data will be generated, the EWG tasked with recommending MLs should not automatically recommend revoking the ML due to the small number of data points unless the ML is inconsistent with current good agricultural and manufacturing practices or current toxicological data. If a potentially significant risk exists from consuming the commodity, an option would be to recommend to CCCF to maintain the existing ML, or if there is no longer a significant health risk or there is no known trade barrier, an option would be to recommend to CCCF to

revoke the ML. In some cases, when reviewing an ML, if there are only a small number of datapoints for a particular commodity, it may be possible to consider merging the commodity under the food group from which the commodity was originally excluded (e.g., removing an exclusion for canned *Brassica* from a canned vegetables ML (ref. REP18/CF paragraph 32)).

Handling datasets where available data on individual commodity(ies)/food(s) are insufficient, but data for the food group are sufficient

120. Even when the data points are sufficient for a whole food group, if the data are separated according to individual foods in that food group, the data points may be too small for development of MLs for individual foods. In general, the analysis of currently available data should begin in development of a discussion paper. Based on available data, the EWG should recommend a preliminary approach to setting MLs for individual foods versus food groups as well as recommending language for a new data call. If after the data call and data collection, it is found that there are fewer data points available than initially expected, the food(s) that the ML should target may need to be changed to a broader range of foods, e.g., individual food(s) to food subgroups or food subgroups to food group.
121. The appropriateness of developing an ML for a food group depends on whether the distribution of contaminant concentration values in the individual foods within the food group are similar. Non-parametric statistical tests, such as Mann-Whitney U test (for 2 datasets) or Kruskal-Wallis H-test (for 2 or more datasets), can be used to determine if the distribution of those foods in the group can be considered to have a similar distribution, even when the number of data points is relatively small. However, as contamination levels among commodities within a food group may vary significantly, it is important to be flexible in the choice of methods and interpretation of statistical testing. If the number of data points is relatively small, comparison of datasets by box-and-whisker plots is also useful, provided the percentage of left-censored data is less than 25% of the respective dataset.
122. If an individual food shows a different distribution of contaminant concentrations from the other foods in the food group, two different MLs may need to be established, one for the food group excluding the individual food, and the other for the individual food. Similar approaches can be taken for subgroup(s) in the food group. If there are insufficient data for individual food(s) to meet the required minimum number of data points, additional data calls can be issued for those foods for which it is considered necessary to develop MLs. If the consumption of an individual food which shows a different distribution pattern from the food group does not contribute significantly to the total exposure to the contaminant of concern – it may be considered negligible from a consumer health protection point of view. In such cases no additional data call is required and its exclusion from ML development for the food group may be considered (e.g. ML for lead in salt, food grade excluding salt from marshes (REP18/CF paragraphs 39-41). As for food groups and their sub-groups, reference can be made to the commodity covered by relevant Codex Commodity Standards, Codex Committee on Pesticide Residues (CCPR) Classification of Foods and Animal Feeds, and other food categorization systems used by Codex Committee on Food Additives (CCFA) on processed foods.
123. When developing an ML for a broader food group due to limited data availability for individual foods or subgroups, some foods (or subgroups) might show different distribution patterns from others in the same food group. If there is not sufficient data to develop a separate ML for these foods, the EWG tasked with recommending MLs could recommend to the CCCF whether these foods should be excluded from the ML development until sufficient data become available.

Handling of datasets (including use of substitution methods) with a large proportion of left-censored data points

124. The “dataset” in this section refers to a dataset or datasets which is (are) among the dataset(s) selected to be used for ML development. This section is particularly relevant when the datasets used for ML development contain a high ratio of non-quantified data points (e.g., due to low sensitivity of available analytical methods for the concentration in the samples; extremely low frequency of occurrence; etc.) after data clean-up.
125. Though no official definition of the term “left-censored” is found in any of the Codex documents, in statistics, individual data without quantified (finite) values are called left-censored data generally referred to as data below the reported LOQs/LODs.
126. For statistical analysis of datasets containing left-censored data, conventional substitution methods should be considered, particularly when calculating statistics such as the 95th percentile, or when estimating sample rejection rates and reductions in exposure for target commodities under hypothetical MLs. If the dataset contains a high ratio of left-censored data, statistical analysis using only quantified values is not recommended because this practice introduces bias into the results of the statistical analysis.

Substitution methods for datasets with a large proportion of left censored data

127. The conventional approach to deal with left-censored data for statistical analysis is the use of one or more of the following substitution scenarios:

- Lower-bound (LB) scenario: results below the LOQ are replaced by zero, or by LOD if the LOD is known (results <LOD are replaced by zero);
- Upper-bound (UB) scenario: results below the LOQ are replaced by the reported LOQ value; and
- The middle-bound (MB) scenario: A point estimate between the two extreme scenarios (LB and UB); assigning a value of LOQ/2, square root of the LOQ, or (LOD + LOQ)/2 if the LOD is known for analytical results below the reported LOQ. In general, LOQ/2 is the most widely used.

For each of these scenarios, if the LOQ is not reported and only the LOD is reported, use the LOD as an alternative.

128. In general, depending on the distribution of data, these substitution methods may be used for calculating measures of central tendency such as the arithmetic mean when estimating dietary exposure (See EHC 240¹¹). If the EWG tasked with recommending MLs is to perform preliminary calculations for the reduction of exposure, the choice of LB, MB or UB scenarios may affect the calculated arithmetic mean and the estimated exposure from target commodities based on the arithmetic mean. However, for ML development, the effect of left-censored values on the empirical estimation of high percentile values may be negligible and there may be little impact on the ML regardless of which scenario is chosen, unless a large majority of data points are left-censored (i.e., <LOD).
129. The datasets with a large proportion of left-censored data should be handled on a case-by-case basis, depending on the result of JECFA risk assessment on the contaminant and the consumption of the food concerned. Ideally, the LB, MB and UB estimates should be calculated and presented. It is very important to know the distribution of quantified values in case of high percentage of left-censored data when estimating high percentile values using a modelled distribution for developing MLs.
130. When the dispersion of quantified values is within a narrow range (values close to each other) and close to the reported LOQ, developing an ML may be unnecessary unless the contaminant is highly toxic. The EWG can make a recommendation to the CCCF on the appropriateness of an ML in this situation.

Handling of multiple datasets

Analysis of individual and combined datasets and making decision on the necessity of comparing individual datasets before combining, especially when distribution patterns are significantly different.

131. As Codex MLs are for global application, they should be ideally based on global datasets. While a default approach for ML development is to use the combined global dataset, individual datasets per year or per region are provided for additional consideration. Whether an ML should be based on a global dataset or a dataset from a specific region/year should be decided by the CCCF on a case-by-case basis following statistical analysis as described in this section.
132. Parametric statistical methods are available for comparing distribution patterns of individual datasets per region/country or per year. The null hypothesis is that all datasets are assumed to be from the same population. Such tests include Mann-Whitney U test (for 2 datasets) or Kruskal-Wallis H-test (for 2 or more datasets).
133. Many templates for non-parametric statistical methods are available on the Internet. Among them, MS Excel templates for performing Mann-Whitney U tests and Kruskal-Wallis H-tests are available for download from the FAO's JMPR website¹².
134. In addition, it is helpful to draw box-and-whisker plots or histograms of each dataset to compare if there are visual differences in the distributions before combining the datasets. It is preferable to draw a histogram only when the dataset contains a sufficient number of data points (see paragraph 114 and Annex II). For a dataset with a smaller number of datapoints, it is difficult to know the shape of the distribution by a histogram, and a box-and-whisker plot is more helpful (See Annex III).
135. Proposing an ML(s) using combined dataset(s) (global datasets) has been done conventionally in EWGs. When there is no significant difference between the distributions of multiple datasets from different sources, it is considered of little importance to perform additional data analysis and comparison for each individual dataset (although it would be ideal to do so if resources and time permit).
136. When the number of data points is significantly different between individual datasets from different regions/countries, the resulting combined dataset reflects primarily the conditions of a country/region with the larger number of datapoints, rather than that of all countries/regions submitting the data. To address this problem, it would be theoretically feasible, although requiring a complex process, to balance the datasets by weighting them by the production or trade volume or on any other reasonable factors. However, the methodology and justification

¹¹ ENVIRONMENTAL HEALTH CRITERIA 240, Principles and methods for the risk assessment of chemicals in food (WHO, 2009)

¹² "Appendix XIV Electronic Attachments (2020_Nov)" and open "XIV 12 Spreadsheet for Kruskal_Wallis 20 group.xls" to carry out Mann-Whitney U test and Kruskal-Wallis H-test. <https://www.fao.org/agriculture/crops/thematic-sitemap/theme/pests/jmpr/jmpr-docs/en/>

for the use of data weighting has not been considered in the past in CCCF, and because of its complexity and related workload, this is not envisaged to be done until a workable guidance is elaborated in the future for this. If there are concerns about distortion in the distribution of the combined dataset due to the presence of a very large dataset, it should be considered on a case-by-case basis. For the time being, it may be recommended to also conduct analyses for each individual dataset, as presented in the following sections or to seek advice from JECFA on data analysis.

Cases where the analysis of individual datasets is recommended

137. If a statistical test indicates a significant difference between the distribution of multiple datasets, and the difference is substantial, it is recommended to analyze individual datasets alongside the combined dataset for ML development. However, this should be decided on a case-by-case basis as the extent of differences in distribution are typically dependent on the combination of commodity and contaminants being examined. A rationale for analysing the specific datasets separately alongside the combined dataset should be provided, and if no rationale can be found, the combined dataset should be used as a default.
138. When considering the use of individual datasets, it is recommended to compare the statistical results, such as high percentile values of the separate datasets to those of the combined dataset. It should be noted that robust high percentile values cannot be obtained for individual or combined datasets whose sample sizes are lower than the minimum required number of data points (see paragraph 114 and Table 1 in Annex II).
139. It should be noted that when multiple datasets are considered individually, multiple possible MLs may be identified. It is outside the scope of this Guidance to provide guidance on which possible ML to be selected as it is the CCCF that takes a decision on the appropriate ML, taking into account elements that can differ case-by-case.
140. If the datasets from different regions/countries are analysed separately through the statistical methods recommended in this Guidance, the EWG tasked with recommending MLs could provide elements for the choice of the dataset on which to base the ML for consideration and decision by CCCF. For example, if there is assurance that the datasets with high concentrations are for commodities produced under good practices (Codex CoP or GAP, GMP, etc.), then the focus could be on the high concentration datasets to consider globally applicable MLs.

Conducting data analysis by visualization

141. There are different methods of illustrating the chart/graph and plots to show the distribution of occurrence data and statistical values for evaluating the appropriateness of the proposed draft ML. Annex III presents examples of exploratory data analyses and statistics. Depending on the contaminant, the number of data points and distribution of occurrence data (parametric vs non-parametric), exploratory data analyses by visualization may be performed on a case-by-case basis.

Data aggregation and calculation of descriptive statistics

142. The following information and summary statistics can be presented for occurrence datasets, meeting the minimum requirements:
 - Number of total data points.
 - Number of data points lower than the reported LOQs (or LODs), and/or ratio of the number of data <LOQ (or < LOD) among the total number of data points.
 - Range of LOQs (LODs) among data excluded or included in the final data set used to recommend MLs.
 - Mean (arithmetic mean), if the dataset contains datapoints below LOQ (LOD), three arithmetic means based on three substitutional scenarios of LB, MB and UB could be prepared (if the distribution is or close to normal and symmetric).
 - If the distribution is highly skewed, a geometric mean using the same approach as above could be an option but has not been used yet within CCCF.
 - Median (50th percentile values), but if more than 50% of datapoints are below LOQ, the median could be reported as "<LOQ" (or <LOQ).
 - High percentile values (e.g., 95th, 97th and 98th percentile values, as necessary, depending on discussions in the EWG on appropriate rejection rate(s)); if more than 95%, 97% etc. of data points are below the LOQ, then the associated percentiles could be reported as "<LOQ" (or <LOD).
 - Minimum.
 - Maximum: in cases where the maximum was identified as a potential outlier and the maximum value was not excluded from the dataset, it may be worth reporting the 2nd highest value, 3rd highest value, etc. for additional context.

- Range of quantified data.
- Standard deviation, which is a measure of the amount of variation of a parametric distribution.
- Interquartile values (see Annex III), which is a measure of the amount of variation of a non-parametric distribution.

143. Many of these statistics can be easily obtained by using Excel Functions, by using a menu of Descriptive Statistics in Data Analysis tools in MS Excel, or from any other statistical application. Different statistical applications use different calculation protocols and as such return different percentile values for the same dataset. Therefore, when calculating percentile values using computer applications, the values obtained should be carefully checked against the functions used and state the name and if appropriate details of the application used for the calculation.
144. When left-censored data comprise most of the dataset, it may not be possible to calculate high percentile values. In such cases, it is recommended to use the substitution method with LB, MB, or UB scenarios. Although this should be done on a case-by-case basis, depending on the number of quantified values and the contaminant distribution, methods such as estimating high percentile values from probability density functions by modelling the distribution of occurrence data can be used.

Calculation of rejection rates at hypothetical MLs

Estimation of hypothetical MLs

145. From a high percentile value (usually the 95th percentile) of the target dataset, a candidate value for an ML is identified, that takes into consideration the precision of the current analytical method and significant figures of the analytical results. Numerical values for MLs should preferably be regular figures in a geometric scale (0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, etc.), unless this may pose problems in the acceptability of the MLs (Annex I to CXS 193-1995)
146. Once the numerical candidate value of an ML has been determined, the next higher and lower values can also be suggested as hypothetical MLs. (For example, for a candidate ML of 0.5 mg/kg, additional hypothetical MLs could be 0.4 and 0.6 mg/kg). In the case of revision to existing MLs, the existing ML should also be added as one of the hypothetical MLs. Further, values obtained by from the high percentile values (e.g., 95th, 97th and 98th percentile values) can also be used directly as hypothetical MLs.
147. When the decision is made to analyze multiple datasets with different distribution patterns separately, hypothetical MLs are determined from the high percentile values of each dataset. If the distribution patterns are significantly different, hypothetical MLs of individual datasets may be significantly different (see paragraph 139).
148. There is no rule for the number of hypothetical MLs to be proposed, but it is preferable to have more than 2 hypothetical MLs (to be considered case-by-case), for consideration of their impacts on dietary exposure from target commodities and economics arising from rejection rates to be further discussed in the EWG tasked with recommending MLs to the CCCF.
149. The rationale for choosing an ML must be explained clearly in the document prepared for CCCF. The rationale should be based clearly on the analysis, not a pre-chosen value.

Calculation of rejection rates at the hypothetical MLs

150. The rejection rate is defined as the equation below. It can be easily obtained using MS Excel functions (such as COUNTIF function) directly or using statistical or modelling/simulation applications after modelling each dataset. If a different method is used to calculate the rejection rate, the method should be clearly stated in the working document.

$$\text{Rejection rate (\%)} = (\text{number of data points} > \text{hypothetical ML}) / (\text{total number of data points}) \times 100$$

151. It should be noted that the rejection rate may be different from that anticipated from the high percentile due to rounding. The smaller the number of data points used to calculate rejection rates, the greater the uncertainty in estimating the rejection rate. In the calculation of rejection rate, it is assumed that samples that exceed the hypothetical ML are excluded from the market with 100% probability by enforcement of the ML.

Assessment of impact of an ML on rejection rate

152. To assess the impact on international trade of the commodity, the combined global dataset should be used, and if necessary, datasets for each region. Calculating rejection rates on a country-by-country basis is not recommended to be done by the EWG tasked with recommending ML but a Codex Member might bring to the attention of the EWG/CCCF the economic impact certain hypothetical MLs have for their country for consideration, in case this is not sufficiently reflected in the assessment of impact of the hypothetical ML on rejection rate for their region/country.
153. For contaminants known to have large annual variation in concentrations, the rejection rate should be calculated for the dataset of each year, if possible, for year-to-year comparison of rejection rates.

Improvement of calculation of rejection rates

154. At different hypothetical MLs, the calculated rejection rates may not change significantly, depending on the contaminant distribution. Given the skewed nature of contaminant data, the number of data points at the high percentile range is often much lower than that at low percentile range, which affects the estimation of hypothetical MLs and rejection rates (See Figure 1 for the shape of distribution).
155. If the distribution pattern of the combined (potentially global) dataset shows a single peak, modelling/simulation application (such as @Risk, Crystal Ball, R, etc.) can be used to model the distribution to continuously estimate the distribution near the high percentile values from the distribution function and more refined and improved estimates of the rejection rate may be possible.
156. If a more detailed or improved impact assessment regarding rejection rates is needed, requesting an evaluation from JECFA is an option.

Preliminary calculation of effects of MLs on the reduction of dietary exposure to the contaminant from the target commodity at hypothetical MLs

Calculation of dietary exposure and reduction from the target commodity at hypothetical MLs

157. To ensure that the proposed ML is protective of consumers' health, it might be appropriate to quantitatively evaluate the effect of a hypothetical ML in reducing dietary exposure from the target commodity by comparing the exposure with and without an ML. In the case of a revision of an existing ML, the dietary exposure from the target commodity under the already established ML is compared with the exposure under the new hypothetical MLs (revised ML). As calculation of dietary exposure is a risk assessment function that should be undertaken by JECFA, CCCF can request JECFA to conduct evaluation of effects of hypothetical MLs on the reduction of dietary exposure if a detailed evaluation is necessary. [It is important to clarify the roles of JECFA and CCCF as risk assessor and risk manager, respectively, in the calculation of dietary exposure reduction rates when considering MLs. With this understanding, preliminary exposure assessment from the target commodity can also be conducted as reference information in EWGs tasked with recommending MLs if resources are available. For examples of how to calculate the reduction rate of dietary exposure from the target commodity, please refer to Annex IV.]

Improvement of calculation of reduction of dietary exposure reduction rates

158. If a more detailed or improved impact assessment regarding dietary exposure is needed, an evaluation from JECFA might be requested.

Preparing final recommendations to CCCF

159. When finalizing an analysis, the EWG tasked with recommending ML may face challenging questions such as whether an ML is needed based on a preliminary exposure assessment, whether multiple draft proposed MLs are needed, which dataset should be used, or which rejection rate is appropriate. Different perspectives may arise than were envisioned when the TOR for the work were developed. The EWG should not make a final decision but prepare questions and recommendations to be put forward to the CCCF as a whole.
160. If estimated UB dietary exposure is well below the health-based guidance value (HBGV) even without an ML, and a proposed ML is at or about the LOQ value, the ML would have little impact on reducing dietary exposure and would be unnecessary. For contaminants where an HBGV has not been established, if all the data are left-censored, it could be recommended to CCCF to establish the ML(s) at the LOQ value for the time being if there is a potential health concern. However, if most of the data are left-censored and there is no or little health concern, it could be recommended to CCCF that there is no need to establish ML(s). For example, a combined dataset of lead in fresh chicken eggs after data clean-up contained 99% left-censored data points, ranging from 0.001 to 0.257 mg/kg. The calculated exposure reduction at the hypothetical ML was low, and the proposed ML was within the range of reported LOQs. Therefore, development of ML for lead in chicken eggs was discontinued (ref. CX/CF 22/15/7).

161. If the estimated UB dietary exposure is close to or above the HBGV or the margin of exposure is low the development of an ML could be recommended to the CCCF even if a proposed ML is close to the reported LOQs, provided that there is a validated analytical method(s) with appropriate LOQ. If necessary, additional calls for data using more sensitive analytical methods (lower LOQs) may be recommended.
162. When quantified values show a large variation and reach significantly high value(s), it could be recommended to CCCF to develop an ML in order to eliminate highly contaminated foods from the international market. If the contaminant is highly toxic or genotoxic/carcinogenic and found in foods that are consumed in high volumes, it could be recommended to CCCF that the establishment of an ML would protect consumer health, even if rejection rate is low. For example, the combined dataset of total aflatoxins in sorghum grains after clean-up contained 94% of left-censored data, and the upper range of quantified concentrations in this dataset exceeded 200 µg/kg. This indicates that an ML based on a low rejection rate would still have a large impact on reducing dietary exposure to aflatoxins from sorghum grains. A draft ML was recommended to CCCF for aflatoxins in sorghum with a rejection rate of 0.9 % and an intake reduction of 46.5 percent(ref. CX/CF 22/15/9).
163. Overall, if dietary exposure is assessed, EWGs tasked with recommending MLs to CCCF should evaluate the balance between the rejection rate and the reduction of dietary exposure from the target commodity at each hypothetical ML and determine which level is as low as reasonably achievable or offer options to CCCF to inform the Committee's decision.
164. While it is out of the scope of this Guidance to determine what rejection rate is the most appropriate, the EWG tasked with recommending MLs to the CCCF should also consider regional and international consumption patterns to determine a proposed draft ML among the hypothetical MLs with respect to protecting consumer health and ensuring food security and fair trade.
165. It is the responsibility of Codex Members to check the impact of the draft ML (or hypothetical ML(s)) against their own national/regional data and to provide comments on the results of their statistical analysis to the EWG or to the plenary.

DATA PRESENTATION IN EWG REPORTS TO CCCF

Presentation of data analysis: statistical analysis

166. It is important that the data are presented in such a way in the EWG report to CCCF as a working document that they enable an informed discussion on appropriate MLs for deliberation through the Codex Step procedure. This means that the data are reported with inclusion of all assumptions e.g., how many data were excluded and the reasons thereof, how were left-censored data managed, whether data outside the GEMS/Food database were considered, etc.
167. Annex V provides the elements and examples of templates that can be used by the EWG when presenting the results of statistical analyses of occurrence data for ML development. The EWG may use the templates or modify them on a case-by-case basis, as the detail reported will depend on the amount of data available and the nature of the contaminant.

ANNEX I

Determination of outliers/extreme values and handling them

The term “outliers” is defined in the Codex document (CXG 72-2009) as follows:

Outliers: A member of a set of values which is inconsistent with other members of that set

Note:

The following practice is recommended for dealing with outliers.

a) Tests such as Cochran’s (for within laboratory variation) or Grubb’s (for between laboratory variation) tests are applied to identify stragglers or outliers:

- if the test statistic is less than or equal to its 5 % critical value, the item tested is accepted as correct;*
- if the test statistic is greater than its 5 % critical value and less than or equal to its 1 % critical value, the item tested is called a straggler and is indicated by a single asterisk;*
- if the test statistic is greater than its 1 % critical value, the item is called a statistical outlier and is indicated by a double asterisk.*

b) It is next investigated whether the stragglers and/or statistical outliers can be explained by some technical error, for example:

- a slip in performing the measurement,*
- an error in computation,*
- a simple clerical error in transcribing a test result,*
- analysis of the wrong sample.*

Where the error was one of the computation or transcription type, the suspect result should be replaced with the correct value; where the error was from analysing a wrong sample, the result should be placed in its correct cell. After such correction has been made, the examination for stragglers or outliers should be repeated. If the explanation of the technical error is such that it proves impossible to replace the suspect test result, then it should be discarded as a “genuine” outlier that does not belong to the experiment proper.

c) When any stragglers and/or statistical outliers remain that have not been explained or rejected as belonging to an outlying laboratory, the stragglers are retained as correct items, and the statistical outliers are discarded unless the statistician for good reason decides to retain them.

Outliers are a type of extreme values, usually determined by statistical tests as describe above.

Some statistical tests to determine outliers (such as Cochran’s or Grubb’s tests) are available but mostly they assume a normal distribution. Therefore, they are not suitable for applying to the occurrence data on contaminants in foods if they do not follow normal distribution or other distribution that can be converted to normal distribution.

Reference is made to paragraphs 82-92 of the main document

Statistical outlier test

The interquartile (IQR) approach is useful to identify outliers as it can be applied to non-parametric data. This method determines values above “75th percentile values + 1.5 × interquartile” in the dataset as outliers. The IQR approach is widely used as an easy method in a variety of fields to identify outliers. The IQR method assumes that data points equivalent to the median $\pm 2.7\sigma$ in a normal distribution fall within the range of median ± 2 IQR. If the distribution is characterized with high kurtosis and skewness, as in the case of occurrence data of contaminants in food, many data on the right side (higher concentrations) may be determined as outliers. This is because high kurtosis results in smaller IQR and high skewness means greater variability.

Since excluding many data points in the higher concentration range from a dataset will significantly affect the calculation of high percentile values depending on the total number of datapoints of the dataset, and subsequently any ML proposals, it is not recommended to exclude data as outliers solely based on the results of the IQR approach without considering pattern of the contamination (e.g., homogeneous vs heterogeneous).

Other methods to identify possible outliers

Another way to identify possible outliers is to visually inspect the data with a frequency distribution and identify those data which appear disconnected from the rest of the data. However, this is not a sufficient basis to exclude disconnected data as outliers. Figure 2 is an example of EU data on the sum of T-2 and HT-2 toxin in oat milling products, where a relatively small number of data points have levels exceeding 500 µg/kg. In such circumstances, and knowing the nature of the contaminant, these data should remain in the dataset, unless it can be determined that such levels are definitively outside the natural variance of the contaminant.

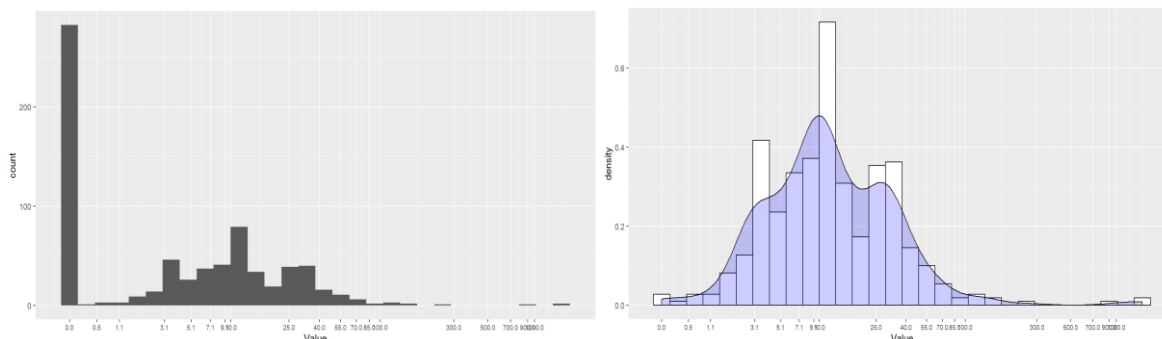


Figure 2. Example EU data on sum of T-2 and HT-2 toxin in oat milling products (717 results of which 438 are quantified results), (left) Histogram: all results, (right) Histogram/probability density curve of quantified results

Decision on the handling possible outliers

There may be cases where extreme values are scientifically valid depending on production and weather conditions and/or other potential factors such as volcanic eruptions, etc. Considering the characteristics of the distribution pattern of occurrence data of a contaminant in food, it is not recommended to simply exclude extreme values based on the results of statistical outlier tests or other methods such as visual inspection. Since the range of concentrations and distribution patterns that can be empirically or theoretically assumed varies significantly depending on the type of contaminant (heavy metals, mycotoxins, etc.), the handling of extreme values must be determined on a case-by-case basis. For example, special consideration should be given to mycotoxins whose concentrations can vary significantly depending on the sampling methods utilized due to the well-known heterogeneous distribution in a lot, as well as very large annual variation.

CXG 72-2009 assumes a dataset of results of repeated analysis of the same sample which shows a normal distribution. It states, “the statistical outliers are discarded unless the statistician for good reason decides to retain them”. In contrast, the datasets addressed in this Guidance are analytical results from a variety of samples and from different analytical methods. Because it is unknown what distribution they will take, and they may be combined from multiple sources, it is difficult to predict the range of variation within a dataset. Therefore, this Guidance recommends, “the statistical outliers are not discarded unless good reason to exclude them is identified and scientifically explained”.

If extreme values are to be excluded as outliers, it is recommended that the reasons for exclusion are clearly reported, and that sensitivity analyses are performed -to show how the exclusion or non-exclusion of outliers may or may not affect the calculation of high percentile values. It should be reiterated that a few extreme values remaining in the dataset will have little effect on the calculation of the high percentile values provided that the total number of data points in the dataset is sufficiently larger than the minimum number of data points required to calculate high percentile values.

Examples of investigation of a set of values inconsistent with other members of that dataset as possible outliers.

Clean-up of data

- a) *If confirmed by the data submitter that the extreme values were caused by errors, decide to exclude these data from further data analysis:*
 - *Outliers (clearly) due to adulteration/fraudulent action or human error (e.g., incorrect data entry) → it can be decided, in consultation and agreement with the data submitter, to exclude these data from further data analysis (or mark them as ‘invalid data’ after confirmation of the data submitter)*
 - *no valid justification to exclude these data can be provided from data submitter or can be explained from EWG for these possible outliers → it can be decided not to exclude these data for further data analysis*
 - *a valid justification can be provided for possible outliers (such as data from a year with extreme weather conditions, data from a specific region/continent, ...) → these data are in principle NOT to be excluded*

Statistical analysis

- b) *Assess the impact of possible outliers on the summary statistics (arithmetic mean, high percentile values). Possible outliers should be retained in the dataset unless a sensitivity analysis results in a significant and meaningful impact on the summary statistics. When there is a significant difference in the results of a sensitivity analysis, the EWG or CCCF will decide on a case-by-case basis whether this difference is meaningful, depending on the toxicity of the contaminant and the type of food, and therefore whether to exclude the possible outliers or not.*
- c) *Conduct an outlier test for extreme values that require further investigation. Statisticians should recommend outlier tests suitable to the dataset for the use by the EWG of the CCCF, if necessary. IQR approach may be an option, but automatic exclusion of outliers should be discouraged unless there is another justification to exclude them. In general, outlier tests that require a normal distribution are not recommended.*

ANNEX II

Minimum number of data points for estimating high percentile values

Reference is made to paragraph 54 of the main document mentioning that for the development of an ML, it is necessary to estimate high percentile values (generally 95 percentile values) of a dataset. A minimum number of 59 datapoints samples is required for a 95th percentile estimation with 95 % confidence.

This Annex II provides details on different alternative options for calculating the minimum number of data points required in relation to estimating high percentile values.

Three options are available for calculating the minimum number of data points required in relation to estimating high percentile values.

- **Option 1:** Calculation based on the concept that a dataset contains one or more values higher than a certain percentile occurring with a probability at a certain confidence level. This is based on the binominal distribution. The minimum number of data points is obtained from the following formula:
 $n = \log(1-CL) / \log(p)$, (CL=1- α , confidence level (0<CL<1); p, percentile/100 (0<p<1); and n, number of data points). In general, a 95% confidence level (CL=0.95) is used in the food safety area. If there is any need for higher confidence level, n can be calculated with a CL higher than 0.95.
- **Option 2:** Calculation based on the rule by Kroes *et al.* (2002)¹³. The minimum number of data points for high percentile (>75th percentile) values can be obtained from the following formula:
 $n \geq 8/(1-p)$, (p, percentile/100 (0<p<1); and n, number of data points).
 No information is available in the reference about confidence levels.
- **Option 3:** Calculation based on the descriptions by Conover (1971)¹⁴ using a binomial distribution and on the concept that a dataset contains one or more values higher than a certain percentile and one or more value lower than the percentile occurring with a probability at a certain confidence level. The minimum number of data points is obtained from the following formula:
 $n \approx 1/4 * x_{1-\alpha} * (1+p)/(1-p) + 1/2$ (CL=1- α , confidence level (0<CL<1); p, percentile/100 (0<p<1); n, number of data points; and $x_{1-\alpha}$, (1- α) quantile of chi-squared random variable and value of $x_{1-\alpha}$ are 9.488 ($\alpha=0.05$), 11.14 ($\alpha=0.025$) and 13.28 ($\alpha=0.01$))

The formula using the same concept as Option 1 (replacing percentile/100 with (1-violation rate/100)) has been used for the compliance tests developed by CCPR (CXG 33-1999) and CCRVDF (CXG 71-2009). Some guidelines¹⁵¹⁶ recommend the numbers obtained from Options 2 and Option 3 (at 99 % confidence level) for a survey design for collection of food consumption data. The following table shows the calculated minimum number of data points for deriving 95th, 96th, 97th, 97.5th and 98th percentile values using the above three options at 95 % confidence level.

Table 1 serves as guide for understanding the minimum number of data points needed to develop hypothetical MLs or to propose MLs. Such a dataset is usually/ideally a global dataset which consists of individual datasets submitted to GEMS/foods or directly to CCCF from countries and/or organizations or observers and covers worldwide occurrence of contamination. If datasets can be combined, it is not necessary that each of individual datasets submitted by member country(ies)/organization(s) or observer(s) contains more data points than the minimum number shown in Table 1. However, when individual datasets per region and/or per year are separately used for deriving high percentile values rather than, or in addition to, the global dataset, the minimum number of data points are required.

Table 1. Minimum number of data points to obtain high percentile values.

	Minimum number of data points to obtain the following percentile values				
Option	95th	96th	97 th	97.5th	98th
Option 1 (CL=0.95)	59	74	99	119	149
[Option 1 (CL=0.99)]	[90]	[113]	[152]	[182]	[228]
Option 2 (CL: not provided*)	160	200	267	320	400
Option 3 (CL=0.95)	93	117	157	188	236
[Option 3 (CL=0.99)]	[130]	[164]	[219]	[263]	[330]

* No information on confidence level is described in the original literature of Kroes *et al.* (2002).

¹³ Kroes et al.2002 "Assessment of intake from the diet." Food and Chemical Toxicology 40.2-3: 327-385.

¹⁴ Conover, 1971. Practical nonparametric statistics. Wiley, New York, USA.

¹⁵ EFSA Journal, 2009. General principles for the collection of national food consumption data in the view of a pan-European dietary survey. 7(12):1435

¹⁶ EFSA Journal, 2014. Guidance on the EU Menu methodology. 12(12):394

Table 1 and the formula for each option can be used to determine how a high percentile value can be calculated with a high confidence level (such as 95%) from the number of data points in the clean-up dataset on a case-by-case basis. Numbers derived from Option 1 have the advantage of requiring the smallest number of data points compared with the other options and are the most feasible and most used in the previous ML development by the CCCF taking into consideration the number of data available to EWGs.

Calculation of confidence level

Confidence level for option 1 in Table 1 can be obtained from the following formula.

$$CL=1-p^n$$

($CL=1-\alpha$, confidence level ($0 < CL < 1$); p , percentile/100 ($0 < p < 1$); and n , number of datapoints)

ANNEX III

Drawing charts/graphs and plots on distribution of occurrence data

As a first step in the statistical analysis, it is recommended to create histograms or box-and-whisker plots for each dataset (e.g., individual and combined datasets) in order to get a perspective on trends in the distribution pattern of the occurrence data. This approach is called exploratory data analysis in statistics. Histograms and box-and-whisker plots can be created using statistical analysis applications, or spreadsheet applications, such as MS Excel.

While various applications can be used, an easy way is provided in Microsoft Excel. To perform drawing and statistical analysis in MS Excel, “Analysis ToolPak” should be installed from the Excel add-ins to use various functions useful for statistical analysis. A ribbon menu for “Data Analysis” will be added to the Data tab of MS Excel after the installation of the add-ins.

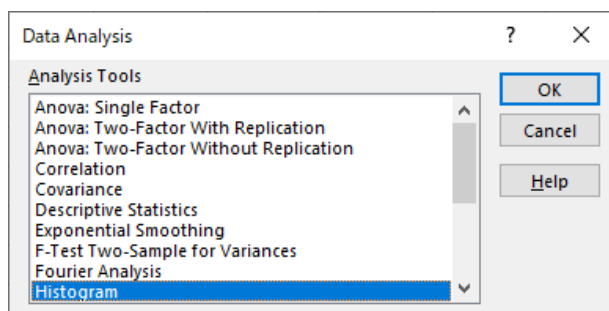


Figure 1. Example of menu of Data Analysis tool in MS Excel

Histograms can be created from the Draw Charts menu as well as from the Data Analysis tool in MS Excel. However, it is recommended to use the Data Analysis tool, which offers greater flexibility in customizing the chart drawing. Box-and-whisker plots can be created from the Draw Chart menu, not from the Data Analysis tool.

In general, histograms provide a good indication of distribution patterns when the number of data points are sufficiently large (ca. 50 or greater). The minimum number of data points needed to calculate high percentile values can be used as guide for the approximate number of data points needed to draw a histogram. If the number of data points contained in the individual datasets differs, the vertical axis should be for relative frequency for easier comparison.

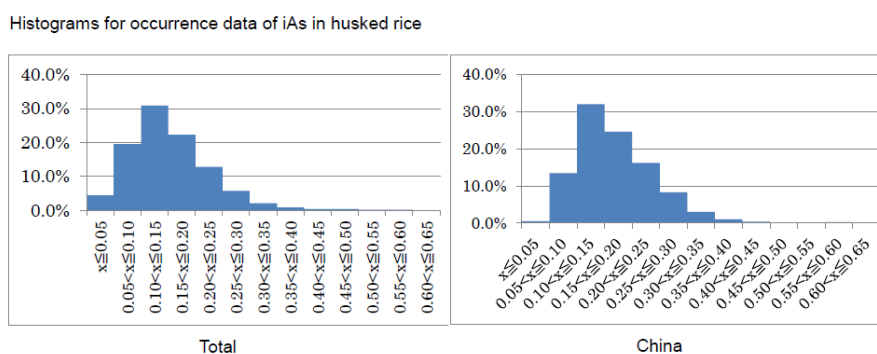


Figure 2. Example of histograms of occurrence data on inorganic As in husked rice (combined and individual dataset) (ref. CX/CF 15/9/7)

A cumulative frequency curve can also be added to a histogram. However, MS Excel alone cannot draw distribution curves. A dedicated statistical analysis application (e.g., SAS, SPSS, R, etc.) should be used, or some add-in applications capable of modelling/simulation (e.g., @Risk, Crystal Ball, etc.) should be used.

For a dataset with certain individual datasets with a small number of data points to be assessed separately, it is difficult to know the shape of the distribution by a histogram for these individual datasets, and a box-and-whisker plot is more helpful because it can be created even when the number of datapoints less than 20. For example, the following box-and-whisker plots for inorganic arsenic concentration data in husked rice from various countries could be drawn when the number of data submitted from several countries was too small (e.g., $n=9$) to draw a histogram. The box-and-whisker plots were drawn for comparisons of individual datasets.

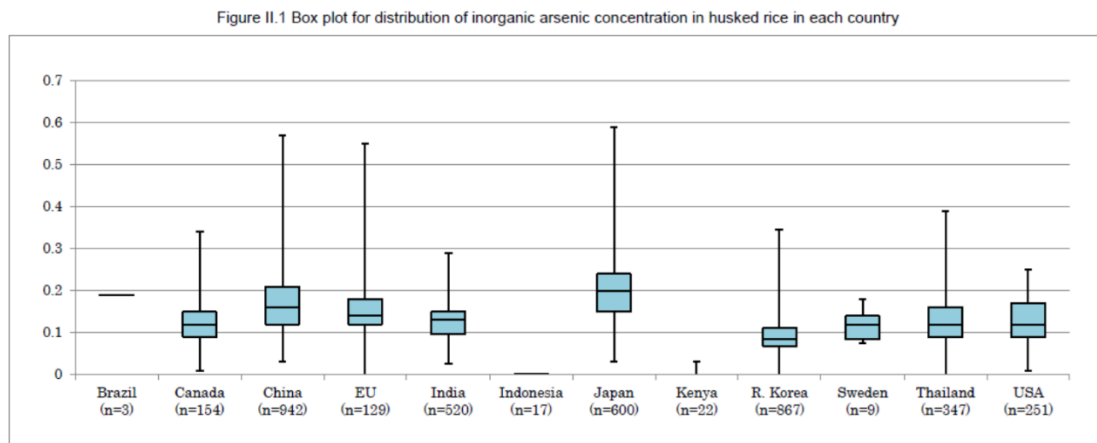


Figure 3. Box-and-whisker plots of individual datasets on inorganic As in husked rice (ref. CX/CF 16/10/5)

After drawing histograms or box-and-whisker plots, it is necessary to check for possible outliers, and differences in the distribution patterns of individual datasets, the shapes of the distributions, and the central tendency and range of the dataset. The presence of multimodalities in the combined dataset should also be checked. When the combined dataset has possible outliers, or clearly shows a multimodal distribution, it is necessary to go back for reconsideration on how to handle the dataset.

[Annex IV

Preliminary calculation of dietary exposure and its reduction rate to the contaminants from the target commodity(ies) at the hypothetical MLs

Annex IV is put in square brackets as it is important to first clearly clarify the roles of JECFA and CCCF as risk assessor and risk manager, respectively, in the calculation of dietary exposure reduction rates when considering MLs.

For all contaminants, long-term/chronic dietary exposure can be calculated using the following equation.

Dietary exposure ($\mu\text{g}/\text{person}\cdot\text{day}$) = concentration ($\mu\text{g}/\text{g}$) $\times$ food consumption ($\text{g}/\text{person}\cdot\text{day}$), or

Dietary exposure ($\mu\text{g}/\text{kg bw}\cdot\text{day}$) = concentration ($\mu\text{g}/\text{g}$) $\times$ food consumption ($\text{g}/\text{kg bw}\cdot\text{day}$)

Reference is made paragraph 96 of the main document.

To estimate average dietary exposure under hypothetical MLs, arithmetic mean concentrations (or geometric mean concentrations in case where the distribution is highly skewed) under hypothetical MLs need to be calculated using a dataset from which higher concentration data than each hypothetical ML are excluded. In the calculation, it is assumed that samples with concentrations above the hypothetical ML are excluded from the market with 100% probability by enforcement of the ML.

Where left-censored data are contained in the occurrence dataset and the distribution permits (e.g., normal distribution), arithmetic means calculated by substitution scenario LB, MB, or UB can be used. For the impact assessment, it is not necessary to calculate using all three scenarios, but the scenario used for calculating arithmetic mean should be reported.

In addition to the arithmetic mean, the geometric mean or median, or high percentile values can be used on a case-by-case basis, particularly when the distribution is highly skewed, to help clearly understand where the central tendency lies for different purposes. Which value was used in the calculation should be clearly stated.

If a dataset contains concentration data that is much higher than the hypothetical ML, the mean concentration will be significantly lowered after excluding those extremely high values (see the part on determination of outliers/extreme values and handling them – paragraphs 82–92). Since the median is a robust statistic, if the number of data points in the dataset does not change significantly after excluding data points that are much higher than the hypothetical ML, the median under the hypothetical ML may change little compared to the median without ML.

A template for calculating the point estimates of chronic dietary exposure (International Estimated Dietary Intake, IEDI) using the GEMS/Food cluster diets (explained below) is available from the URL of the GEMS/Food Programme¹⁷. In this template, all countries are grouped into 17 clusters and for each cluster there are food consumption data derived from the data in the FAO Food Balance Sheet and some additional data provided by the governments. After entering the concentration data (in this case, the arithmetic mean concentrations calculated for each hypothetical MLs or other alternative values) and clicking the button “Make table”, there will be calculated mean intakes for 17 clusters.

For foods for which consumption data are not available in the GEMS/Food cluster diet, it may be possible to estimate consumption per capita in the population from available data, such as production volumes. As for food consumption data, other databases are available (e.g., FAO/WHO GIFT, FAO/WHO CIFOCS, etc.). What data were used and how it was processed should be indicated in the EWG’s report.

Percentage of decrease in dietary exposure of the contaminant from the foods or food groups concerned under hypothetical ML compared to exposure without ML is regarded as a reduction rate of exposure.

Reduction rate of exposure (%) = (exposure with no ML – exposure with ML) / (exposure with no ML) $\times$ 100

For a contaminant for which an ARfD has been established by JECFA, acute/short term exposure should be calculated using the International Estimated Short-term Intake (IESTI) template available at the same URL as the IEDI template. Acute/short term exposure under a hypothetical ML should be well below the ARfD for the general population or children 6 and below, or if ARfD is set for women of child-bearing age, should be well below this ARfD. An ARfD has been recommended only for DON and related compounds, how to conduct the IESTI calculation is not explained here in detail.

If a HBGV is established by JECFA for a contaminant/toxin (PMTDI/PTDI, PTWI, PTMI or ARfD, etc.), it may be useful to evaluate the impact on the percentage of exposure from the food to which the ML applied relative to the HBGV. Information on average body weight should be obtained when comparing dietary exposure to HBGV.

Ratio to the HBGV (%) = exposure with ML ($\mu\text{g}/\text{person}\cdot\text{day}$) / average body weight (kg/person) / HBGV ($\mu\text{g}/\text{kg bw}/\text{day}$) $\times$ 100

If HBGVs are set on a per week or per month basis, the exposure should be multiplied by the appropriate factor, e.g., 7 or 30.]

¹⁷ <https://www.who.int/teams/nutrition-and-food-safety/databases/global-environment-monitoring-system-food-contamination>

ANNEX V

Presentation of data analysis/statistical analysis

(Annex III could be integrated into this Annex given that Annex III relates to visualise the distribution of occurrence data that is an important aspect of presentation of occurrence data).

This annex provides the elements and examples of templates that can be used by the EWG when presenting the results of statistical analysis of occurrence data for ML development. The EWG may use the templates or modify them on a case-by-case basis, as the detail of reporting depends on the number of data points available and the nature of the contaminant.

For each dataset provided by Members (and Observers) used in statistical analysis, the following elements should be reported:

- Number of data points and proportion of <LOQ data.
- Arithmetic mean (LB, MB and/or UB), median, minimum and maximum.
- Relevant percentile values (e.g., 95th, 97th, 98th); and
- Charts/graphs or plots showing distribution patterns (such as histograms or box-and-whisker plots)

In reporting the above elements, Table 1 can be used as a template.

Table 1. Example Template: Summary of statistics of the datasets

Food or food group	Total number of data points	Number or percent <LOQ	Mean (LB-UB) (mg/kg)	Median (LB-UB) (mg/kg)	95 th %ile (LB-UB) (mg/kg)	97 th %ile (LB-UB) (mg/kg)	98 th %ile (LB-UB) (mg/kg)	Min (mg/kg)	Max (mg/kg)

Note to table 1: To be filled for the combined dataset (potentially global dataset), %ile: percentile value.

If necessary, the following elements may also be useful for interpreting the dataset:

- range of reported LOQs; and
- standard deviation (unbiased) most relevant for the mean LB and mean UB.

If there is significant year-to-year variation in occurrence for a food or food group, it is appropriate to provide an analysis of the data per year using Table 2b and in case of a statistically significant difference in distribution pattern by geographical region, the analysis should consider presenting data per region (e.g., continent or Codex region) using Table 2a.

Table 2a. Example Template: Summary of statistics of the datasets per region for a specific commodity/food (category)

Region/Continent/World	Total number of data points	Number or percent <LOQ	Mean (LB-UB) (mg/kg)	Median (LB-UB) (mg/kg)	95 th %ile (LB-UB) (mg/kg)	97 th %ile (LB-UB) (mg/kg)	98 th %ile (LB-UB) (mg/kg)	Min (mg/kg)	Max (mg/kg)

Note to Table 2a: To be filled in by region or global, %ile: percentile value.

Table 2b. Example Template: Summary of statistics of the datasets per year for a specific commodity/food (category)

Year	Total number of data points	Number or percent <LOQ	Mean (LB-UB) (mg/kg)	Median (LB-UB) (mg/kg)	95 th %ile (LB-UB) (mg/kg)	97 th %ile (LB-UB) (mg/kg)	98 th %ile (LB-UB) (mg/kg)	Min (mg/kg)	Max (mg/kg)

Note to Table 2b: To be filled in by per year/global or per year/region, %ile: percentile values

If the number of data points is higher than the minimum number of data points required for an individual food or an indication that the distribution of concentrations for an individual food differs significantly from other foods despite the lower number of data points, it is important to present a summary of data for all individual foods within a food group, in addition to summary data for the broader food group. This type of analysis allows for an understanding on how a proposed ML impacts the individual foods and to determine if the development of an ML for a broad food category is more appropriate than an ML for individual foods within the broad category.

In general, two to four hypothetical MLs are estimated based on the distribution. Data presentation should cover these hypothetical MLs showing how the dietary exposure estimates are affected by them, and what the rejection rates would be using Table 3.

If necessary, a table, using Table 3 as a template, should be prepared for not only combined (potentially global) dataset but also regional datasets to consider the impact of MLs by geographical area.

Table 3. Example Template: Summary of impact assessment at hypothetical MLs

Hypothetical MLs (mg/kg)	Mean concentration (mg/kg)	Dietary intake (µg/person/day)	Exposure reduction (%)	Rejection rate (%)
Name of Food or food group (total number of data points)				

Note to table 3: Exposure reduction = percentage of decrease in dietary intake of the contaminant from the food or food group concerned

Finally, a working document prepared for discussion at the plenary session of CCCF should accompany a table, using Table 4 as a template and in line with the format of GSCTFF, which includes information on the proposed draft MLs as well as explanation notes for the commodities under consideration. The explanation of this format and its contents are provided in the GSCTFF.

Table 4. Format of the schedule in the GSCTFF

Commodity/Product name	Proposed draft MLs (mg/kg)	Portion of the commodity/Product to which the ML applies	Notes/Remarks

ANNEX VI

Glossary of terms

Term	Definition/Explanation
Acute reference dose (ARfD)	The estimate of the amount of a substance in food or drinking water, expressed on a body weight basis, that can be ingested in a period of 24 h or less without appreciable health risk to the consumer. It is derived on the basis of all the known facts at the time of evaluation. The ARfD is expressed in milligrams of the chemical per kilogram of body weight. (WHO, EHC 240)
Box-and-whisker plot	A graphical method of displaying the important characteristics of a set of observations. The display is based on the five-number summary of the data with the 'box' part covering the inter-quartile range, and the 'whiskers' extending to include all but outside observations, these being indicated separately. It is often particularly useful for comparing the characteristics of different samples. (Cambridge dictionary of statistics) A box contains data that falls between 25th and 75th percentile and the ends of whisker show in general the minimum and maximum values. Outliers are plotted away from whisker. The median (50th percentile) is shown in the box above 25th percentile and below 75th percentile. The mean is plotted with a symbol, such as "X".
Cochran's test	One of the tests used to identify outliers, which is a test of the within-laboratory variabilities. It should be applied first, then any necessary action should be taken, with repeated tests if necessary. (ISO 5725-2)
Confidence level	A measure of the reliability of a result. A confidence level of 95% or 0.95 means that there is a probability of at least 95% that the result is reliable.
Critical value	The value of the net concentration or amount the exceeding of which leads, for a given error probability α, to the decision that the concentration or amount of the analyte in the analysed material is larger than that in the blank material. It is defined as: $\Pr(\hat{L} > L_c L=0) \leq \alpha$ Where: $\hat{L}$ is the estimated value, L is the expectation or true value, and L_c is the critical value. (CXG72-2009)
Distribution curve	A graph of the frequencies of different values of a variable in a statistical distribution.
Exposure assessment	Exposure assessment is the qualitative and/or quantitative evaluation of the likely intake of biological, chemical, and physical agents via food as well as exposures from other sources if relevant. (Codex Alimentarius Commission Procedural Manual) For the purpose of this document, the term "dietary exposure" refers to the intake of a substance by a person as part of its diet (via food, beverages, drinking water and food supplements).
Extreme value	The largest and smallest variables among a sample of observations. (Cambridge dictionary of statistics) For the purpose of this document, the maximum and nearby values in a dataset are referred to as extreme values.
Quantified value	A number that is not infinite, i.e., it could be measured or given a value. For the purpose of this document, finite value means any analytical results at or higher than reported LOQ.
Food group	A set of individual food commodities with similar biological and morphological characteristics and therefore similar potential for concentrations of a chemical of concern and for which a common group maximum level can be set.

Term	Definition/Explanation
Gamma distribution	The probability distribution, $f(x)$, given by $f(x) = \frac{x^{\gamma-1} \exp(-x\beta)}{\beta \Gamma(\gamma)}, \quad 0 \leq x < \infty, \beta > 0, \gamma > 0$ β is a scale parameter and γ a shape parameter. The mean, variance, skewness and kurtosis of the distribution are as follows. $\begin{aligned} \text{mean} &= \beta\gamma \\ \text{variance} &= \beta^2\gamma \\ \text{skewness} &= 2\gamma^{-\frac{1}{2}} \\ \text{kurtosis} &= 3 + \frac{6}{\gamma} \end{aligned}$ The distribution of $u = x/\beta$ is the standard gamma distribution with corresponding density function given by $f(u) = \frac{x^{\gamma-1} e^{-x}}{\Gamma(\gamma)}$ The function Γ defined by $\Gamma(r) = \int_0^{\infty} t^{r-1} e^{-t} dt$ Where $r > 0$ (r need not be an integer). The function is recursive satisfying the relationship. $\Gamma(r+1) = r\Gamma(r)$ (Cambridge Dictionary of Statistics)
GEMS/Food	The World Health Organization's Global Environment Monitoring System – Food Contamination Monitoring and Assessment Programme (GEMS/Food Programme), which maintains databases on contaminant levels in foods and estimates of dietary exposure to food chemicals. (WHO, EHC 240)
Grubbs' test	One of the tests to identify outliers, which is primarily a test of between-laboratory variability and can also be used (if $n > 2$) where Cochran's test has raised suspicions as to whether the high within-laboratory variation is attributable to only one of the test results. (ISO 5725-2)
Health-based guidance value (HBGV)	A numerical value derived by dividing a point of departure (a no-observed-adverse-effect level, benchmark dose or benchmark dose lower confidence limit) by a composite uncertainty factor to determine a level that can be ingested over a defined time period (e.g., lifetime or 24 h) without appreciable health risk. (WHO, EHC 240)
Histogram	A graphical representation of a set of observations in which class frequencies are represented by the areas of rectangles centred on the class interval. If the latter are all equal, the heights of the rectangles are also proportional to the observed frequencies. (Cambridge Dictionary of Statistics)
International estimated daily intake (IEDI)	A prediction of the long-term daily intake of a pesticide residue on the basis of the assumptions of average daily food consumption per person and median residues from supervised trials, allowing for residues in the edible portion of a commodity and including residue components defined by the Joint FAO/WHO Meeting on Pesticide Residues for estimation of dietary intake. Changes in residue levels resulting from preparation, cooking or commercial processing are included. When information is available, dietary intake of residues resulting from other sources should be included. The IEDI is expressed in milligrams of residue per person. (WHO, EHC 240)

Term	Definition/Explanation
International estimated short-term intake (IESTI)	A prediction of the short-term intake of a pesticide residue on the basis of the assumptions of high daily food consumption per person and highest residues from supervised trials, allowing for residues in the edible portion of a commodity and including residue components defined by the Joint FAO/WHO Meeting on Pesticide Residues for estimation of dietary intake. The IESTI is expressed in milligrams of residue per kilogram of body weight. (WHO, EHC 240)
Interquartile range	A measure of spread given by the difference between the first quartile (25th percentile) and third quartile (75th percentile) of a sample. (Cambridge Dictionary of Statistics)
Joint FAO/WHO Expert Committee on Food Additives (JECFA)	An expert committee that has been meeting since 1956. JECFA has been engaged in collecting and evaluating scientific data on food additives and making recommendations on safe levels of use. This has been accomplished 1) by elaborating specifications for the identity and purity of individual food additives that have been toxicologically tested and are in commerce and 2) by evaluating toxicological data on these food additives and estimating acceptable intakes by humans. In 1972, the scope of the evaluations was extended to include contaminants in food, whereas in 1987, the scope was extended even further to include residues of veterinary drugs in food. (WHO, EHC 240)
Joint FAO/WHO Meeting on Pesticide Residues (JMPR)	The abbreviated title for the Joint Meeting of the FAO Panel of Experts on Pesticide Residues in Food and the Environment and the WHO Core Assessment Group on Pesticide Residues, which has been meeting since 1963. The meetings are normally convened annually. The FAO Panel of Experts is responsible for reviewing residue and analytical aspects of the pesticides considered, including data on their metabolism, fate in the environment and use patterns, and for estimating the maximum residue levels and supervised trials median residue levels that might occur as a result of the use of the pesticide according to Good Agricultural Practice. The WHO Core Assessment Group on Pesticide Residues is responsible for reviewing toxicological and related data on the pesticides and, when possible, for estimating acceptable daily intakes and long-term dietary intakes of residues. As necessary, acute reference doses for pesticides are estimated along with appropriate estimates of short-term dietary intake. (WHO, EHC 240)
kurtosis	The extent to which the peak of a unimodal probability distribution or frequency distribution departs from the shape of a normal distribution, by either being more pointed (<i>leptokurtic</i>) or flatter (<i>platykurtic</i>). (Cambridge Dictionary of Statistics)
Kruskal-Wallis H-test	A distribution free method that is the analogue of the analysis of variance of a one-way design. It tests whether the group to be compared have the same population median. The test statistic is derived by ranking all the N observations from 1 to N regardless of which group they are in, and then calculating $H = \frac{12 \sum_{i=1}^k n_i (\bar{R}_i - \bar{R})^2}{N(N-1)}$ Where n_i is the number of observations in group i, $\bar{R}_i$ is the mean of their ranks, $\bar{R}$ is the average of all the ranks, given explicitly by $(N+1)/2$. When the null hypothesis is true the test statistic has a chi-squared distribution with $k-1$ degrees of freedom. (Cambridge Dictionary of Statistics)
Left-censored data	Data which are not finite values (quantified values) or are data less than reported LOQs (or LODs in case the LOQ is not reported).

Term	Definition/Explanation
Limit of detection (LOD)	The true net concentration or amount of the analyte in the material to be analysed which will lead, with probability $(1-\beta)$, to the conclusion that the concentration or amount of the analyte in the analysed material is larger than that in the blank material. It is defined as: $\Pr(\hat{L} \leq LC L=LOD) = \beta$ Where $\hat{L}$ is the estimated value, L is the expectation or true value, and LC is the critical value. (CXG 72-2009) OR The limit of detection (LOD) is the minimum concentration of a contaminant that can be qualitatively measured in the specific food. The limit of detection is reported by a laboratory, or a value calculated from the LOQ. (GEMS/Food Programme)
Limit of quantification (LOQ)	A method performance characteristic generally expressed in terms of the signal or measurement (true) value that will produce estimates having a specified relative standard deviation (RSD), commonly 10% (or 6%). LOQ is estimated by: $LOQ = kQ \sigma_Q$, $kQ = 1/RSD_Q$ Where LOQ is the limit of quantification, σ_Q is the standard deviation at that point and kQ is the multiplier whose reciprocal equals the selected RSD (The approximate RSD of an estimated σ, based on v-degrees of freedom is $1/\sqrt{2v}$). (CXG 72-2009) OR The limit of quantification (LOQ) is the minimum concentration of a contaminant that can be quantitatively measured in the specific food with an acceptable level of accuracy and precision. The limit of quantification is reported by a laboratory, or a value calculated from the LOD. (GEMS/Food Programme)
Lognormal distribution	The probability distribution of a random variable, X, for which $\ln(X)$ has a normal distribution with mean μ and variance σ^2. The distribution is given by $f(x) = \frac{1}{x\sigma(2\pi)^{\frac{1}{2}}} \exp\left[-\frac{1}{2\sigma^2} \{-(\ln x - \mu)^2\}\right] \quad 0 \leq x < \infty$ The mean, variance, skewness and kurtosis of the distribution are $\begin{aligned} \text{mean} &= \exp\left(\mu + \frac{1}{2}\sigma^2\right) \\ \text{variance} &= \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1) \\ \text{skewness} &= (\exp(\sigma^2) + 2)(\exp(\sigma^2) - 1)^{\frac{3}{2}} \\ \text{kurtosis} &= \exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 3 \end{aligned}$ For small the distribution is approximated by the normal distribution. (Cambridge Dictionary of Statistics)
Maximum level (ML)	Codex Maximum Level for a contaminant in a food or feed commodity is the maximum concentration of that substance recommended by the Codex Alimentarius Commission to be legally permitted in that commodity. (Codex Alimentarius Commission Procedural Manual)
Mann-Whitney U test	A distribution free method used as an alternative to the Student's t-test for assessing whether two populations have the same location. Given a sample of observations from each population, all the observations are ranked as if they were from a single sample, and the test statistic is the sum in the smaller group. Tables giving critical values of the test statistic are available, and for moderate and large sample sizes, a normal approximation can be used. (Cambridge Dictionary of Statistics)
Measurement uncertainty	Parameter, associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand (i.e., the quantity intended to be measured). (CXG54-2004, amended in 2021)

Term	Definition/Explanation
Mean (Arithmetic)	A measure of location or central value for a continuous valuable. For a sample of observations $x_1, x_2, \dots, x_n$ the measure is calculated as $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ Most useful when the data have a symmetric distribution and do not contain outliers. (Cambridge Dictionary of Statistics)
Mean (Geometric)	A measure of location, g, calculated from a set of observations $x_1, x_2, \dots, x_n$ as $g = \left(\prod_{j=1}^n x_j \right)^{\frac{1}{n}}$ (Cambridge Dictionary of Statistics)
Median	The value in a set of ranked observations that divides the data into two parts of equal size. When there is an odd number of observations the median is the middle value. When there is an even number of observations the measure is calculated as the average of the two central values. Provides a measure of location of a sample that is suitable for asymmetric distributions and is also relatively insensitive to the presence of outliers. (Cambridge Dictionary of Statistics)
Multimodal distribution	A probability distribution or frequency distribution that has two or more modes (peaks). Multimodality is often taken as an indication that the observed distribution results from the mixing of the distributions of relatively distinct groups of observations. (Cambridge Dictionary of Statistics)
Normal distribution	A probability distribution, $f(x)$ of a random variable, X, that is assumed by many statistical methods. Specifically given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right]$ Where μ and σ^2 are, respectively, the mean and variance of x. (Cambridge Dictionary of Statistics)
Outlier	A member of a set of values which is inconsistent with other members of that set. (CXG72-2009)
Parametric test/non-parametric test	Parametric tests assume the data follows a certain distribution model, which is mostly a normal distribution. Non-parametric tests do not assume the data have any particular distribution and can analyse data where no parametric test is applicable.
Percentile	The set of divisions that produce exactly 100 equal parts in a series of continuous values, such as concentration of a certain contaminant in food. For example, a sample with concentration above the 95th percentile has a higher concentration than over 95 % of the other contaminant levels. (Cambridge Dictionary of Statistics, modified for the purpose of this guidance)
Primary producing county	A country where the commodities are grown and harvested
Provisional Maximum Tolerable Daily Intake (PMTDI)/Provisional Tolerable Daily Intake (PTDI)	A type of HBGV. An endpoint used for contaminants with no cumulative properties. Its value represents permissible human exposure as a result of the natural occurrence of the substance in food and in drinking-water. In the case of trace elements that are both essential nutrients and unavoidable constituents of food, a range is expressed, the lower value representing the level of essentiality and the upper value the PMTDI. (CXS193-1995)

Term	Definition/Explanation
Provisional Tolerable Weekly Intake (PTWI)	A type of HBGV. An endpoint used for food contaminants such as heavy metals with cumulative properties. Its value represents permissible human weekly exposure to those contaminants unavoidably associated with the consumption of otherwise wholesome and nutritious foods. (CXS193-1995)
Provisional Tolerable Monthly Intake (PTMI)	A type of HBGV. An endpoint used for a food contaminant with cumulative properties that has a very long half-life in the human body. Its value represents permissible human monthly exposure to a contaminant unavoidably associated with otherwise wholesome and nutritious foods. (CXS193-1995)
Quality assurance (in analytical laboratories)	All those planned and systematic actions necessary to provide adequate confidence that analytical results will satisfy given requirements for quality (CXG72-2009) There are a number of Codex recommendations on quality assurance based on the considerations by the Codex Committee on Methods of Analysis and Sampling (e.g., CXG27-1997, CXG28-1995, CXG64-1995, and CXG65-1997.)
Rejection rate/violation rate	Rejection rate (or violation rate) (%) = (number of data points with higher concentrations of a contaminant than ML)/total number of data points x 100.
Sensitivity analysis	Sensitivity analysis shows how different values of an independent variable affect a dependent variable under a given set of assumptions. In the case of occurrence data it is used to assess the impact that exclusion of data have on e.g. average, median and high percentile values.
Skewness	The lack of symmetry in a probabilistic distribution. (Cambridge Dictionary of Statistics)
Standard deviation	The most commonly used measure of the spread of a set of observations. Equal to the square root of variance s^2 , that is given by the following formula: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ where $x_1, x_2, \dots, x_n$ are the n sample observations and $\bar{x}$ is the sample mean. (Cambridge Dictionary of Statistics)
Statistical power	In statistics, the power of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis when a specific alternative hypothesis is true. The null hypothesis of Mann-Whitney U test or Kruskal-Wallis H-test is that each dataset comes from the same population.
Random sampling	Is a type of sampling. The term “random sampling” should be chosen for routine sampling, even if targeted at specific food types or specific importing countries. Testing a wide range of imported samples of a certain food category for the presence of a certain contaminant would be “random”
Targeted sampling	Is a type of sampling. The term “targeted sampling” should be chosen for follow-up sampling following specific findings of contamination. For example, if a country identifies a sample from a particular manufacturer as having high levels of a contaminant, additional sampling of the same lot or lots produced at the same time by the same manufacturer would be “targeted.”

APPENDIX II

Issues identified for possible future discussion

- Minimum number of samples for estimating high percentile values with high confidence (REP23/CF16, paras 93 and 94, para 98 (vi) (a)).
- Further guidance to be provided on which dataset the ML should be based or to which database should be given priority for ML development (combined dataset, dataset showing the higher contamination patterns as long as the commodity was produced through good practice, datasets from major producing countries or regions, datasets from importing countries reflecting the levels of a contaminant in a commodity in international trade, dataset to be used to be decided on a case-by-case) (REP23/CF16, para 98 (vi) (b)).
- Further consider the role of the Committee in calculating dietary exposure reduction rates when considering MLs. (calculation of dietary exposure is a risk assessment function that should be undertaken by JECFA and JECFA provides the scientific advice on which the risk management decisions of the Committee are based – it is important to clarify the roles of JECFA and CCCF as risk assessor and risk manager respectively, in the calculation of dietary exposure reduction rates when considering MLs (REP23/CF16, paras 90 and 91 and para 98 (vi) (c). (see also paragraph 17 (ii) of Recommendations to CCCF18 – in case a conclusion is reached on this topic at CCCF18, no future work would be needed).
- More structured process for elaborating calls for data (REP23/CF16, para 98 (viii)).
- Identification of appropriate rejection rates in ML establishment (guidance on elements which should be considered to define the appropriate rejection rate) (CX/CF 22/15/14 chapter IV of appendix I, CF16/CRD06 para 32).
- Appropriateness of GEMS/Food market-based cluster diets for ML elaboration (reconcile realistic estimates from national consumption data with the “supply utilization market data in GEMS/Food cluster diets (e.g., sugar, spices/herbs, teas, coffees).
- Development of guidance on data weighting in analyses where the number of data points is significantly different between individual datasets from different countries or regions.

APPENDIX III**List of Participants****CHAIR****European Union**

Frans Verstraete
European Commission

CO-CHAIRS**Japan**

Tetsuo Urushiyama
Ministry of Agriculture, Forestry and Fisheries

Yoshiyuki Takagishi
Ministry of Agriculture, Forestry and Fisheries

The Netherlands

Nikki Emmerik
Ministry of Health, Welfare and Sport

Moniek Ringenier
Ministry of Health, Welfare and Sport

United States of America

Lauren Posnick Robin
U.S. Food and Drug Administration

Eileen Abt
U.S. Food and Drug Administration

MEMBER COUNTRIES AND MEMBER ORGANIZATIONS**ARGENTINA**

Maria Alejandra Rodriguez
Instituto Nacional de Tecnologia Industrial

Martin Fernández
Instituto Nacional de Alimentos

Gisele Simondi
Instituto Nacional de Alimentos

Silvana Ruarte
Instituto Nacional de Alimentos

AUSTRALIA

Matthew Joseph O'Mullane
Food Standards Australia New Zealand

Nick Fletcher
Food Standards Australia New Zealand

BELGIUM

Christine Vinkx
FPS Health, Food Chain Safety and Environment

Andrea Carletta
FPS Health, Food Chain Safety and Environment

Fabio Enrico Occhetti
Federal Agency for the Safety of the Food Chain

BRAZIL

Larissa Bertollo Gomes Porto
ANVISA

Giselle Kindlein
Ministry of Agriculture

BRUNEI

Nur Halimatussaadia Mohd Alias
Brunei Darussalam Food Authority

CANADA

Ian Richard
Health Canada | Santé Canada

Carla Hilts
Health Canada

Stephanie Glanville
Health Canada

Rosalie Awad
Health Canada

CHILE

Lorena Delgado Rivera
Instituto de Salud Pública de Chile

CHINA

Yi Shao
China National Center of Food Safety Risk Assessment (CFSA)

Yongning Wu
National Center of Food Safety Risk Assessment (CFSA)

Dawei Chen

Shuang Zhou
National Center of Food Safety Risk Assessment (CFSA)

ECUADOR

Saul Flores
AGROCALIDAD

EUROPEAN UNION

Veerle Vanheusden
European Commission

FRANCE

Jean-Cédric Reninger
ANSES

Niels Enslen
Ministry of Agriculture

INDIA

Reeba Abraham
Ministry of Commerce and Industry (APEDA)

Aditya Jain
National Dairy Development Board

INDONESIA

Yeni Restiani
Indonesian Food and Drug Authority

IRAN

Mansoor Mazaheri
ISIRI-Standard Research Institute

JAPAN

Daisuke Aoki
Consumer Affairs Agency

Mio Toda
National Institute of Health Sciences

MEXICO

Tania Daniela Fosado Soriano
Secretaría de Economía

Codex Committee on Fresh Fruits and Vegetables (CCFFV)
secretariat
CCFFV

MOROCCO

Ouazzani Sanae
ONSSA

NETHERLANDS

Weiluan Chen
RIVM

Astrid Bulder
RIVM

NEW ZEALAND

Jeane Nicolas
Ministry for Primary Industries, New Zealand Food
Fiapaipai Ruth Auapaau
Ministry for Primary Industries

NORWAY

Olga Soleng
Norwegian Food Safety Authority

NIGERIA

Abu Rachel Kakataidii
NAFDAC

Mazai Maymunah Ummjamil
Standards Organisation of Nigeria

PHILIPPINES

Phelan Apostol
Food and Drug Administration Philippines

REPUBLIC OF KOREA

Republic of Korea/ Codex Secretariat
Ministry of Agriculture, Food and Rural Affairs

Kim Jooyeon
Ministry of Food and Drug Safety

Heewon Moon
Ministry of Food and Drug Safety

RUSSIAN FEDERATION

Darya Bagreeva
Federal Scientific Center of Hygiene

RWANDA

Rosine Niyonshuti
Rwanda Standards Board

SAUDI ARABIA

Mohammed Mousa Ali Alshehri
Saudi Food and Drug Authority

Lama Almainan
Saudi Food and Drug Authority

SENEGAL

Mame Diarra Faye
Ministère de la Santé et l'Action sociale

SINGAPORE

How Chee Ong
Singapore Food Agency

Er Jun Cheng
Singapore Food Agency

Wesley Zongrong Yu
Singapore Food Agency

SOUTH AFRICA

Juliet Masuku
Department of Health

SUDAN

Liza Nelson Michael Taban
Sudan Bureau of Standards

SWEDEN

Nurun Nahar
Swedish Food Agency

SWITZERLAND

Judit Valentini
Federal Food Safety and Veterinary Office

THAILAND

Chutiwan Jatupornpong
Ministry of Agriculture and Cooperatives, National Bureau of
Agricultural Commodity and Food Standards

TÜRKIYE

Bengi Akbulut Pınar
Ministry of Agriculture and Forestry

Sinan Arslan
Ministry of Agriculture and Forestry

UGANDA

Francis Enaru
Ministry of Trade, Industry and Cooperatives

UNITED KINGDOM

Mark Willis
Food Standards Agency

Craig Jones
Food Standards Agency

Helen Twyble
Food Standards Agency

Izaak Fryer-Kanssen
Food Standards Agency

Holly Howell-Jones
Food Standards Agency

Clare McCartney-Collard
Food Standards Agency

UNITED STATES OF AMERICA

Sara McGrath
U.S. Food and Drug Administration

Linda Benjamin
U.S. Food and Drug Administration

Quynh-Anh Nguyen
FDA/CFSAN

Alexandra Ferraro
U.S. Codex Office | U.S. Department of Agriculture

Armen Mirzoian
Alcohol and Tobacco Tax and Trade Bureau (TTB) – Scientific
Services Division

URUGUAY

Santiago Viera
Laboratorio Tecnológico del Uruguay

Natalie Merlinski
Ministry of Livestock, Agriculture and Fisheries

ZAMBIA

Henry Njapau
Zambia Academy of Sciences

NON-GOVERNMENTAL ORGANIZATIONS**INTERNATIONAL COUNCIL OF BEVERAGES ASSOCIATIONS**

Simone SooHoo

Maia Jack

INTERNATIONAL ORGANISATION OF SPICE TRADE ASSOCIATIONS

Shannen Kelly

INTERNATIONAL SPECIAL DIETARY FOODS INDUSTRIES

Marian Brestovansky

Jean Christophe Kremer

FOODDRINKEUROPE

Alejandro Rodarte

THIE | TEA & HERBAL INFUSIONS EUROPE

Farshad Rostami

FEDIOL | EU VEGETABLE OIL AND PROTEINMEAL INDUSTRY ASSOCIATION

Kalila Hajjar

AMERICAN PEANUT COUNCIL

Jim Elder