

WORKSHOP ON DATA COLLECTION AND ANALYSIS FOR CODEX PROCEEDINGS

Data Analysis for the Development of Maximum Levels (MLs) – Part 2

Dr. Elie Bou Yazbeck, GForSS

2025 Global Food Regulatory Science Society (GForSS). All rights reserved.

Learning Objectives

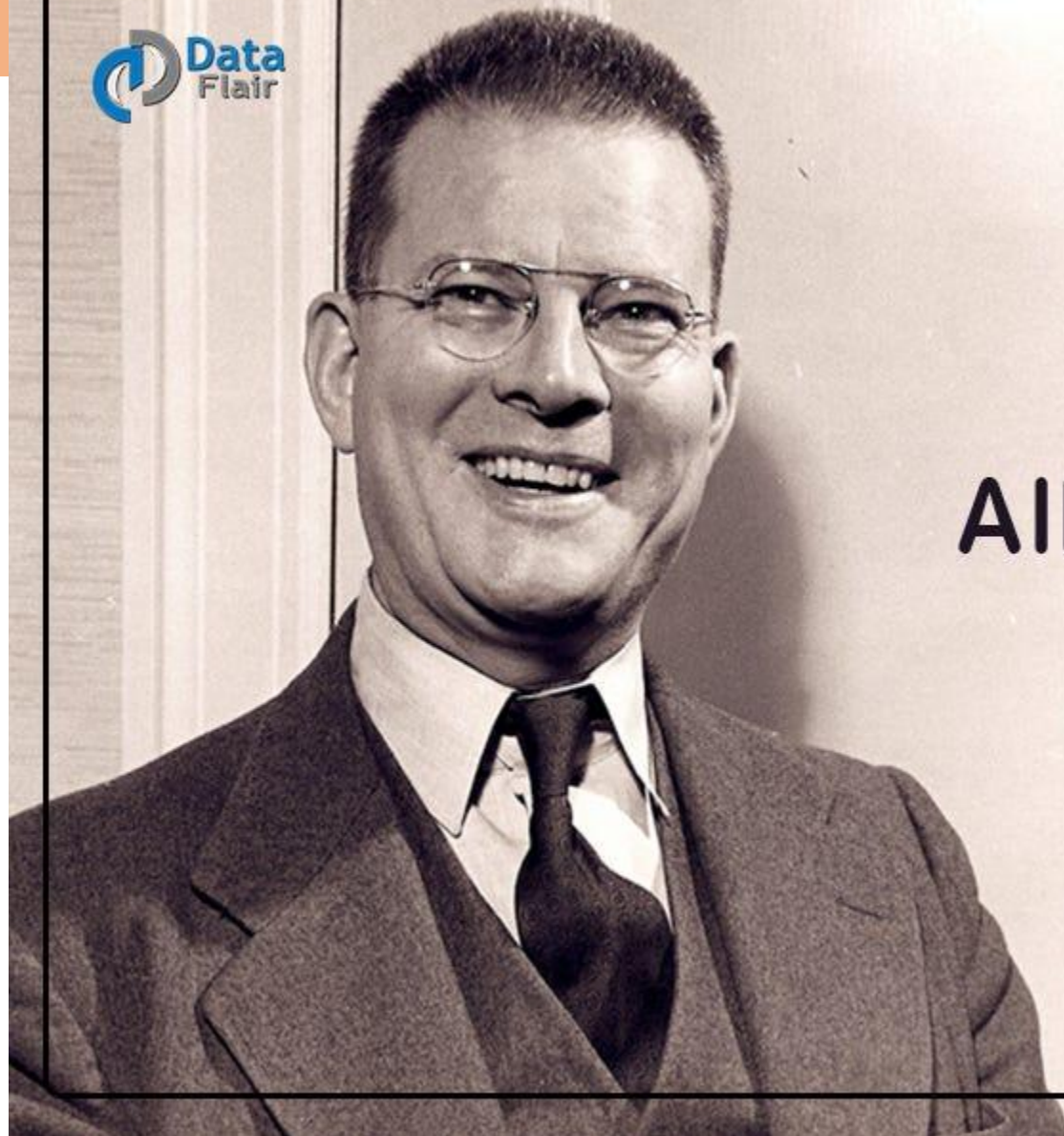
By the end of this session, participants will be able to:

1. Define Maximum Levels (MLs) and explain their relevance in Codex food safety standards.
2. Understand the importance of occurrence data and its role in developing science-based MLs.
3. Identify data distribution types and their implications for statistical analysis.
4. Apply appropriate statistical methods to analyze contaminant occurrence data, particularly non-parametric approaches.
5. Handle data challenges such as low sample sizes and left-censored data (below LOD/LOQ).
6. Evaluate and compare multiple datasets to assess suitability for combination or separate analysis.
7. Implement the five-step process for statistically robust ML development.

Agenda

1. Introduction to Maximum Levels (MLs)
2. Principles of Statistical Analysis of Occurrence Data
3. Understanding Data Distribution
4. Practical Exercise: 95th Percentile Estimation
5. Dealing with Data Limitations
6. Comparing and Combining Datasets
7. Statistical Tools and Final Recommendations





**In God we trust.
All others must bring data**

- W. Edwards Deming

**Statistician, Professor, Author,
Lecturer, and Consultant**

Introduction to ML Development

What is an ML?

- **Maximum Level (ML):** A regulatory upper limit for contaminants in food or feed.
- Based on **scientific data**, good practices, and risk assessment.
- **Occurrence data is the evidence base**
- Why is statistical analysis important?

To ensure MLs are protective, realistic, and reflect real-world contamination patterns.



Introduction to Statistical Analysis of Occurrence Data

Importance of Statistical Analysis

Statistical analysis is essential for handling occurrence datasets in the development of Maximum Levels (MLs) for contaminants in food and feed.

It ensures that MLs are based on robust, scientifically sound data reflecting real-world contaminant levels.

Steps in the Analytical Process

The process involves **data extraction, cleaning, distribution assessment, and presentation of results.**

These steps are crucial for preparing data for informed decision-making by Expert Working Groups (EWGs).



Introduction to Statistical Analysis of Occurrence Data

Global Applicability of MLs

- Proper analysis supports **global applicability of MLs** by accounting for data variability and distribution characteristics.
- This ensures that MLs are **relevant and effective** across diverse regions and conditions.

Role of Expert Working Groups

Results from statistical analysis are presented to Expert Working Groups (EWGs) for **informed decision-making**.

EWGs rely on **scientifically sound data** to establish Maximum Levels (MLs) for contaminants.



Why Statistics Save Lives ?

Core Message: 

Robust MLs = Consumer Protection + Fair Trade



Goal: Set science-based Maximum Levels (MLs) to **minimize health risks**

Foundation: “MLs should be set at a level which is (slightly) higher than the normal range of variation in levels in food and feed.” (*Codex GSCTFF Guidance*)



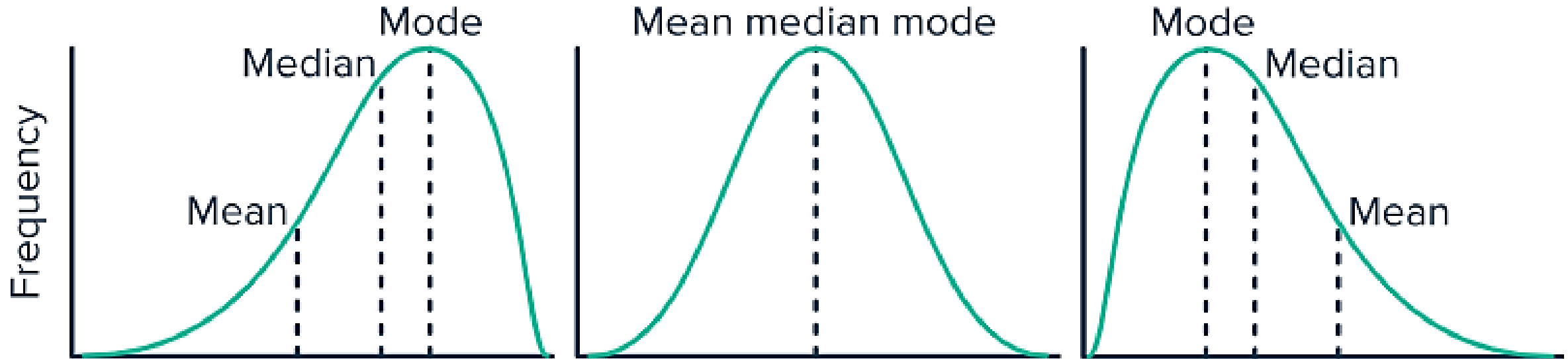
Example: If the selected value is 12.3 µg/kg, ML could be set at 13 or 15 µg/kg.

Data distribution assessment

Negatively skewed
(Left skewed)

Symmetrical distribution

Positively skewed
(Right skewed)



$\text{mean} < \text{median} < \text{mode}$

$\text{mean} = \text{median} = \text{mode}$

$\text{mean} > \text{median} > \text{mode}$

Which distribution to which scenario ?

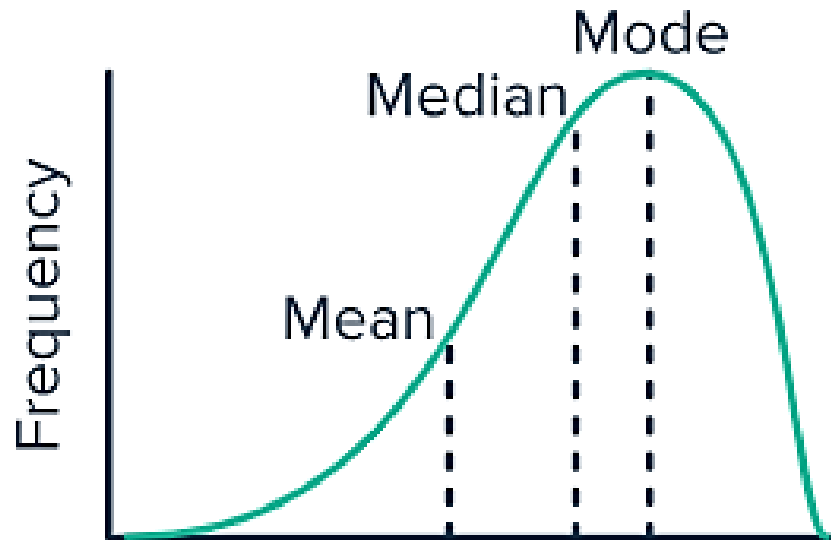
**Aflatoxin levels
in peanuts**

**Fiber levels in
Whole wheat bread**

**pH levels in
pasteurized milk**

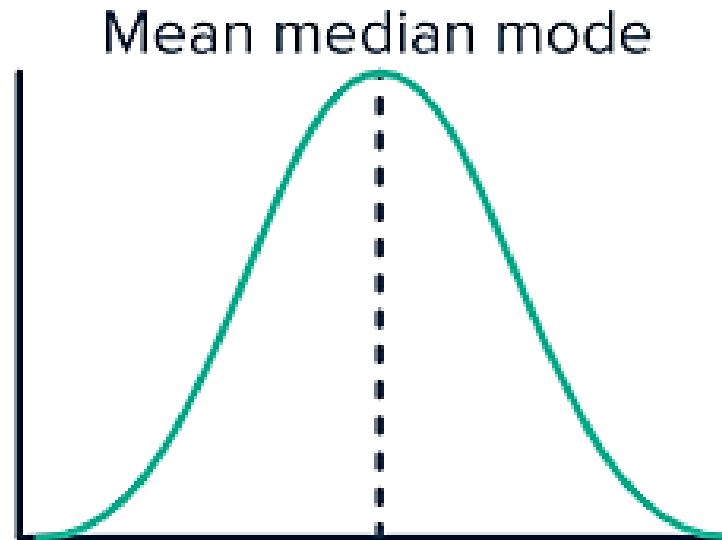
Data distribution assessment

Negatively skewed
(Left skewed)



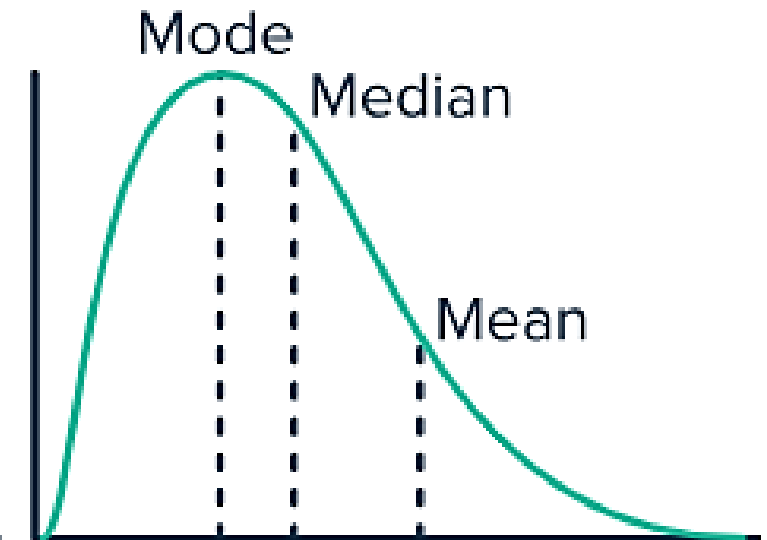
mean < median < mode

Symmetrical distribution



mean = median = mode

Positively skewed
(Right skewed)



mean > median > mode



**Fiber levels in
Whole wheat bread**

**pH levels in
pasteurized milk**

**Aflatoxin levels
in peanuts**

Data Distribution

Contaminant Data Distribution:

Often right-skewed , e.g., log-normal.

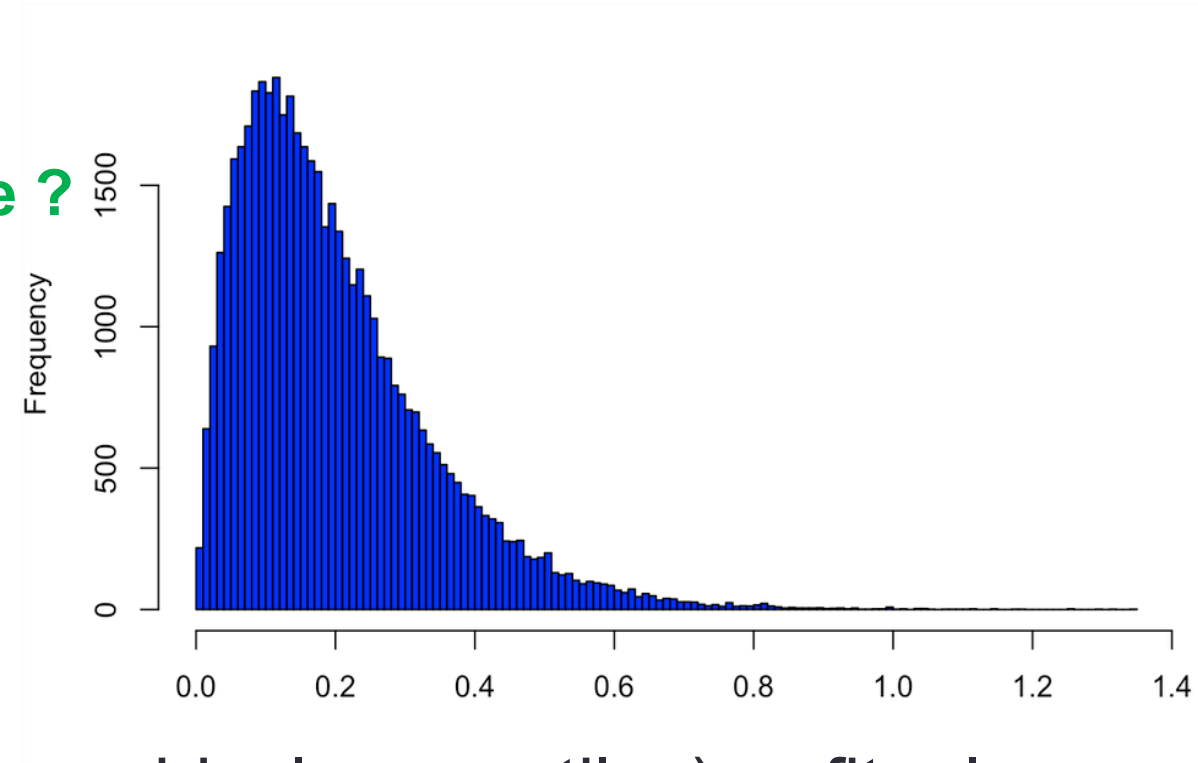
Why parametric methods are not suitable ?

Assume normal distribution.

May underestimate high percentiles.

Recommended Approach:

Use non-parametric methods (e.g., empirical percentiles) or fit a log-normal distribution .




Why not use the mean or median?

MLs must cover the upper end of contamination levels.

Consequences of Ignoring Skewness

Why Parametric Methods Fail ?

Example: Lead in Spices Dataset

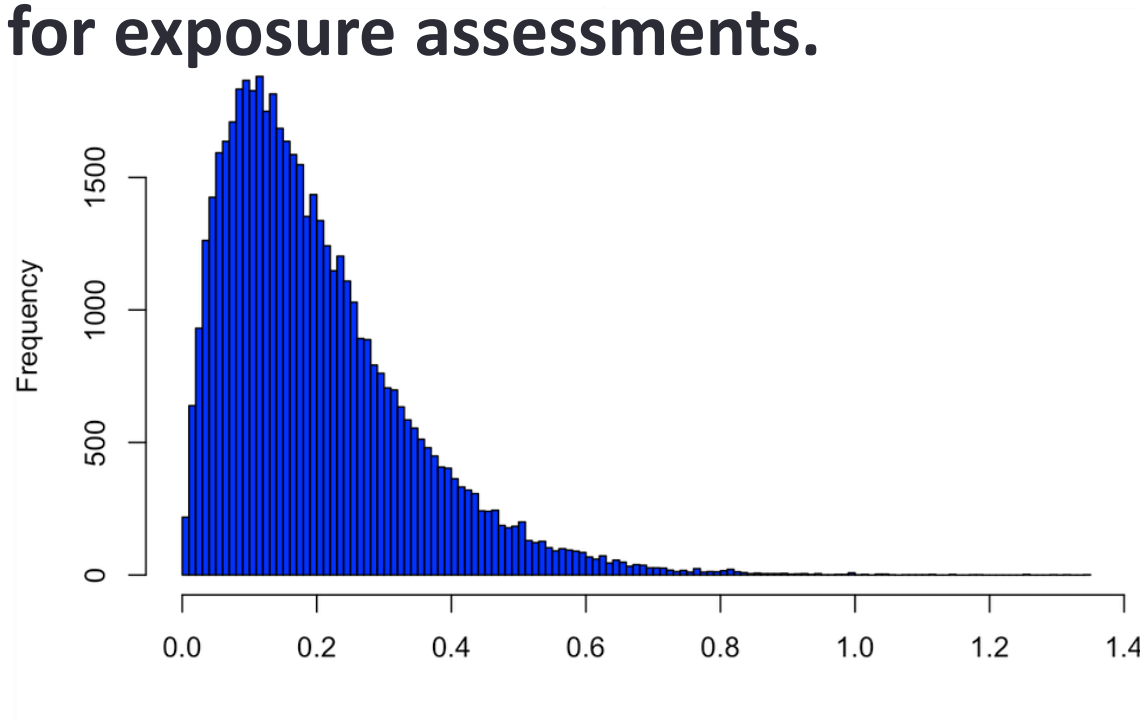
Method	Result (ppm)	Problem
Mean	0.8	Underestimates risk!
Mean + 2SD	2.1	Still misses extreme values
95th Percentile	5.3	Captures tail risk 

non-parametric methods = 95th Percentile Value

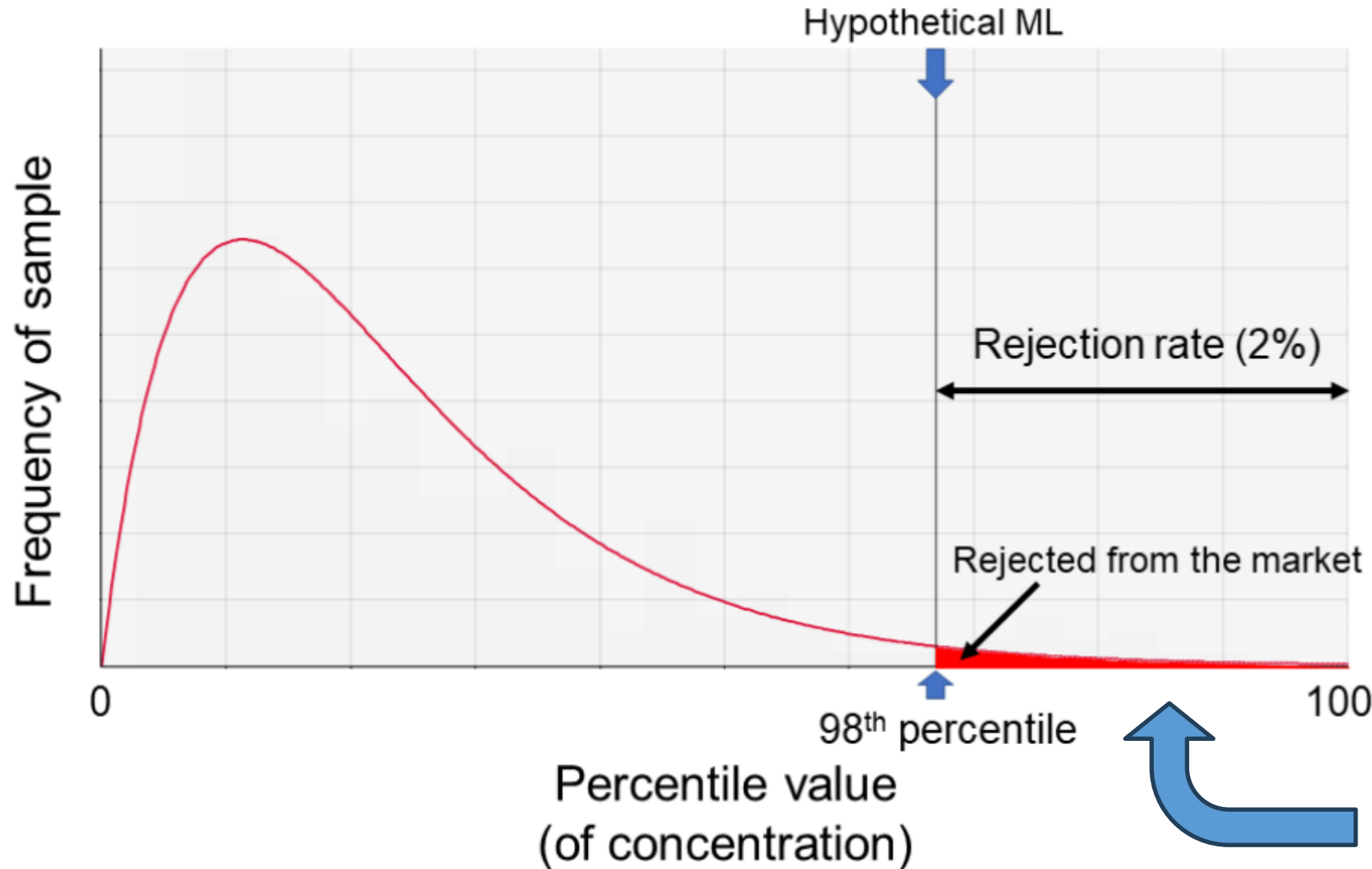
What is the 95th Percentile?

The 95th percentile is the value below which 95% of the data fall.

In food safety, it helps identify high-level consumers or high contamination levels—important for exposure assessments.



Handling Non-Normal Distributions



>90-95% of contaminant data are **RIGHT-SKEWED** (e.g., Mycotoxins, Heavy Metals, Marine Toxins)

✓ Most samples: **Low/undetectable contamination**

✗ Few samples: **Extreme contamination (the "TAIL")**

How to Calculate the 95th Percentile ?

Measuring Contaminant X in Matrix Y (in µg/kg)

We have a dataset of **20 Y samples**, and the results are:

[30, 22, 28, 35, 25, 27, 40, 33, 29, 31, 24, 36, 34, 38, 26, 32, 23, 37, 39, 41]



Step 1: Sort the Data in Ascending Order:

[22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41]



Step 2: Use the Percentile Position Formula

$$P = \frac{n + 1}{100} \times 95$$

$$P = \frac{20 + 1}{100} \times 95 = 0.21 \times 95 = 19.95$$

This means the **95th percentile lies between the 19th and 20th values** in the sorted list.

How to Calculate the 95th Percentile ?

From the sorted list: [22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41]

19th value = 40

20th value = 41

95th percentile = $40 + 0.95 \times (41 - 40) = 40.95$

So, the 95th percentile = 40.95 µg/kg

Round up to 41 or 45 µg/kg = ML

Key Notes:

- Percentiles help avoid **underestimating risks**.
- International bodies like **EFSA, WHO, and Codex** often rely on **95th percentile consumption or contamination levels** in risk assessments.
- It gives insight into **realistic worst-case scenarios**.

How to Calculate the 95th Percentile ?

Tip for Excel Users

In Excel, you can quickly calculate the 95th percentile using:

=PERCENTILE.EXC(A1:A20, 0.95)



What is the minimum number of data points to obtain percentiles values to generate MLs ?

Estimating High Percentile Values for ML Development

A **confidence level of 95%** is standard in food safety to ensure reliability of these estimates.

MLs are generally set slightly above the normal variation range, requiring estimation of high percentile values, typically the **95th percentile**.

A minimum of **59 data points** is necessary to estimate the 95th percentile with 95% confidence, ensuring statistical robustness.

Alternative calculation methods exist for **minimum data requirements**, but insufficient data can compromise ML accuracy.

A minimum data points

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= <u>0.95</u>)	<u>59</u>	74	99	119	149
[Option 1 (CL=0.99)]	[90]	[113]	[152]	[182]	[228]
Option 2 (CL: not provided*)	160	200	267	320	400

Handling datasets with low number of data points

- 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 would be **considered adequate** for developing an ML
- 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.
- **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.

Managing Left-Censored Data in Datasets

1. Definition of Left-Censored Data

- Left-censored data refer to non-quantified values below the limit of detection (LOD) or limit of quantification (LOQ).
- Such data arise from analytical limitations or low contaminant occurrence frequency.

2. Impact on Statistical Analysis

- Ignoring left-censored data biases statistical analysis.
- Substitution methods are employed to assign values for these data points.

3. Substitution Methods Overview

- Common substitution scenarios include lower-bound (zero or LOD), middle-bound (LOQ/2 or similar), and upper-bound (LOQ) approaches.
- These methods help mitigate the bias introduced by left-censored data.

Substitution Methods and Their Impact on Exposure Estimates

3 substitution scenarios

Scenario 1: Lower-bound (LB) scenario: results below the LOQ are replaced by zero, or by LOD if the LOD is known (results $< \text{LOD}$ are replaced by zero);

Scenario 2: Upper-bound (UB) scenario: results below the LOQ are replaced by the reported LOQ value; and

Scenario 3: The middle-bound (MB) scenario: A point estimate between the two extreme scenarios (LB and UB); assigning a value of $\text{LOQ}/2$, square root of the LOQ, or $(\text{LOD} + \text{LOQ})/2$ if the LOD is known for analytical results below the reported LOQ.

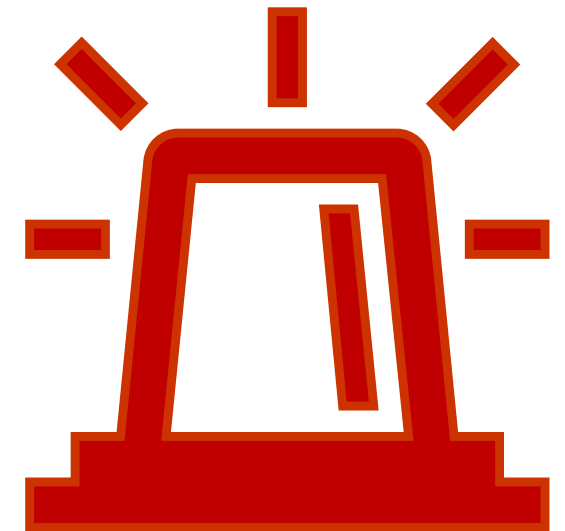
In general, $\text{LOQ}/2$ is the most widely used in dietary exposure assessments.

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

ATTENTION

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.



Combining and Comparing Multiple Datasets

Global Applicability of Codex MLs:

Codex MLs aim for **global applicability**, ideally based on combined global datasets. This ensures a broader and more inclusive understanding of patterns and trends.

Regional and Temporal Dataset Variations

Datasets may vary by region or year, necessitating **statistical comparison** before combining. Such comparisons ensure the integrity and reliability of the combined data.

Statistical and Visualization Tools

Non-parametric tests like **Mann-Whitney U** and **Kruskal-Wallis H** help assess distribution similarities. . The **null hypothesis** is that all datasets are assumed to be from the same population.

Aflatoxins in Peanuts

Nigeria
40 inputs

Argentina
40 inputs

DATASET FOR COMPAIRING.sav [DataSet1] - IBM SPSS Statistics Data Editor

	SampleID	Country	AflatoxinTotalppb	SamplingDate	SourceType	var	
1	NG-001	Nigeria	33.2	01-Mar-25	Open Market		
2	NG-002	Nigeria	8.6	02-Mar-25	Warehouse		
3	NG-003	Nigeria	14.5	03-Mar-25	Rural Farm		
4	NG-004	Nigeria	5.9	04-Mar-25	Open Market		
5	NG-005	Nigeria	18.2	05-Mar-25	Storage Unit		
6	NG-006	Nigeria	9.1	06-Mar-25	Retail Store		
7	NG-007	Nigeria	25.4	07-Mar-25	Warehouse		
8	NG-008	Nigeria	6.3	08-Mar-25	Rural Farm		
9	NG-009	Nigeria	11.8	09-Mar-25	Processing Unit		
10	NG-010	Nigeria	4.2	10-Mar-25	Open Market		
11	NG-011	Nigeria	15.3	11-Mar-25	Rural Farm		
12	NG-012	Nigeria	29.0	12-Mar-25	Storage Unit		
13	NG-013	Nigeria	10.0	13-Mar-25	Open Market		
14	NG-014	Nigeria	7.1	14-Mar-25	Rural Farm		
15	NG-015	Nigeria	12.6	15-Mar-25	Warehouse		
16	NG-016	Nigeria	3.8	16-Mar-25	Processing Unit		
17	NG-017	Nigeria	19.2	17-Mar-25	Storage Facility		
18	NG-018	Nigeria	6.5	18-Mar-25	Open Market		
19	NG-019	Nigeria	9.9	19-Mar-25	Retail Store		
20	NG-020	Nigeria	16.4	20-Mar-25	Rural Farm		
21	CN-001	Nigeria	22.7	01-Jan-25	Storage Center		
22	CN-002	Nigeria	8.4	02-Jan-25	Retail Store		
23	CN-003	Nigeria	10.6	03-Jan-25	Rural Farm		
24	CN-004	Nigeria	6.8	04-Jan-25	Wholesale Market		
25	CN-005	Nigeria	17.9	05-Jan-25	Open Market		
26	CN-006	Nigeria	4.5	06-Jan-25	Processing Plant		
27	CN-007	Nigeria	15.0	07-Jan-25	Warehouse		

Combined vs individual dataset

Histograms for occurrence data of iAs in husked rice

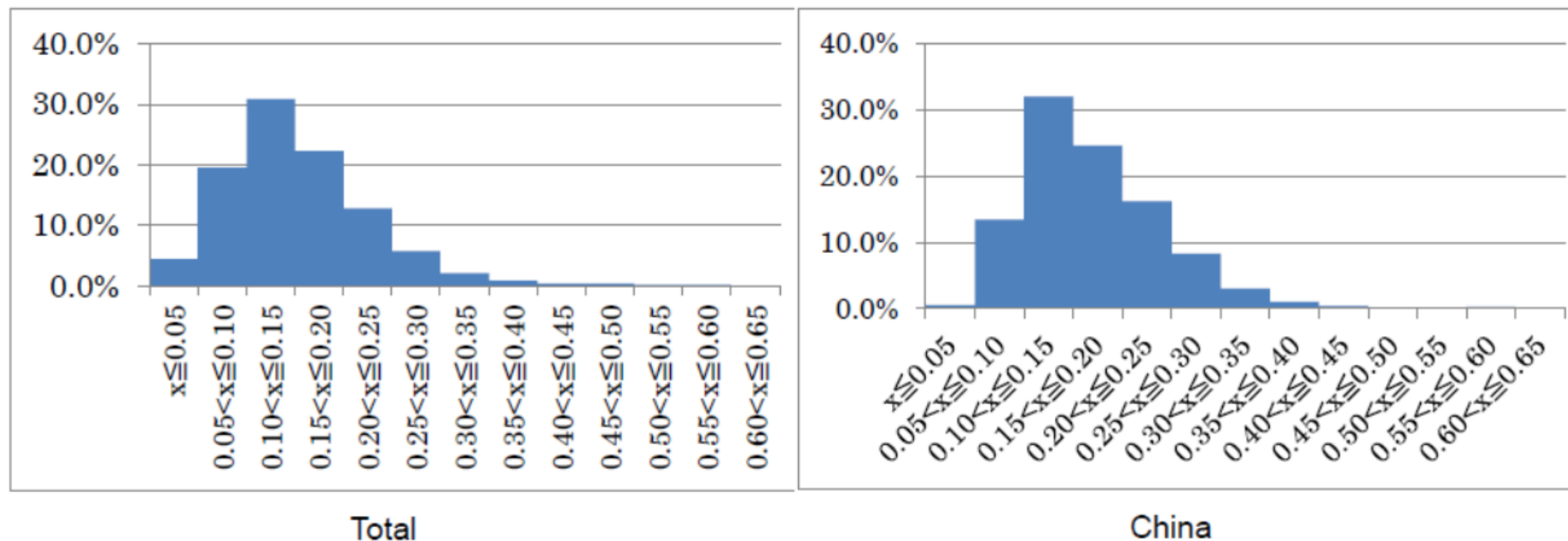


Figure 2. Example of histograms of occurrence data on inorganic As in husked rice (combined and individual dataset) (ref. CX/CF 15/9/7)

Cases where the analysis of individual datasets is recommended

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

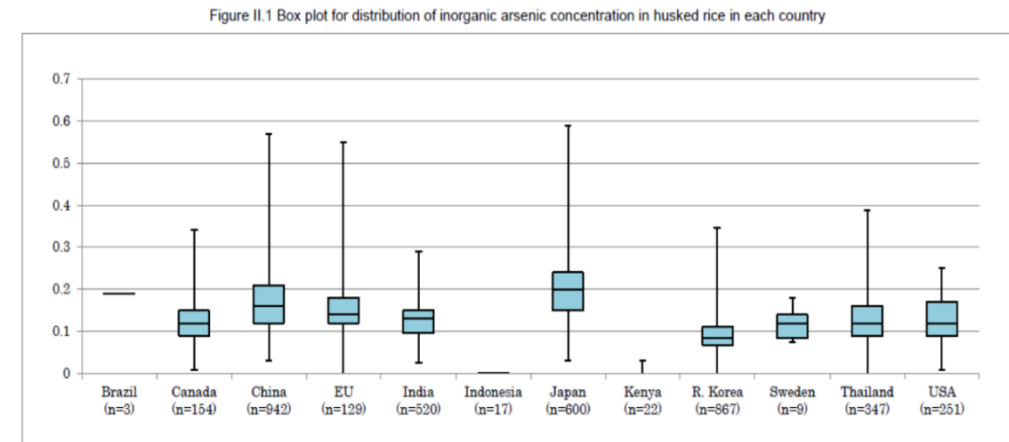


Figure 3. Box-and-whisker plots of individual datasets on inorganic As in husked rice (ref. CX/CF 16/10/5)

Statistical Analysis – 5 STEPS

Steps in Statistical Analysis:

1. **Clean** and **verify** the dataset.
2. Assess the **distribution** of the data.
3. Use **appropriate** statistical **methods**.
4. Estimate **high percentiles** (e.g., 95th or 98th).
5. **Round up** to a reasonable ML.

Take-home messages

Robust Dataset Development

Develop MLs using statistically robust datasets, prioritizing minimum data requirements and appropriate distribution assessments.

Dataset Comparability Evaluation

Use visualization and statistical tests to evaluate dataset comparability before combining or separating data sources. This ensures that the data sources are compatible and maintain the integrity of the analysis.

Case-by-Case Decision Making:

Make case-by-case decisions on ML development when data are limited. Balance consumer health protection and data quality in these scenarios.

Transparent Documentation Practices:

Transparently document methods, software, and assumptions to ensure reproducibility. This approach supports the global applicability of MLs.

Handling Skewed Data

Employ non-parametric methods and substitution techniques to handle skewed and left-censored data effectively. These techniques are crucial for maintaining the accuracy of the ML models.

STATISTICS: BECAUSE GUESSING IS NOT A FOOD SAFETY STRATEGY!

