SUMMARY STATISTICS TOOL BACKGROUND DOCUMENTATION

VITO

Introduction

The summary statistics tool calculates the summary statistics of the HBM data in a standardized and comparable way. This document provides an overview of which summary statistics are calculated and the methodology on how they are calculated.

The summary statistics are calculated for the biomarkers in the original units (volume-based) and standardized for creatinine and normalized for specific gravity for urinary markers and standardized for lipids in case of lipid soluble biomarkers measured in blood/breast milk. All calculations are performed per sample group (meaning that biomarker data in all urine types (urine spot, morning urine and urine 24 hours) is combined, biomarker data in all blood types (blood whole blood, blood serum, blood plasma) is combined etc.), per time point, per 'relation' variable (this is, per group of subjects, e.g., participants, mother of the participant, siblings, or father of the participant), and case vs. control groups. These groups are used in the summary statistics tool as default strata.

The summary statistics tool calculates descriptive statistics for each single biomarker, within each of the defined groups described above, imputed via single random imputation based on a lognormal distribution (see background documentation derived variables tool). For sum parameters, the summary statistics are obtained for the imputed variables via the medium bound imputation. Note that only values below LOD/LOQ are imputed; missing values, e.g., in case of a lost blood sample, are not imputed.

Statistics calculated

For each exposure biomarker, the tool calculates:

- Number of observations (sample size),
- Arithmetic mean with 95% confidence intervals (CIs),
- Standard deviation and the standard arithmetic mean error,
- Geometric mean with its 95% CIs,
- Observed percentiles (5th, 10th, 25th, 50th, 75th, 90th and 95th) and their 95% confidence intervals,
- Number and percentage of missing values,
- Number and percentage of values below LOD/LOQ.

Calculation details

Percentiles and their confidence intervals are only presented if observed, otherwise they are indicated as -1 (below LOD), -2 (between LOD and LOQ) or -3 (below LOQ). In the case of multiple LODs or LOQs for a specific biomarker within the data collection, the overall percentage of values below LOD or LOQ is used to present the final value of the percentile. As an example, if 25% of values are below LOD, independently of there being one or multiple LODs, then the 25th percentile will be below LOD. Same with between LOD and LOQ and below LOQ.

For single biomarkers, the calculation of arithmetic mean with 95% CI, standard deviation and the standard mean error, geometric mean with its 95% CIs, is only done if at least 60% of the values are detected. If <60% of values are detected, those summary statistic for that specific biomarker are not provided (empty).

Summary statistics are only provided if N is at least 50 (for GDPR reasons).

Accompanying the descriptive statistics calculated for the biomarker data, frequencies of other variables are added to the output, that is the frequency of the specific LOD/LOQ values and summary of sociodemographic characteristics will be calculated. All output variables are explained in **Annex Table 1: summary statistics codebook.**

Confidence intervals

The CIs for percentiles are computed using binomial distribution, taking the large sample into consideration (see (Conover, 1999)). The formulas and procedures used can be found https://www-users.york.ac.uk/~mb55/intro/cicent.htm. Missing values are not used when calculating percentiles or the confidence intervals.

Data stratifications

Besides the overall summary statistics for a biomarker within a data collection, further stratification is also possible based on specified characteristics (e.g., stratified for sex, etc.).
The tool calculates single stratifications based on a single variable and double stratifications based on a combination of 2 variables.

For PARC, the following default stratifications are integrated in the tool:
Single stratifications:

- Age of the subject (in years)
- Degree of urbanization: a subject's living environment is classified by Eurostat in three levels of urbanization[1]:
    - Densely populated area (cities)
    - Intermediate density area (towns or suburbs)
    - Thinly populated area (rural area)
- Educational level

  For stratification according to educational level, highest educational level of the household or, if not available, highest educational level of the mother, is used for stratification of children and teenagers. For adults, educational level of the subject is used. For educational level the International Standard Classification of Education (ISCED) developed by United Nations Educational, Scientific and Cultural Organization[2] is used. A subject's educational level is re-categorised. The tools includes 2 different educational level strata, educational level (3 levels: low, medium, high) and educational level (2 levels: low/medium, high).

  For educational level (**3 levels**) we categorize into:
    - Lower educational level denotes individuals with no to lower secondary education (ISCED 0-2);
    - Medium level of education includes individuals with upper secondary to post-secondary non-tertiary education (ISCED 3-4);
    - High level of education represents individuals with tertiary education and higher (ISCED ≥ 5)

  For educational level (**2 levels**) we categorize into:
    - Lower/medium educational level (ISCED 0-4);
    - Higher educational level (ISCED ≥5).
- NUTS 1
- NUTS 2
- NUTS 3
- Season of sampling

  Season of sampling, is defined as follows:
    - Spring: From March 21st to June 20th;
    - Summer: From June 21st to September 20th;

[1] https://ec.europa.eu/regional_policy/sources/work/2014_01_new_urban.pdf
[2]https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf

    - Autumn: From September 21st to December 20th;
    - Winter: From December 21st to March 20th.
- Sex of the subject
    - male vs. female participants
- Smoking habit of subject
    - Current smoker vs. current non-smoker
- Smoking of the mother during pregnancy
    - Mother non-smoker during pregnancy vs. mother smoker during pregnancy
- Year of sampling

Double stratifications

- Age of the subject x Degree of urbanization
- Age of the subject x Educational level
- Age of the subject x Season of sampling
- Age of the subject x Sex of the subject
- Age of the subject x Smoking habit of subject
- Age of the subject x Year of sampling
- Degree of urbanization x Educational level
- Degree of urbanization x Season of sampling
- Degree of urbanization x Sex of the subject
- Degree of urbanization x Smoking habit of subject
- Degree of urbanization x Smoking of the mother during pregnancy
- Degree of urbanization x Year of sampling
- Educational level x Season of sampling
- Educational level x Sex of the subject
- Educational level x Smoking habit of subject
- Educational level x Smoking of the mother during pregnancy
- Educational level x Year of sampling
- Season of sampling x Sex of the subject
- Season of sampling x Smoking habit of subject
- Season of sampling x Smoking of the mother during pregnancy
- Season of sampling x Year of sampling
- Sex of the subject x Smoking habit of subject
- Sex of the subject x Smoking of the mother during pregnancy
- Year of sampling x Sex of the subject
- Year of sampling x Smoking habit of subject
- Year of sampling x Smoking of the mother during pregnancy

References

UNESCO Institute for Statistics
International Standard Classification of Education ISCED 2011, vol. 88, UNESCO Institute for Statistics, Montreal, Quebec H3C 3J7 Canada (2012)

Annex 1: Summary statistics codebook.

Description of the different columns in the aggregated data sheet:	
Column header	**Description**
Version.script	Documentation of the tool version number.
Project	If the study is part of a project funded by the European Commission, then it is indicated here. Projects include now: • HBM4EU Aligned Studies • HBM4EU Occupational Studies • HBM4EU MOM Study • PARC Aligned Studies • DEMOCOPHES
Population.type	Indicates if the population sample in the row belongs to general population or to a subpopulation.
Subpopulation	In case the population belongs to a subpopulation, it is specified here to which one (Hotspot, pregnant women, occupationally exposed, Clinical, other)
Age.group	Age category to which the population sample in the row belongs to: Toddlers (0-2y) Children (3-11y) Teenagers (12-17y) Adults (18+) If a row cannot be categorized into one age group, it will be indicated as 'Multiple age groups'.
Country	Country Name
Region	European region of the study, countries are assigned to one of the following regions based on the United Nations Geoscheme: Northern Europe, Eastern Europe, Southern Europe, Western Europe. With an exception for Cyprus which is assigned to Southern Europe.
Data.collection	Name of the study
Repeated.sampling	In the case of repeated measurements, the time point is indicated here.
Subject	In the case of combined populations, this column indicates whether the data reported in the row are related to the participants or related subjects (mother, father, siblings).
Case or Control	In the case of case and control populations, this column indicates if the data in the row are related to the case or control subjects.
Matrix	Matrix name according to codebook (e.g., Urine, Blood)
Matrix.type	For urine and blood matrices, matrix type (For example, for urine, options Spot urine, 24h urine and First Morning urine are available).
Substance.group	Substance group
Biomarker	Biomarker name
Unit	Unit in which the concentration levels are expressed (e.g., µg/L, µg/g crt, . .))
Biomarker.abbreviation	Abbreviation of biomarker according to biomarker list
CAS.nr	CAS number
INCHI.key	INCHI key
CHEBI.key	CHEBI Key
Stratification	Shows whether the information in the corresponding row is non stratified data, single stratified data or double stratified data.
Stratification.name	For single and double stratifications, this column contains text describing for which variable(s) the stratification has been performed.
Stratification.value	For single and double stratifications, this column contains text describing for which strata the data are given. E.g., when stratification is done by sex, in "FEMALE", the data for the subgroup of females is provided.

N	Sample size for that row: excluding missing biomarker values but including observations below or within LOD-LOQ range.
Data.Filtered.on.N	If a data collection has a sample size of N<50 in the stratifications, then it is indicated for this stratum in this column as “N<50” and the descriptive statistics and frequencies are not calculated.
N.BelowLOD	Number of samples below LOD
N.BetweenLOD_LOQ	Number of samples between LOD and LOQ
N.BelowLOQ_LODunknown	Number of samples below LOQ
Perc.below.limit	Percentage of samples below LOD or LOQ. This can be used to know the percentage of values that were imputed and therefore to know the reliability of the imputation.
FREQ.LOD	LOD values frequency
FREQ.LOQ	LOQ values frequency

P05
P10
P25
P50
P75
P90
P95
	Observed percentiles: If the percentile value X lies below LOD or LOQ, it is substituted by the following values:
	if LOD as well as LOQ is provided:
	-1 for X < LOD
	-2 for LOD <= X < LOQ
	if LOQ is provided, but LOD is not:
	-3 for X < LOQ
	if LOD is provided, but LOQ is not:
	-1 for X < LOD

P05_95CI_Lower
P05_95CI_Upper
P10_95CI_Lower
P10_95CI_Upper
P25_95CI_Lower
P25_95CI_Upper
P50_95CI_Lower
P50_95CI_Upper
P75_95CI_Lower
P75_95CI_Upper
P90_95CI_Lower
P90_95CI_Upper
P95_95CI_Lower
P95_95CI_Upper
	Lower and upper confidence limits of the observed percentiles. If the value X lies below LOD or LOQ, it is substituted by the following values:
	if LOD as well as LOQ is provided:
	-1 for X < LOD
	-2 for LOD <= X < LOQ
	if LOQ is provided, but LOD is not:
	-3 for X < LOQ
	if LOD is provided, but LOQ is not:
	-1 for X < LOD

Mean SD SEM Mean_Lower95CI Mean_Upper95CI	Arithmetic mean, standard deviation (SD), standard error of the mean (SEM), and lower and upper limit of the 95%CI of the mean. Random imputation from a censored lognormal distribution is performed for values below LOD/LOQ . Missing when >60% of the biomarkers values are below LOD/LOQ. **For the sum parameters, the sum is calculated by substituting values below limit by medium bound imputation.
Geomean Geomean_Lower95CI Geomean_Upper95CI	Geometric mean, and lower and upper limit of the 95%CI of the geometric mean. Random imputation from a censored lognormal distribution is performed for values below LOD/LOQ. Missing when >60% of the biomarkers values are below LOD/LOQ. **For the sum parameters, the sum is calculated by substituting values below limit by medium bound imputation.

FREQ.Age of subject	Frequency tables of relevant variables in the data collection. To enable better interpretation, e.g., is the smoker / non-smoker ratio comparable across different analyses, check the parameter FREQ.smoking. The names that are used link to the variables of the PARC basic codebook.
FREQ.Degree of urbanization	
FREQ.Educational level of the subject (three classes)	
FREQ.Educational level of the father (three classes)	
FREQ.Highest educational level of the household (three classes)	
FREQ.Educational level of the mother (three classes)	
FREQ.NUTS1	
FREQ.NUTS2	
FREQ.NUTS3	
FREQ.Season of sampling	
FREQ.Year of sampling	
FREQ.Sex of the subject	
FREQ.Smoking habit of subject	
FREQ.Smoking of the mother during pregnancy	
Metadata.IPCHEM	Link to the metadata page in IPCHEM corresponding to each data collection.
Lab.institute	Institution name (acronym) of the laboratory that performed the chemical analysis.
Lab.group	Name of the group within the institution/laboratory that performed the chemical analysis (if applicable).
Lab.country	Country of the laboratory that performed the chemical analysis.
Accredited.method	Information about successful participation in proficiency tests (PTs) by the time of analysis.
Analytical.method	Analytical method used.