Overview of probabilistic estimation of exposure

DAY 3 (25th January)
9:30 – 11:00

Objectives

- To understand how to estimate dietary exposures by probabilistic estimation

Outline of the presentation

- Probabilistic models
- Principles of the Monte Carlo simulation
- Fitting occurrence data to an appropriate distribution model
- Fitting occurrence data for estimation of maximum level
- Estimation of dietary exposure

PROBABILISTIC MODELS

What is probabilistic estimation?

- To estimate distribution of dietary intake from both data of food consumption and its contamination level taking into account those variability and provability
- Provides more detailed information and realistic results on dietary intake of population of interest than deterministic approach such as a point estimation

Characteristics of probabilistic estimation

- Can reflect individual variability in dietary pattern, amount of consumption and body weight by use actual data
- Can simulate various scenarios such as whether an ML is established or not
- Can provide a probability distribution and high percentile exposure
- Require many data, time and cost
When we use probabilistic estimation?
- Available adequate data on both contaminants concentration in foods and food consumption
- Evaluation of risk management measures such as maximum levels or codes of practice
- If point estimates indicate no health concern for all populations, probabilistic estimation are not necessarily required

What is needed for probabilistic estimation?
- Many data points for both food consumption and contaminant concentrations
  - Statistically designed survey
  - Quality controlled / quality assured data
  - The larger data points the higher reliability
- Computer and probabilistic modeling software for simulation (e.g. @RISK, Crystal Ball, MCRA, Analytica, etc.)
- Understanding basic statistics

Probabilistic distributions
- Two approaches are used in probabilistic estimation of exposure
  - Parametric approach: Use probabilistic distribution model by fitting to a dataset
    - In general, lognormal, gamma, inverse Gaussian, exponential or Pearson type V or IV are selected as distribution models
  - Non-parametric approach: Use actual distribution of data set as is

Examples of probabilistic model
Normal distribution
RiskNormal(μ,σ)* specifies a normal distribution with parameters mean μ and standard deviation σ
- Domain: -∞ < x < +∞

Examples of probabilistic model
Lognormal distribution
RiskLognorm(μ,σ) specifies a lognormal distribution with parameters mean μ and standard deviation σ
- Domain: 0 ≤ x < +∞

Examples of probabilistic model
Gamma distribution
RiskGamma(α, β) specifies a gamma distribution with shape parameter α and scale parameter β
- Domain: 0 < x < +∞
Examples of probabilistic model

**Inverse Gaussian distribution**

RiskInvgauss(μ, λ) specifies an inverse Gaussian distribution with mean μ and shape parameter λ.

**Pearson type V distribution**

RiskPearson5(α, β) specifies a Pearson type V distribution with shape parameter α and scale parameter β.

**Pearson type VI distribution**

RiskPearson6(α₁, α₂, β) specifies a Pearson type VI distribution with shape parameter α₁ and α₂, and scale parameter β.

**Exponential distribution**

RiskExpon(β) specifies an exponential distribution with parameter β, the mean of the distribution.

Examples of distribution model

**Examples of probabilistic model**

**Exponential distribution**

RiskExpon(β) specifies an exponential distribution with parameter β, the mean of the distribution.

**Advantages of probabilistic models**
- Interpolate among the data points
- Extrapolate beyond the data points
- Provide continuous distribution form
- Represent variable as a distribution function

**Situation to use parametric approach**
- Interpolation and extrapolation by assuming a distribution form are preferable in order to fill in gaps of uncertainty of actual data set.
- Modeling of distribution form is necessary to estimate a value of high percentile of the distribution from a limited data point.

Recommended for distributions of concentration data of contaminants in foods in probabilistic estimation of dietary intake.
**Situation to use non-parametric approach**

- A data set does not fit simple distributions (e.g. multimodal)
- A data set does not have a sufficient data point can be assumed to represent the distribution of interest

Recommended for food consumption data in order to avoid taking random unrealistic values that would never occur in real life into account (in general, upper end of model is infinity)

---

**PRINCIPLES OF THE MONTE CARLO SIMULATION**

---

**What is Monte Carlo simulation?**

- A computerized mathematical technique that allows people to account for risk in quantitative analysis
- Used to model the probability of different outcomes in a process that cannot easily be estimated due to the intervention of random variables
- Named after Monte Carlo, the Monaco resort town renowned for its casinos

---

**How Monte Carlo simulation works?**

- Samples a random value from each input distribution of variables and runs the model using those values
- Calculates results over and over, each time using a different set of random values from the probabilistic distribution model
- After iterating the process a number of times, estimates probability distributions for the outputs of the model
- The more accurate estimations require more iterations

---

**Monte Carlo simulation**

- Select consumption from distribution at random
- Select concentration from distribution at random
- Calculate dietary intake

- A sufficient number of iterations (> 10000)

---

**FITTING OCCURRENCE DATA TO AN APPROPRIATE DISTRIBUTION MODEL**
Preparing occurrence data for probabilistic estimation  
(Review of the lectures of day1 and day2)
- If several data sets are gathered, no significant difference between data sets should be checked before data aggregation.
- If a histogram of data set indicates multimodality, inappropriate data (including outlier) might be incorporated in the data set and should be excluded when possible.
- Data from evidently contaminated situations shall be excluded from the data set.

Handling of left-censored data  
(Review of the lecture of day2)
- If the data set has values falling below analytical LOD or LOQ, it is recommended to prepare following two data set:
  - lower bound (e.g. \( x < \text{LOQ} = 0 \))
  - upper bound (e.g. \( x < \text{LOD} = \text{LOD}, \text{LOD} \leq x < \text{LOQ} = \text{LOQ} \))
- If the data set includes a significant number of data of \(< \text{LOQ}\), more sensitive analytical methods for collecting occurrence data might be required for use to exposure assessment depending on toxicity.

Fitting distribution to occurrence data
- Parametric approach
- @RISK allows to fit probability distributions in MS Excel for use statistically modeled concentration.

Fitting distributions to the data set by use of @RISK
- The fitted distributions can be assigned to an uncertain input in the spreadsheet.
- The fitted distribution can be linked to the data, so that the fit will automatically update whenever sample data change.
- Several options are available for controlling the fitting process in @RISK.

Fitting Options
- Specific distributions can be selected to fit (@RISK automatically selects applicable distributions).
- As negative values are impossible for occurrence data, “Lower Limit” could be set 0 as fixed bound.

Evaluation of fitting reports
- The following graphs, statistics are available:
  - Statistics on both the fitted distribution and the input data.
  - Graph comparison, P-P, and Q-Q plots.
  - The goodness-of-fit (GOF) tests on the fit, including the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the chi-square statistic (Chi-Sq), the Kolmogorov-Smirnov statistic (K-S), the Anderson-Darling statistic (A-D).
Example of fitting report in @RISK

- Graph comparison
- Fit Ranking
- Delimiters
- Probabilities
- Statistics

Comparison of the two graphs
- A “Comparison Graph” displays two curves: the fitting distribution and the distribution of the sample data
- Two delimiters are available for a comparison graph
- These delimiters set the Left X and Left P values, along with the Right X and Right P values. Values returned by the delimiters are displayed in the probability bar above the graph

P-P (Probability-Probability) plots
(Review of the lecture of day2)
- Plots the p-value of the fitted distribution vs the p-value of the input data
- If the fit is good, the plot will be nearly linear
- Emphasis on fitting in high probability values

Q-Q (Quantile-Quantile) plots
(Review of the lecture of day2)
- Plots the quantiles of the fitted distribution vs the quantiles of the input data
- If the fit is good, the plot will be nearly linear
- Emphasis on fitting in high quantile values

Fit ranking
- Ranks the fitted distributions according to the GOF tests
- In general, a lower statistic indicates a better fit
- The Fit Ranking selector specifies the GOF statistic to use for ranking
- Indicates how well a potential fitted distribution matches the distribution of the sample data
- The following statistics are available in @RISK: AIC/BIC, Chi-sq, K-S and A-D

Selection of the best distribution model
- Taking into account evaluation for all graphs, statistics, and reports, select the best choice for models
- Risk managers or risk assessors have a responsibility on the selection of models
- Selected models affect results of Monte Carlo simulation
FITTING OCCURRENCE DATA FOR ESTIMATION OF MAXIMUM LEVEL

Establishment maximum levels for contaminants in foods (Review of the lectures of day1 and day2)
- Maximum levels should be set based on occurrence data follow ALARA principle
- Parametric approach for assuming distribution of occurrence data can be used for estimation proposed MLs
- Probabilistic modelling software such as @RISK can derive high percentile values from distribution models

Maximum levels should be set based on occurrence data follow ALARA principle
- Parametric approach for assuming distribution of occurrence data can be used for estimation proposed MLs
- Probabilistic modelling software such as @RISK can derive high percentile values from distribution models

Percentile values from parametric or non-parametric distribution (Review of the lecture of day2)
- Non-parametric distribution can not assume values beyond the range of actual data set
  - Percentile values shall be assigned lower than the max value of the actual data
- Parametric distribution can extrapolate data beyond the range of the actual data set depending on selected models
  - High percentile values could result in higher than the max value of the actual data

Example of distribution of occurrence data of a certain contaminant in food

Estimation of dietary exposure
- Target population groups assessed (i.e. whole population, young children, women childbearing age, etc.)
- Foods contributing exposure (single source or multiple sources)
- Degradation of contaminants during food processing
- Duration of exposure (acute or chronic)
- Cut off limit of contaminant concentration (existence or non-existence of maximum levels)
Distribution model of occurrence data

- Select a distribution model of contaminants concentration by fitting evaluation
- Examples of distribution function in @RISK are as follows:
  - RiskLognormal(a,b) (returns a lognormal distribution)
  - RiskInvgauss(a,b) (returns an inverse Gaussian distribution)

Note: functional forms are defined by selected distribution models

Cut off limit of concentration data

- Because sampling is at random, the simulation will be in accurate at the extreme ends of distribution if using parametric distribution
- A cut-off limit in the distribution tail correspond to maximum level is available to avoid taking unrealistic samples into account

Distribution of food intake

- Use actual consumption weights of foods of the interest and body weight of individuals

Note: Types of food intake distribution are different between single day consumption and average consumption of multiple days

Run Monte Carlo simulation

- After preparing the distribution models of variables, develop an equation to calculate the intake based on the exposure scenario
- Be careful about unit of intake (µg/kg bw/day, /week or /month) (mg=1,000, µg=1,000,000 ng)
- Run Monte Carlo simulation in a computer using a probabilistic modelling software
- If necessary, consider a different exposure scenario for simulation

Evaluation of health risk based on probabilistic estimate

- Determine the mean and the 95th percentile value of dietary intake (and the 99th percentile value for acute toxic substance)
- Compare those values with health based guidance values (PTDI, PTWI, PTMI, BMDL, ARfD, etc.)
- Evaluation the impact to the dietary intake of proposed maximum levels
- Conduct uncertainty analysis

Refer to the lectures of day1 and day2