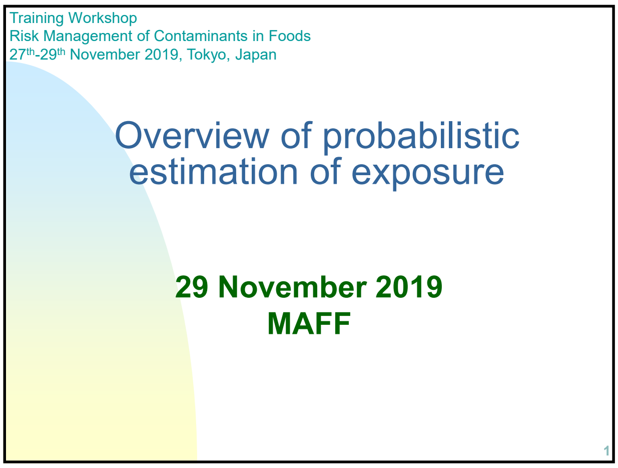
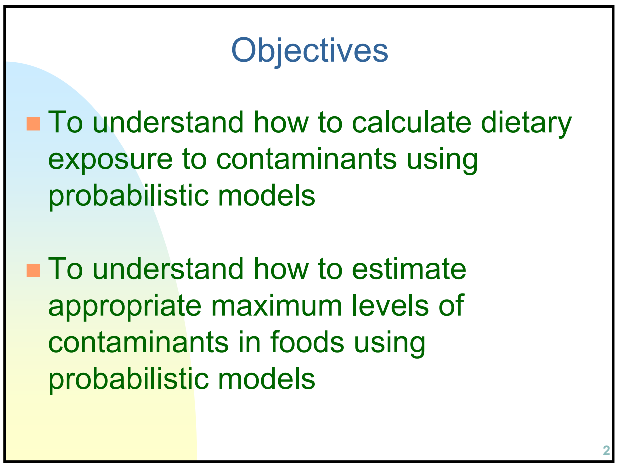


Training Workshop
Risk Management of Contaminants in Foods
27th-29th November 2019, Tokyo, Japan

Overview of probabilistic estimation of exposure

29 November 2019
MAFF

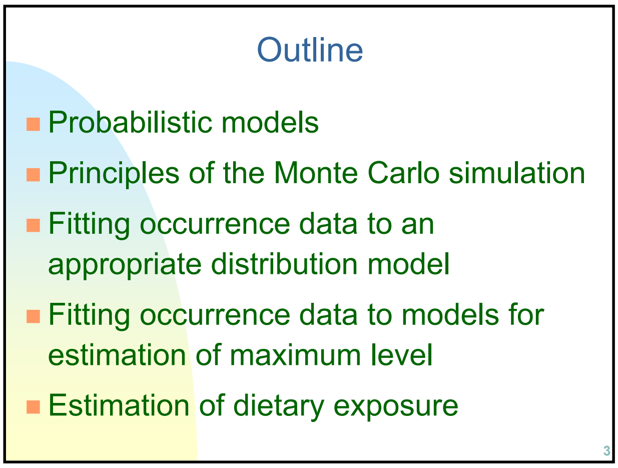
1



Objectives

- To understand how to calculate dietary exposure to contaminants using probabilistic models
- To understand how to estimate appropriate maximum levels of contaminants in foods using probabilistic models

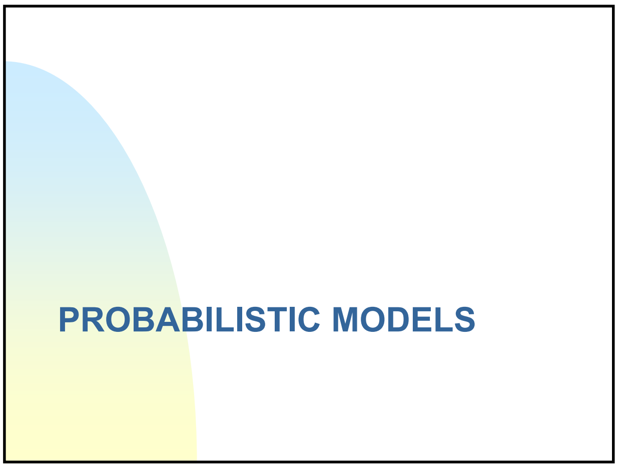
2



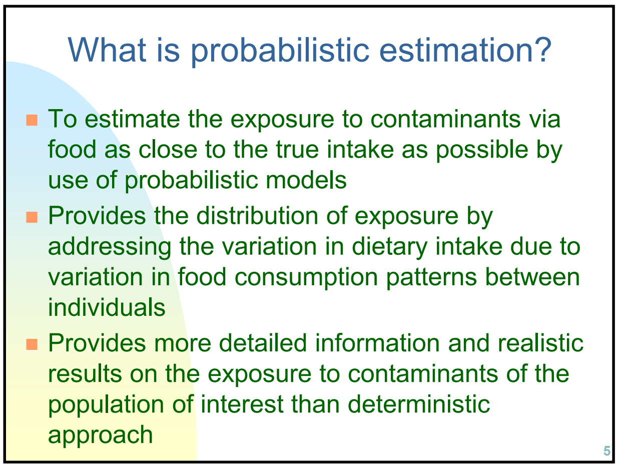
Outline

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

3



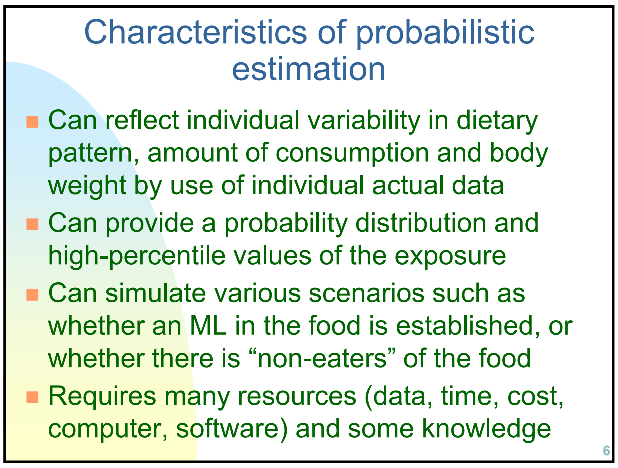
PROBABILISTIC MODELS



What is probabilistic estimation?

- To estimate the exposure to contaminants via food as close to the true intake as possible by use of probabilistic models
- Provides the distribution of exposure by addressing the variation in dietary intake due to variation in food consumption patterns between individuals
- Provides more detailed information and realistic results on the exposure to contaminants of the population of interest than deterministic approach

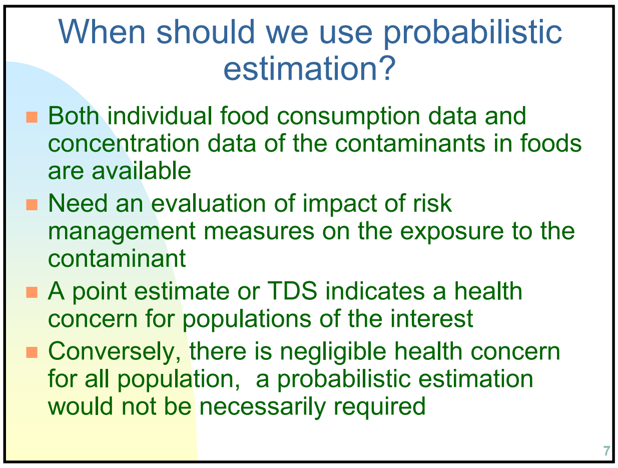
5



Characteristics of probabilistic estimation

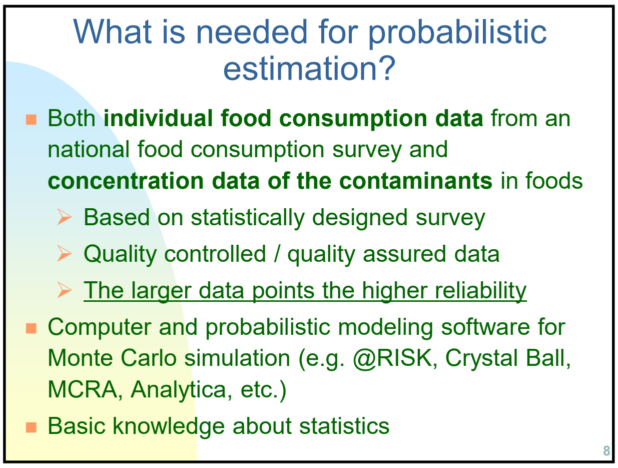
- Can reflect individual variability in dietary pattern, amount of consumption and body weight by use of individual actual data
- Can provide a probability distribution and high-percentile values of the exposure
- Can simulate various scenarios such as whether an ML in the food is established, or whether there is “non-eaters” of the food
- Requires many resources (data, time, cost, computer, software) and some knowledge

6



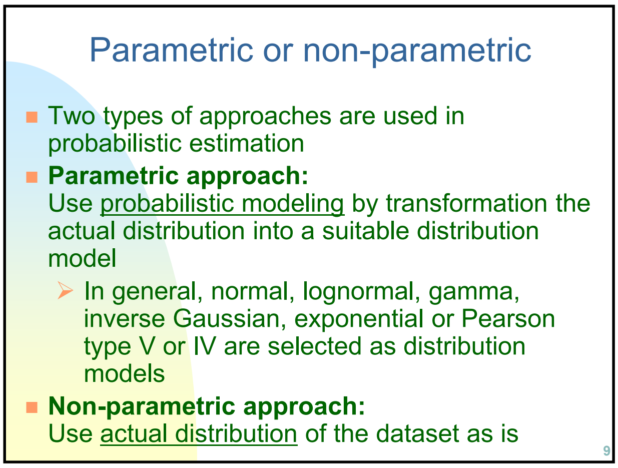
When should we use probabilistic estimation?

- Both individual food consumption data and concentration data of the contaminants in foods are available
- Need an evaluation of impact of risk management measures on the exposure to the contaminant
- A point estimate or TDS indicates a health concern for populations of the interest
- Conversely, there is negligible health concern for all population, a probabilistic estimation would not be necessarily required



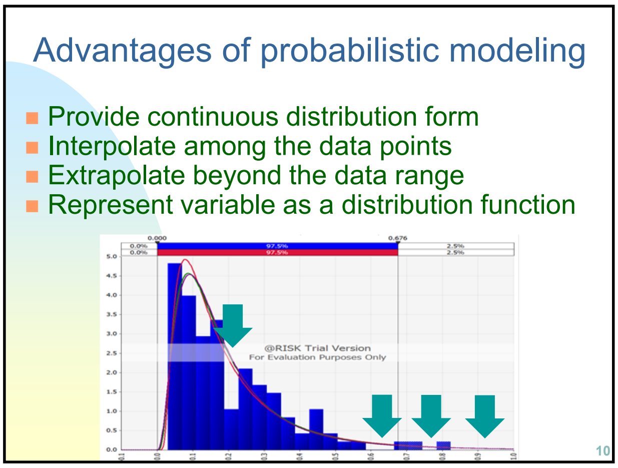
What is needed for probabilistic estimation?

- Both **individual food consumption data** from an national food consumption survey and **concentration data of the contaminants** in foods
  - Based on statistically designed survey
  - Quality controlled / quality assured data
  - The larger data points the higher reliability
- Computer and probabilistic modeling software for Monte Carlo simulation (e.g. @RISK, Crystal Ball, MCRA, Analytica, etc.)
- Basic knowledge about statistics



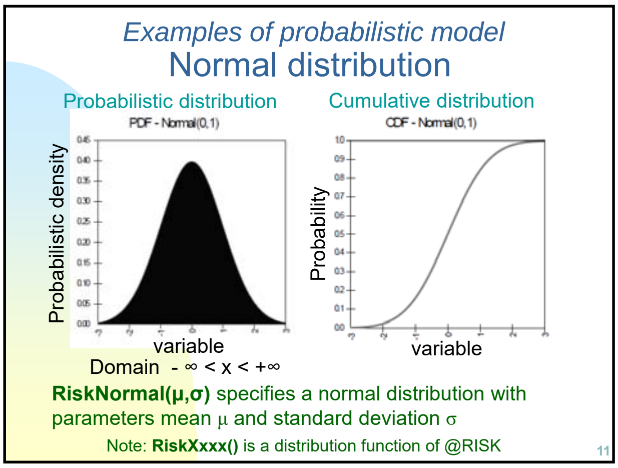
Parametric or non-parametric

- Two types of approaches are used in probabilistic estimation
- **Parametric approach:**
  Use probabilistic modeling by transformation the actual distribution into a suitable distribution model
  - In general, normal, lognormal, gamma, inverse Gaussian, exponential or Pearson type V or IV are selected as distribution models
- **Non-parametric approach:**
  Use actual distribution of the dataset as is



Advantages of probabilistic modeling

- Provide continuous distribution form
- Interpolate among the data points
- Extrapolate beyond the data range
- Represent variable as a distribution function

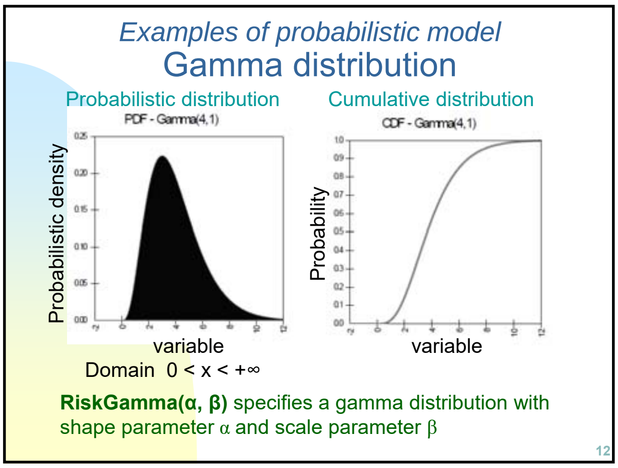


Examples of probabilistic model
Normal distribution

Domain - ∞ < x < +∞

RiskNormal(μ,σ) specifies a normal distribution with parameters mean μ and standard deviation σ

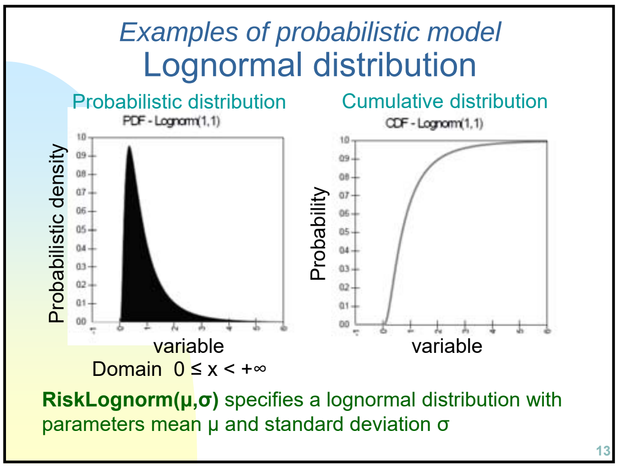
Note: **RiskXxxx()** is a distribution function of @RISK



Examples of probabilistic model
Gamma distribution

Domain 0 < x < +∞

RiskGamma(α, β) specifies a gamma distribution with shape parameter α and scale parameter β



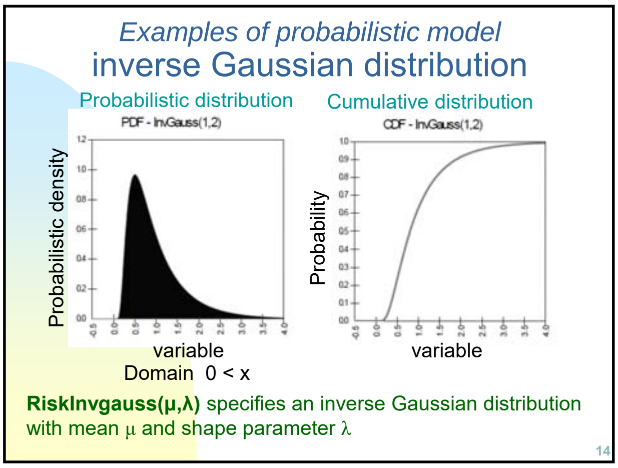
Examples of probabilistic model

Lognormal distribution

Domain 0 ≤ x < +∞

RiskLognorm(μ,σ) specifies a lognormal distribution with parameters mean μ and standard deviation σ

13



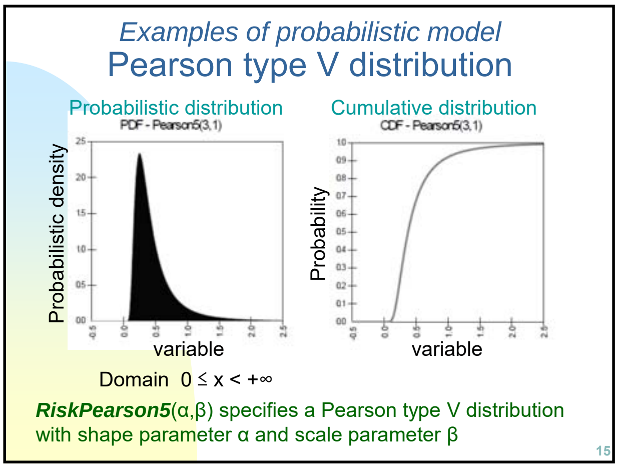
Examples of probabilistic model

inverse Gaussian distribution

Domain 0 < x

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

14



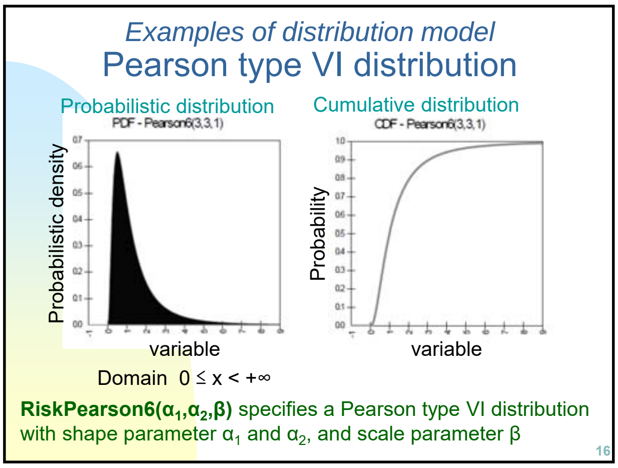
Examples of probabilistic model

Pearson type V distribution

Domain 0 ≤ x < +∞

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

15



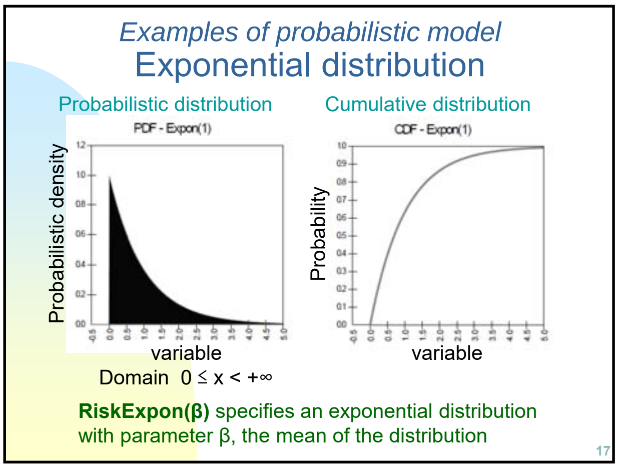
Examples of distribution model

Pearson type VI distribution

Domain 0 ≤ x < +∞

RiskPearson6(α_1,α_2,β) specifies a Pearson type VI distribution with shape parameter α_1 and α_2, and scale parameter β

16



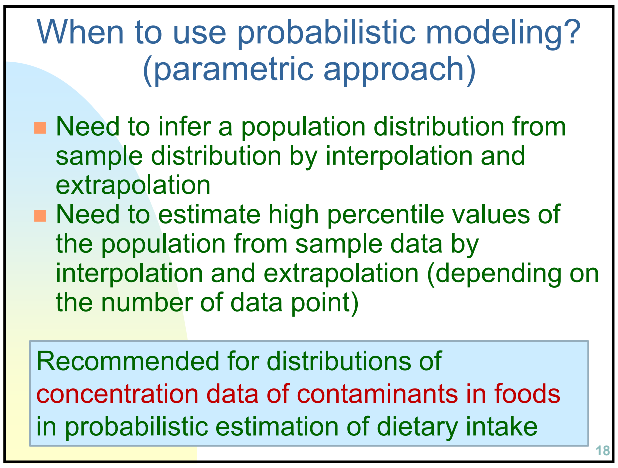
Examples of probabilistic model

Exponential distribution

Domain 0 ≤ x < +∞

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

17



When to use probabilistic modeling? (parametric approach)

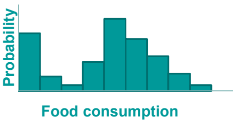
- Need to infer a population distribution from sample distribution by interpolation and extrapolation
- Need to estimate high percentile values of the population from sample data by interpolation and extrapolation (depending on the number of data point)

Recommended for distributions of concentration data of contaminants in foods in probabilistic estimation of dietary intake

18

When to use sample distribution? (non-parametric approach)

- A sample data does not fit simple distributions (e.g. multimodal)
- A sample size is not sufficient for inference the population of interest



Recommended for food consumption data in order to avoid using unrealistic values that would never occur in real (In general, upper end of probabilistic models are infinity)

PRINCIPLES OF THE MONTE CARLO SIMULATION

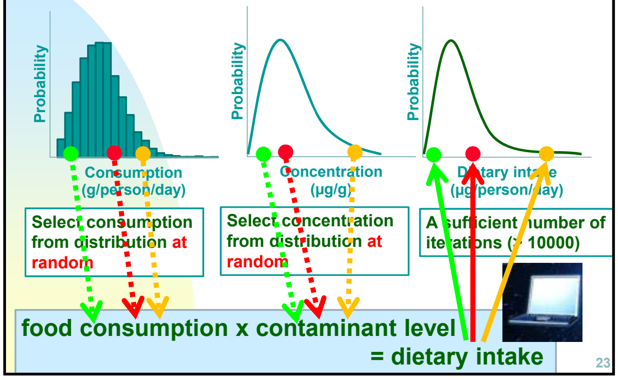
What is Monte Carlo simulation?

- 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 simulation 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

A sufficient number of iterations (> 10000)

food consumption x contaminant level = dietary intake

FITTING OCCURRENCE DATA TO AN APPROPRIATE DISTRIBUTION MODEL

Preparing occurrence data for probabilistic estimation

(Review of the lectures of day1 and day2)

- When there are multiple sample dataset, aggregate after confirming that there is no significant difference in distribution
- If a histogram of the dataset indicates multimodality, inappropriate data might be incorporated in the dataset, and should be excluded before further analysis, where possible
- Data obtain from evidently contaminated situations which can not be used as foods shall be excluded from the dataset.

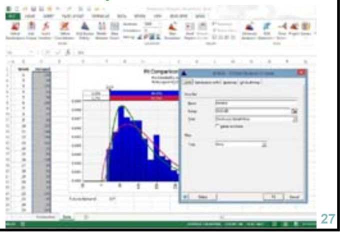
Handling of below LOQ data

(Review of the lecture of day2)

- If the data set has values falling below analytical LOD or LOQ, it is recommended to prepare following datasets:
  - Lower bound (e.g. “x < LOQ” = 0)
  - Upper bound (e.g. “x < LOD”= value of LOD and “LOD ≤ x < LOQ”= value of LOQ , **or** “x <LOQ” = value of LOQ)
- If the data set includes a significant number of data of below LOQ, more sensitive analytical methods for collecting occurrence data might be required for exposure assessment depending on a toxicity of the contaminant

Fitting occurrence data to a distribution model

- Parametric approach
- Need a modeling software
- @RISK allows to fit probability distributions in MS Excel for use statistically modeled concentration

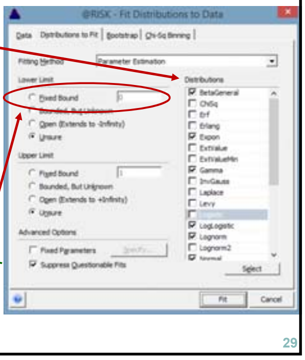


Fitting distribution of a dataset to models by use of @RISK

- The fitted distributions can be assigned to an uncertain input in the spreadsheet of MS Excel
- The fitted distribution can be linked to the data, so that the fit will automatically update whenever sample data are changed on the spreadsheet
- Several options are available for controlling the fitting process in @RISK

Fitting Options of @RISK

- Specific distributions can be selected to fit (@RISK automatically selects applicable distributions in default condition)
- As negative values are impossible for occurrence data, “Lower Limit” could be set to 0 or other positive values as “Fixed Bound”



Fitting Method: Parameter Estimation

Lower Limit
○ Fixed Bound 0
○ Bounded, But Unknown
○ Open (Extends to -Infinity)
● Unsure

Upper Limit
○ Fixed Bound 1
○ Bounded, But Unknown
○ Open (Extends to +Infinity)
● Unsure

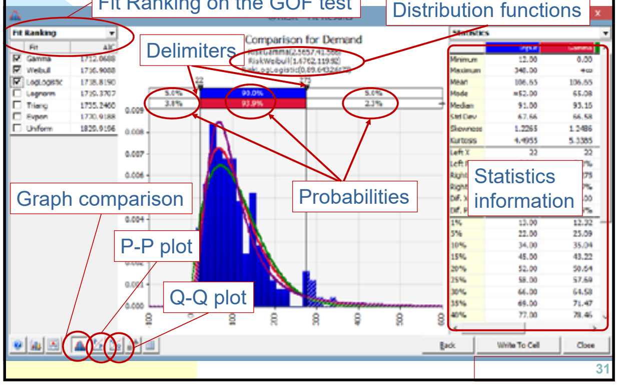
Advanced Options
☐ Fixed Parameters
☑ Suppress Questionable Fits

Distributions
☑ BetaGeneral
☐ ChiSq
☐ Erf
☐ Erlang
☑ Expon
☐ ExtValue
☐ ExtValueMin
☑ Gamma
☐ InvGauss
☐ Laplace
☐ Levy
☐ [illegible]
☑ LogLogistic
☑ Lognorm
☐ Lognorm2
☑ Normal

Evaluation of fitting reports

- @RISK returns a fitting report on the selecting distribution models
- Some models may have good fit, others may have low fit to the input data
- The following options are available for evaluation of models
  - Statistics on both the fitted model and the input data
  - Graph comparison, Probability-Probabiity plot (P-P plot), and Quantile-Quantile plots (Q-Q plot)
  - The goodness-of-fit (GOF) tests

Example of fitting report in @RISK



Fit Ranking

Fit	AIC
☑ Gamma	1712.0688
☑ Weibull	1716.9088
☑ LogLogistic	1718.8190
□ Lognorm	1730.3707
□ Triang	1735.2460
□ Expon	1730.9188
□ Uniform	1820.9156

Statistics

	Input	Gamma
Minimum	12.00	0.00
Maximum	348.00	+∞
Mean	106.65	106.65
Mode	≈52.00	65.08
Median	91.00	93.15
Std Dev	67.66	66.58
Skewness	1.2265	1.2486
Kurtosis	4.4955	5.3385
Left X	22	22
1%	13.00	12.32
5%	22.00	25.09
10%	34.00	35.04
15%	45.00	43.22
20%	52.00	50.64
25%	58.00	57.60
30%	66.00	64.58
35%	69.00	71.47
40%	77.00	78.46

Fit Ranking on the GOF test

Distribution functions

Delimiters

Graph comparison

Probabilities

Statistics information

P-P plot

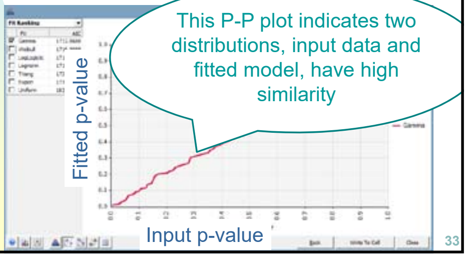
Q-Q plot

Comparison of the two graphs

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

P-P (Probability-Probability) plots

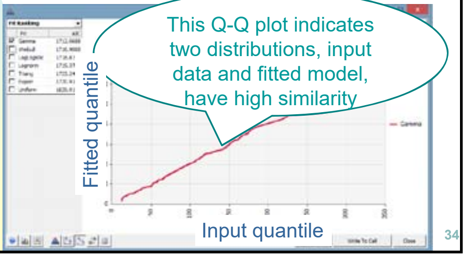
- Plots the p-value of the fitted distribution vs the p-value of the input data
- If the fit is good, the plot will be nearly linear of y = x



This P-P plot indicates two distributions, input data and fitted model, have high similarity

Q-Q (Quantile-Quantile) plots

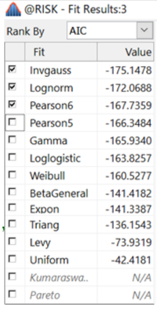
- Plots the quantiles of the fitted distribution vs the quantiles of the input data
- If the fit is good, the plot will be nearly linear of y =x



This Q-Q plot indicates two distributions, input data and fitted model, have high similarity

Fit ranking

- Ranks the fitted distributions according to the GOF test, such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Chi-square, Kolmogorov-Smirnov(K-S), Anderson-Darling (A-D)
- A GOF statistic provides a quantitative measure of how closely the fitted distribution matches the distribution of the input data.
- In general, a lower statistic values indicates a better fit


@RISK - Fit Results:3

Rank By AIC

Fit	Value
☑ Invgauss	-175.1478
☑ Lognorm	-172.0688
☑ Pearson6	-167.7359
□ Pearson5	-166.3484
□ Gamma	-165.9340
□ Loglogistic	-163.8257
□ Weibull	-160.5277
□ BetaGeneral	-141.4182
□ Expon	-141.3387
□ Triang	-136.1543
□ Levy	-73.9319
□ Uniform	-42.4181
□ *Kumaraswa..*	*N/A*
□ *Pareto*	*N/A*

Selection of the best fitting distribution model

- Taking into account evaluations for all graphs, statistics, and reports, select the best fitting distribution model for models
- Risk managers or risk assessors have a responsibility on the selection of the model
- There is no absolute correct answer for the best fitting
- Selection of the model affects the result of Monte Carlo simulation as a model uncertainty

FITTING OCCURRENCE DATA FOR ESTIMATION OF MAXIMUM LEVEL

Establishment maximum levels for contaminants in foods

(Review of the lectures of day1 and day2)

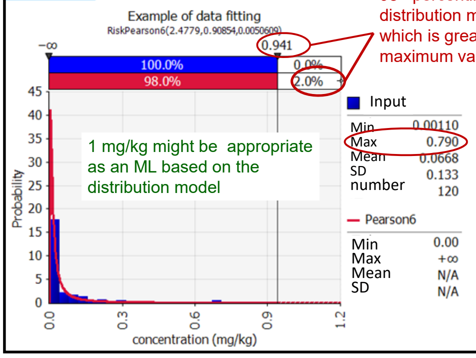
- Maximum levels should be set based on occurrence data follow ALARA principle
- Parametric approach for assuming distribution of occurrence data can be used for estimation draft maximum levels
- Probabilistic modelling software such as @RISK can derive high percentile values from distribution models
- Effective for estimating population distribution and high percentiles from a limited number of sample data (Ideally there is a minimum number of samples using the concept of binomial distribution)

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 maximum 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 maximum value of the actual data

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



1 mg/kg might be appropriate as an ML based on the distribution model

98th percentile of the distribution model is 0.941, which is greater than maximum value 0.790

ESTIMATION OF DIETARY EXPOSURE

Scenario setting for estimation

- 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)

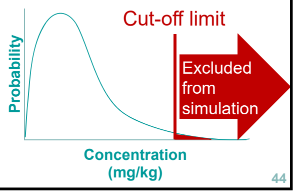
Formula of distribution model

- Select a distribution model of contaminants concentration by fitting evaluation
- Distribution model is described as a formula of distribution function
- Example of distribution function of @RISK
  =RiskLognormal(a,b)
  (returns a lognormal distribution)
  =RiskInvgauss(a,b)
  (returns an inverse Gaussian distribution)

Note: formula are defined by selected distribution models and input data

Cut-off limit of concentration data

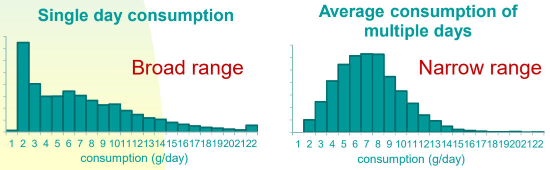
- Because sampling from a distribution model is at random, samples of extreme upper ends of distribution might be used
- A cut-off limit in the distribution tail can be set to avoid using unrealistic samples or samples higher than an ML



Distribution of food intake

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

Note: Shape and range of 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 a mathematical expression to calculate the intake based on the exposure scenario
- Be careful about unit of variables and intake (µg/kg bw/day, /week or /month)
  (0.001 kg = 1 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 and run simulation again

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 of scenario setting to the dietary intake
- Conduct uncertainty analysis

Refer to the lectures of day1 and day2

Let's try the exercise!