January 24, 2019
Training Workshop for Building Capacities
"Risk management of contaminants in foods"
Tokyo, Japan

Exercise 3 Analysis of Occurrence data

Takanori UKENA, Ph.D.
takanori_ukena130@maff.go.jp

MAFF
Ministry of Agriculture Forestry and Fisheries
Food Safety and Consumer Affairs Bureau

Exercise 3

3.1 Data aggregation

calculation of basic statistics

maximum, minimum, mean, median

3.2 Creating a frequency table, histogram

3.3 Calculation of high percentile

Exe 3.1 Data aggregation

Data analysis using occurrence data

Purpose:

- To estimate population (e.g. nationwide situation) form sample data
- For further consideration
  - setting maximum level
  - evaluating effectiveness of risk management measures
  - time-course analysis

Analysis of surveillance results

1. laboratory conditions
   - Sampling plan
   - Internal quality control
     - ✓ LOD, LOQ and their definitions
     - ✓ Calibration curve
     - ✓ Recovery
     - ✓Control material(CM) and frequency to test CM
   - Analytical results
     - ✓ Possibility of outliers
       Do not remove results without evidence.

Analysis of surveillance results

2. Dataset
   - Basic statistics
     mean, median, maximum and minimum value
   - Results below LOD or LOQ
     Replace <LOD and <LOQ with appropriate value for further data analysis.
     (depend on ratio of <LOD, <LOQ)
3. Making frequency table
4. Making histogram to check distribution
   parametric or non-parametric? multimodal?

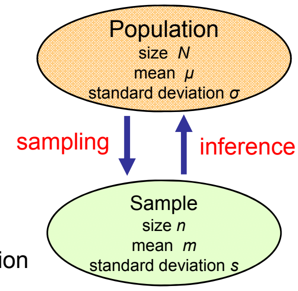
Analysis of surveillance results

5. Statistical analysis
- distribution model
- estimation of high percentiles
- exposure assessment

Basic statistics

Estimation of population from sample data.

- ✓ Maximum value
- ✓ Minimum value
- ✓ Range
- ✓ Average
- ✓ Mean (arithmetic mean)
- ✓ Median
- ✓ Variance
- ✓ Sample standard deviation



Median, Mode

- Median
  - the middle value in a set of values arranged in order of size:
  - the average of the two middle values if there is no one middle value.
  - a robust measure of central tendency
    Comparing to mean, median is robust to outlier value.
- Mode
  - a set of data values in a dataset that appears most often (most-frequently occurring value)

Percentile (%ile)

- The values are ranked in ascending order, i.e. from smallest to largest.
- Percentile is a number where a certain percentage of observations fall below that number. For example, the 20th percentile is the value below which 20% of the observations may be found.

  - ✓ 0 percentile (0%ile): minimum value
  - ✓ 25 percentile (25%ile): first quartile (Q_1)
  - ✓ 50 percentile (50%ile): median or second quartile (Q_2)
  - ✓ 75 percentile (75%ile) : third quartile (Q_3).
  - ✓ 100 percentile (100%ile) : maximum value

Exercise

- **Let's calculate basic statistics of Data1**

  - ✓ Maximum value
  - ✓ Minimum value
  - ✓ Range
  - ✓ Average
  - ✓ Mean (arithmetic mean)
  - ✓ Median
  - ✓ Variance
  - ✓ Sample standard deviation

parameter

Mean
- population mean μ
- sample mean $\bar{x}$

Variance
- population σ^2
- sample s^2

Standard deviation
- population σ
- sample s

Mean

population mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

sample mean

$$m \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

13

Deviation

- Deviation: difference between the observed value of a variable and mean

$$(x_i - \bar{x})$$

- Squared deviation from the mean

needed to calculate sample variance

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

14

Unbiased estimation of variance

- use of *n − 1* for sample variance formula instead of sample size *n*

Population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

15

Standard deviation (SD)

a measure to quantify the amount of variation or dispersion of a set of data values

Population standard deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$

Sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

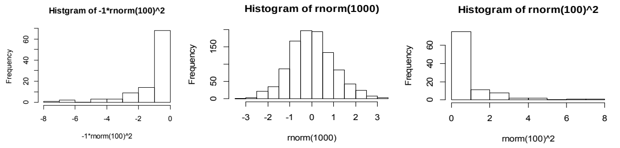
N : number of population
n : number of observations in the sample
x_i : observed values of the items
μ : population mean *$\bar{x}$: sample mean*

16

Skewness

degree of distortion from symmetrical curve

Sk < 0 : *left-skewed* Sk = 0: no skew Sk > 0: *right-skewed*



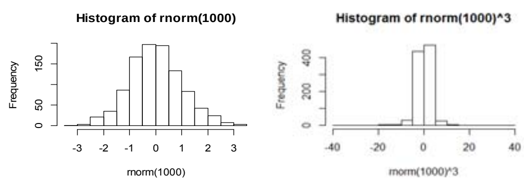
Skewness $(Sk) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$

n : sample size
x_i : observed values
$\bar{x}$: sample mean
s : sample SD

17

Kurtosis

High kurtosis is an indication of an outlier (or outliers)



$$kurtosis = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}$$

n : sample size
x_i : observed values
$\bar{x}$: sample mean
s : sample SD

defined as 0 or 3 for normal distribution

18

Analytical results below LOD, LOQ

How to deal with results below LOD, LOQ?

Assume
LOD = 0.03 mg/kg
LOQ = 0.05 mg/kg

Some results are below LOD or LOQ.

Calculation of LB, MB and UB

Examples

- Lower bound (LB)
  replacing all the results reported as below the LOD/LOQ by 0
- Medium bound (MB)
  i. replacing all the results reported as below the LOD/LOQ by half their respective LOD/LOQ
  ii. replacing all the results reported as below the LOD by half their respective LOD, and retain all the results reported between LOD and LOQ
- Upper bound (UB)
  replacing all the results reported as below the LOD/LOQ to their respective LOD/LOQ.

Aggregation of two datasets

Can we combine two datasets (data1 and data2) for further analysis?

For example, two datasets obtained by

a. completely different sampling plan for different purpose
b. slight different sampling plan for the same purpose
c. same sampling plan but different target
d. multi-years surveillance
e. same sampling plan but different basic statistics

Statistical test to compare two datasets

Parametric (normal distribution) or non parametric distribution?

a. Statistical normality test

For contaminants, datasets by surveillance usually have non parametric distribution

b. Statistical test to compare median
c. Statistical test to compare two distributions

Exercise two major non parametric statistical tests

a. Mann-Whitney *U* Test
- ✓ sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test
- ✓ test whether medians of two independent datasets come from the same population

(Kruskal–Wallis *H* test is used to compare medians for more than three independent datasets.)

b. Two-sample Kolmogorov–Smirnov test
- ✓ test whether the two independent datasets come from the same distribution

Mann-Whitney U Test (1)

Assumptions of Mann-Whitney U test

1. All the observations from both datasets are independent of each other,
2. The responses are ordinal (i.e., one can at least say, of any two observations, which is the greater),
3. Under the null hypothesis H_0, the distributions of both populations are equal.
4. The alternative hypothesis H_1 is that the distributions are not equal.

Mann-Whitney U Test (2)

1. Assign numeric ranks to all the observations (put the observations from both datasets to one set), beginning with 1 for the smallest value. Where there are groups of tied values, assign a rank equal to the midpoint of unadjusted rankings.
2. Add up the ranks for the observations which came from dataset 1.
3. Add up the ranks for the observations which came from dataset 2.
4. Statistic U is then given by:

$$U_1 = n_1 n_2 - \frac{n_1(n_1+1)}{2} - R_1$$

where n_1, n_2 is the sample size for dataset 1 and dataset 2 respectively, and R_1 is the sum of the ranks in dataset 1.

$$U_2 = n_1 n_2 - \frac{n_2(n_2+1)}{2} - R_2$$

where n_1, n_2 is the sample size for dataset 1 and dataset 2 respectively, and R_2 is the sum of the ranks in dataset 2.

5. The smaller value of U_1 and U_2 is the one used when consulting significance test.

Two-sample Kolmogorov–Smirnov test (1)

Assumptions of two-sample Kolmogorov–Smirnov test

1. All the observations from both detaset are independent of each other,
2. Under the null hypothesis H_0, both samples come from a population with the same distribution
3. The alternative hypothesis H_1 is that both samples do not come from a population with the same distribution

Two-sample Kolmogorov–Smirnov test (2)

1. First dataset has size m with an observed cumulative distribution function of $F(x)$, and the second dataset has size n with an observed cumulative distribution function of $G(x)$.
2. Calculate

$$D_{m,n} = \max_x |F(x) - G(x)|$$

3. The null hypothesis is rejected at level α if

$$D_{m,n,} \geq c(\alpha)\sqrt{\frac{(m+n)}{mn}}$$

Two-sample Kolmogorov–Smirnov test (2)

The value of $c(\alpha)$ is given in the table below for the most common levels of α.

α	0.10	0.05	0.025	0.01	0.005	0.001
c (α)	1.073	1.224	1.358	1.517	1.628	1.858

In general

$$c(\alpha) = \sqrt{-\frac{1}{2}\ln \alpha}$$

Data aggregation

Exercise:
Let's try to test using data1 and data2 if they can combine for further analysis.

- ✓ Mann Whitney U test
- ✓ Two-sample Kolmogorov–Smirnov test

Exe 3.2 Creating a frequency table, histogram

MAFF 31

Graphical expression

- Histogram
  - ✓ drawing histograms with various bin width
  - ✓ kernel density estimation
- P-P plot (Probability-Probability Plot)
- QQ plot (Quantile-Quantile Plot)
- Box plot (box and whisker plot)

32

Creating a frequency table, and histogram

Exercise:
Let's try to make frequency table and histogram using new dataset, combining dataset 1 and dataset 2.

MAFF 33

Making histogram

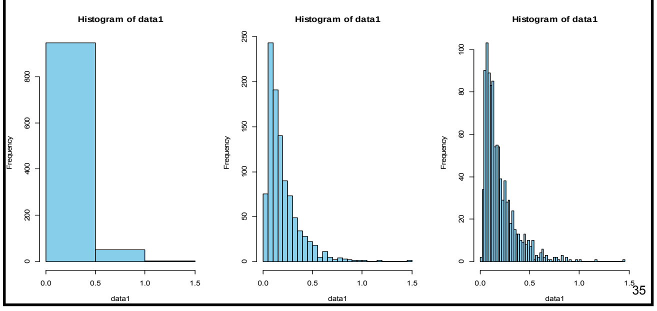
1. Purpose
   to graphically summarize the distribution of a data set
2. Steps to make histogram
   - ✓ Making frequency table
   - ✓ Frequency table
     - decide class interval or bin size, usually ten or more
     - need to consider border value to include lower or upper class
   - ✓ Calculating relative frequency, cumulative frequency
   - ✓ Making bar plot with no gap width between each bar

34

Effect of bin size using same dataset

Try to make histograms with various bin size

Too large bin width | Good bin width | Too small bin width



35

Examples to selecting bin size

$$Bin\ size = \frac{\max(x) - \min(x)}{k}$$

Square-root choice $k = \sqrt{n}$

Sturges' formula $k = \log_2 n + 1$

Scott's choice

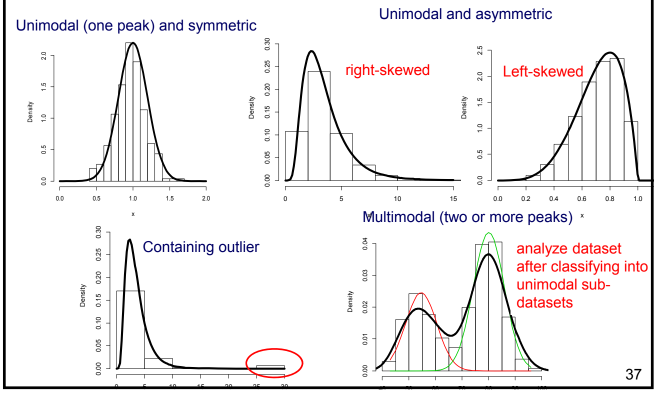
$$Bin\ size = \frac{3.5 \times \sigma}{n^{1/3}}$$

Freedman–Diaconis' choice

$$Bin\ size = \frac{2 \times IQR(x)}{n^{1/3}}$$

36

Shape of histogram and distribution



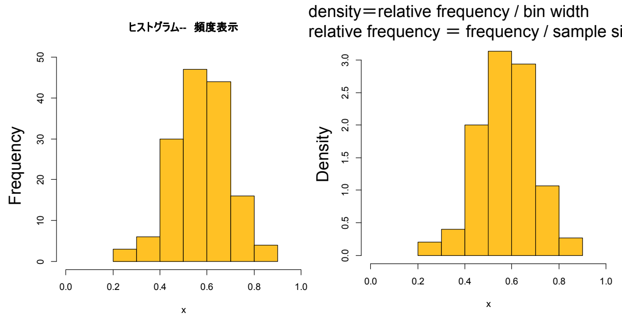
37

Histogram and Density histogram

Y：Frequency

Y：Frequency density

density＝relative frequency / bin width
relative frequency ＝ frequency / sample size



Density histogram is comparable to theoretical probability density distributions.

38

Kernel density estimation (1)

- Estimation of distribution not depend on bin number or class interval of histogram
- The KDE smoothes each data point *Xi* into a small density bumps and then sum all these small bumps together to obtain the final density estimate.

$\hat{f}_k(x)$ below is called the kernel function that is generally a smooth, symmetric function.

$$\hat{f}_K(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

$\hat{f}_k(x)$: Kernel density estimator
K: kernel function
h: h> 0, smoothing bandwidth that controls the amount of smoothing

39

Kernel density estimation (2)

$$\hat{f}_K(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

$\hat{f}_k(x)$: Kernel density estimator
K: kernel function
h: h> 0, bandwidth

Example of kernel functions

Gaussian kernel	$K(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$	
Epanechnikov kernel	$K(z) = \frac{3}{4}\left(1 - \frac{1}{5}z^2\right)/\sqrt{5}$	$(z < 5)$
	$K(z) = 0$	$(z \geq 5)$
Rectangular kernel	$K(z) = \frac{1}{2}$	$(\lvert z \rvert < 1)$
	$K(z) = 0$	$(\lvert z \rvert \geq 1)$

Band width (h) plays key role same as bin width of histogram.

40

Kernel density estimation (3)

Similar to bin size of histogram, KSD need to decide bandwidth to control amount of smoothing.

- ✓ When *h* is too small, there are many wiggly structures on our density curve.
- ✓ When *h* is too large, some important structures are obscured by the huge amount of smoothing.

Try to change *h* value to choose appropriate bandwidth.
The following is one example formula for selecting *h*.

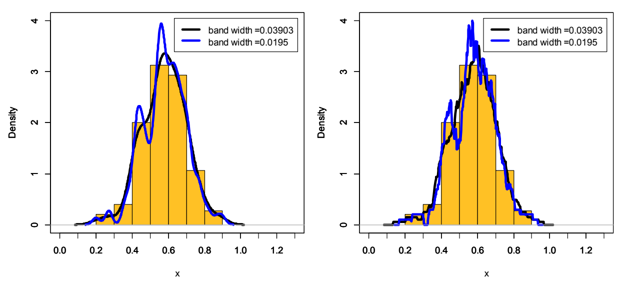
$$h = \frac{0.9\sigma}{n^{1/5}}$$

σ: standard deviation (sd) or IQR instead of sd

41

Examples of kernel density estimation

Gaussian kernel　　Rectangular kernel



42

Probability plot

Comparing two distribution *F(x)* and *G(x)* in graphical expression

⇒ Usually compare an empirical distribution with a theoretical distribution.

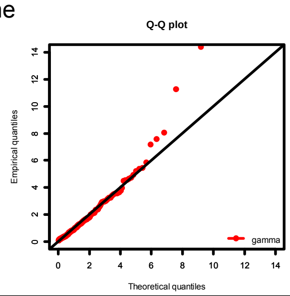
- Q-Q plot, P-P plot, CDF plot
  (Normal Q-Q plot and normal P-P plot is used to compare whether empirical distribution follow a *normal* distribution. The general QQ plot or PP plot is used to compare the distributions of any two datasets.)
- Q-Q plot and P-P plot follows the 45° line $y = x$ if the two distributions agree.

43

Q-Q (Quantile-Quantile) Plot (1)

- comparing two probability distributions by plotting their quantiles against each other

Q-Q plot follows the 45° line y = x if the two distributions agree.



44

Q-Q (Quantile-Quantile) Plot (2)

Example of calculating quantiles of items

- Order items from minimum (1) to maximum(n)
- Calculate quantiles using the following formula

$$f_i = \frac{i - 0.5}{n} \quad \text{or} \quad f_i = \frac{i}{n+1} \quad (i = 1 \sim n)$$

example

Observed value	fi	Observed value	fi
0.21	0.05	0.90	0.55
0.35	0.15	1.00	0.65
0.50	0.25	1.01	0.75
0.64	0.35	1.12	0.85
0.79	0.45	5.56	0.95

45

P-P plot, CDF plot

PP (probability–probability) plot

- Plots the two cumulative distribution functions (CDF) against each other

CDF plot

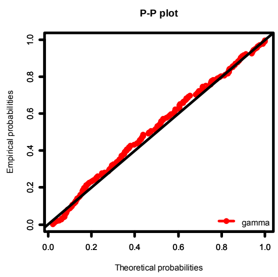
The cumulative distribution function (cdf) is the probability that the variable takes a value less than or equal to x.

46

P-P (probability–probability) Plot

- plots the two cumulative distribution functions (CDF) against each other

P-P plot follows the 45° line y = x if the two distributions agree.



47

P-P plot

Example of calculating CDF

example

Data1	Rank	Cumulative Probability
0.21	1	0.11
0.35	2	0.22
0.50	3	0.33
0.64	4	0.44
0.79	5	0.56
0.90	6	0.67
1.00	7	0.78
1.01	8	0.89
1.12	9	1.00

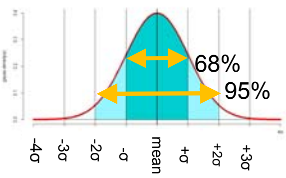
48

Cumulative distribution function (CDF)

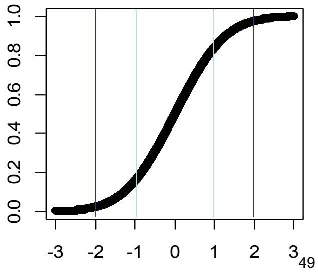
CDF: The area under the probability distribution function from $-\infty$ to x

Example: normal distribution

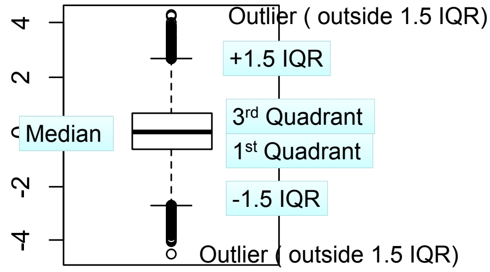
- 68% in $\pm 1\sigma$
- 95% in $\pm 2\sigma$



CDF of normal distribution

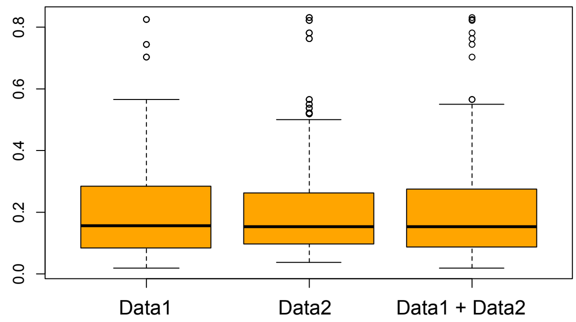


Graphical Distribution, Box-plot



- ✓ Inter-Quartile Range (IQR)
  3rd quartile (75%th quantile) - 1st quartile (25%th quantile)
- ✓ Normalized IQR: 0.7413 × IQR
  If data is normally distributed, converge to standard deviation

Example of box plot



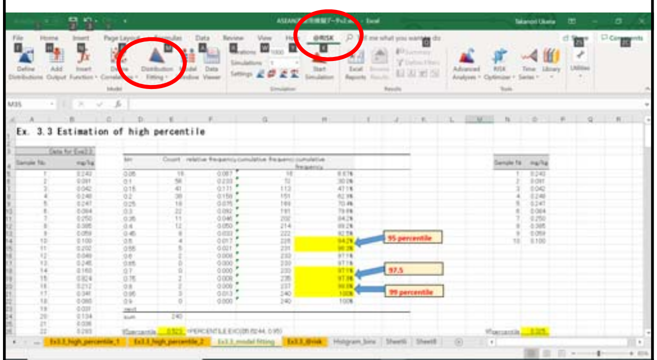
Exe 3.3 Calculation of high percentile

Exercise 3.3

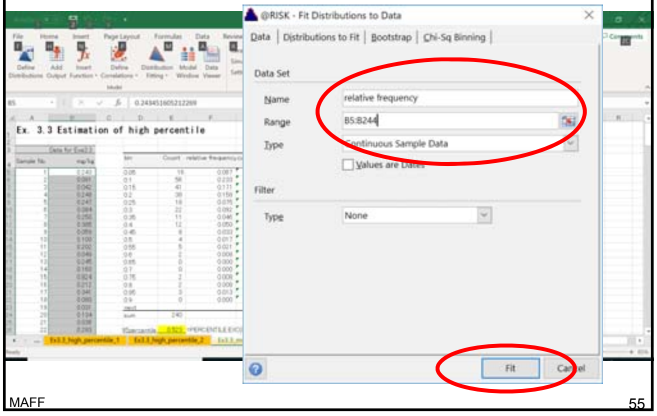
Not only high percentiles of sample data, we want to estimate high percentiles of population.

High percentiles of population are estimated from model distribution from real dataset using statistical computing software.

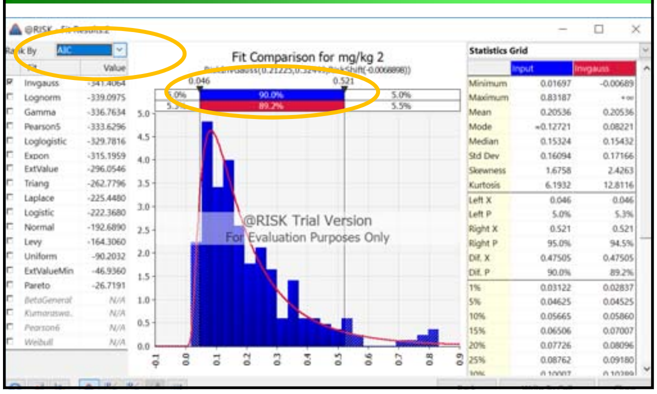
Using @risk for curve fitting



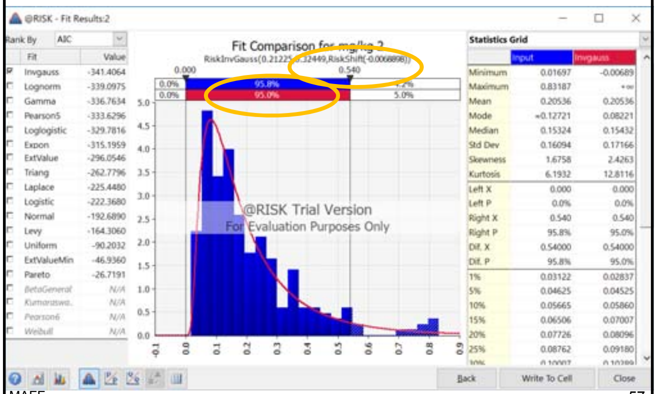
Select data set for fit distribution



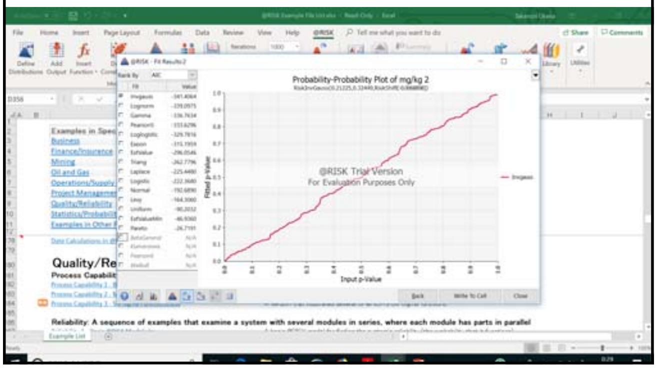
Moving slider to calculate high percentile in theoretical distribution



Example estimate 95 percentile



Check P-P plot for assessing how closely two data sets agree



Exercise 3.3

Let's calculate high percentiles fitting with inverse gaussian distribution.

- ✓ 97.5 percentile
- ✓ 99 percentile

Well done !

Summary– important points

- Data analysis is critical for risk management.
- Exercise by yourself for better understanding.
- In actual situation, collaboration with government scientists, laboratory analytical chemists, and statisticians is needed.

61