

# Comparison of Methods for Analyzing and Interpreting Censored Exposure Data

Paul Hewett Ph.D. CIH

Exposure Assessment Solutions, Inc.

Gary H. Ganser Ph.D.

West Virginia University



# Comparison of Methods for Analyzing and Interpreting Censored Exposure Data

- I. Introduction
- II. Methods
- III. Results
- IV. Recommendations
- V. Research Opportunities

# I. Introduction

- ◆ A review of the censored data literature revealed ...
  - poorly described methods
  - contradictory recommendations
  - confusing articles
  - few articles directed toward the IH scenario.
- ◆ The purpose of this study was to ...
  - compare four standard methods and a proposed method for analyzing low to medium censored datasets.



◆ For more information ...

- Hewett, P. and Ganser, G.H.:
  - ◆ **A Comparison of Several Methods for Analyzing Censored Data**
  - ◆ (submitted to AOH, April 2006)
- Ganser, G.H. and Hewett, P.:
  - ◆ **An Accurate Substitution Method for Analyzing Censored Data**
  - ◆ (in preparation)

# Definitions

## ◆ Left censored

- measurements occurred that were below the limit of detection (LOD)

## ◆ Right censored

- measurements occurred that were above the *maximum measurable concentration*

## ◆ Left truncated

- measurements less than the LOD were removed from the dataset

## ◆ Right truncated

- measurements above the *maximum measurable concentration* were removed from the dataset

# Type of Censored Dataset

## ◆ Simple Censored

- A dataset with a single or multiple censoring points, but all are at the low end.
- e.g.,  $x = \{<LOD, <LOD, x_3, x_4, \dots, x_n\}$
- e.g.,  $x = \{<LOD_1, <LOD_2, x_3, x_4, \dots, x_n\}$

## ◆ Complex Censored

- A dataset with multiple censoring points spread throughout the data.
- e.g.,  $x = \{<LOD_1, x_2, x_3, <LOD_2, \dots, x_n\}$

## Degree of Censoring

Degree of Censoring	Percent Censored
<b>Low</b>	<20%
<b>Medium</b>	20% - 50%
High	>50%
Severe	80% to 100%

## Example Datasets (OEL = 100 $\mu\text{g}/\text{m}^3$ )

Case	<b>Dataset 1*</b> ( $\mu\text{g}/\text{m}^3$ )	Dataset 2 ( $\mu\text{g}/\text{m}^3$ )	Dataset 3 ( $\mu\text{g}/\text{m}^3$ )	Dataset 4 ( $\mu\text{g}/\text{m}^3$ )
1	<3	<3	<3	<30
2	<3	<3	<3	<30
3	<3	<3	<3	<30
4	3.06	<3.1	<3.1	<30
5	4.41	4.4	<4.4	<30
6	7.23	<7.2	<7.2	<30
7	8.29	8.3	<8.3	<30
8	9.52	<9.5	<9.5	<30
9	19.94	19.9	<19.9	<30
10	20.25	20.3	<20.3	<30

\* Data source: Finkelstein and Verma, 2001.

# How can we analyze a simple censored dataset (i.e., Dataset 1)?

- ◆ Simple substitution (LOD/2 & LOD/ $\sqrt{2}$ )
- ◆ Log-Probit Regression (LPR)
- ◆ Maximum Likelihood Estimation (MLE)
- ◆  $\beta$ -Substitution ( $\beta$ -Sub; proposed)

◆ There are other methods, as well as variations of the LPR and MLE methods.

# $\beta$ -Substitution (a proposed method)

Step 1: Create an array of the observed, uncensored data. Let  $n$  = total sample size and  $k$  = number of measurements  $<$  LOD.

Step 2: Calculate input values:

$$\bar{y} = \frac{1}{n-k} \sum_{i=1}^{n-k} y_i \quad \text{where } y_i = \ln(x_i) \text{ and } x_i \text{ is the } i\text{th uncensored datum}$$

$$z = \Phi^{-1} \left[ \frac{k}{n} \right] \quad \text{b}$$

$$f(z) = \frac{\text{pdf}(z, 0, 1)}{1 - \text{cdf}(z, 0, 1)} \quad \text{c d}$$

$$\hat{s}_y = \frac{\bar{y} - \ln(\text{LOD})}{f(z) - z}$$

Step 3: Calculate  $\beta_{GM}$  :

$$\beta_{GM} = \exp\left[-\hat{s}_y \cdot \left(z + \frac{n-k}{k} \cdot f(z)\right)\right]$$

Step 4: Substitute each LOD with  $\beta_{GM} \cdot \text{LOD}$ .

Step 5: Calculate the simple arithmetic mean and sample GM using the array of uncensored and substituted measurements.

Step 6: Calculate  $\beta_{GSD}$  :

$$\beta_{GSD} = \exp\left[\hat{s}_y \cdot \left(-z + f(z) - \sqrt{\frac{n-1}{n-k} + \frac{n-1}{k} \cdot f(z)^2 - \frac{n-1}{k} \cdot z \cdot f(z)}\right)\right]$$

Step 7: Substitute each LOD with  $\beta_{GSD} \cdot \text{LOD}$ .

Step 8: Calculate the sample GSD using the array of uncensored and substituted measurements.

Step 9: Calculate the sample 95<sup>th</sup> percentile using the sample GM and GSD and Equation 2.

	A	B	C	D	E	F	G	H	I	J	K
1											
2		n =	10	y bar =	2.1357						
3		k =	3	z =	-0.5244						
4		n-k =	7	f(z) =	0.4967		beta_gm =			beta_gsd =	
5		LOD =	3	sy =	1.0156		0.524923			0.514465	
6											
7		Case	x	LOD	y		x	y		x	y
8		1	3	1			1.5748	0.4541		1.5434	0.4340
9		2	3	1			1.5748	0.4541		1.5434	0.4340
10		3	3	1			1.5748	0.4541		1.5434	0.4340
11		4	3.06	0	1.1184		3.06	1.1184		3.06	1.1184
12		5	4.41	0	1.4839		4.41	1.4839		4.41	1.4839
13		6	7.23	0	1.9782		7.23	1.9782		7.23	1.9782
14		7	8.29	0	2.1150		8.29	2.1150		8.29	2.1150
15		8	9.52	0	2.2534		9.52	2.2534		9.52	2.2534
16		9	19.94	0	2.9927		19.94	2.9927		19.94	2.9927
17		10	20.25	0	3.0082		20.25	3.0082		20.25	3.0082
18											
19					mean =		7.74				
20							y bar =	1.63		s_y =	1.004196
21							gm =	5.11		gsd =	2.730
22											
23										X95 =	26.66

# Minimum Censored Datasets

## ◆ LOD/2 and LOD/ $\sqrt{2}$ Substitution

- $n \geq 2$  and at least 1 meas.  $\geq$  LOD

## ◆ Log-Probit Regression (LPR)

- $n \geq 3$  and at least 2 meas.  $\geq$  LOD

## ◆ Maximum Likelihood Estimation (MLE)

- $n \geq 3$  and at least 2 meas.  $\geq$  LOD

## ◆ $\beta$ -Substitution

- $n \geq 3$  and at least 2 meas.  $\geq$  LOD

# Analysis of Dataset 1

(n' = number of uncensored data)

Method	n'	GM	GSD	$X_{0.95}$	Mean
Substitute LOD/2	7	5.04	2.76	26.78	7.72*
Substitute LOD/ $\sqrt{2}$	7	5.59	2.42	23.93	7.91*
Log-Probit Regression	7	5.31	2.77	28.38	8.10**
$\beta$ -Substitution	7	5.11	2.73	26.66	7.74*
Max. Likelihood Estimation	7	5.17	2.64	25.53	7.78**

\* Simple arithmetic mean

\*\* MVUE where n=10 for MLE and  $\beta$ -Sub and n=7 for LPR

## Issue

- ◆ *In the long run*, across a variety of datasets where the true GSD, true %censored, and sample size varies, which method is preferable for estimating...
  - GM and GSD (i.e., distribution parameters)
  - 95<sup>th</sup> percentile (i.e., a compliance statistic)
  - Mean (useful when construction a JEM).

## II. Methods

### ◆ Computer simulation 1:

- Generated 100,000 datasets for each combination of:
  - GSD: 1.5, 2, 3, and 4
  - n: 3, 5, 10, 20, 50, and 100
  - %censored: 0%, 10%, 20%, 30%, 40%, and 50%

### ◆ Calculated bias, precision, and overall accuracy for each method and the four parameters:

- GM, GSD,  $X_{0.95}$ , and Mean

### ◆ Note: the root mean square error (rMSE) is used to indicate the overall accuracy

# Accuracy = Bias + Precision

$$\text{Bias} = (\bar{x} - \theta)$$

Where  $\bar{x}$  = mean of results of N repeated simulations and  $\theta$  = true value.

$$\text{Precision} = \sqrt{\frac{\sum (x - \bar{x})^2}{N-1}}$$

$$rMSE = \text{Accuracy} = \sqrt{(\bar{x} - \theta)^2 + \frac{\sum (x - \bar{x})^2}{N-1}}$$

## ◆ Computer simulation 2:

- Generated 3 sets of 100,000 datasets where n, %censored, and the GSD were allowed to vary:
- n: **10** to 100
- %censored: 0% to 50%
- GSD:
  - ◆ 1.2 – 2 (low variability)
  - ◆ 2 – 3 (medium variability)
  - ◆ 3 – 4 (high variability)

## ◆ Calculated bias, precision, and overall accuracy (i.e., rMSE)



Note:

- **Valid, censored datasets** were analyzed using the selected censored data method.
- **Valid, uncensored datasets** were analyzed using standard formulae.
- **Invalid datasets** were not analyzed:
  - ◆ Completely censored
  - ◆ Too few uncensored data
  - ◆ [An issue when n was small (<10)]

# III. Results – Computer simulation 1

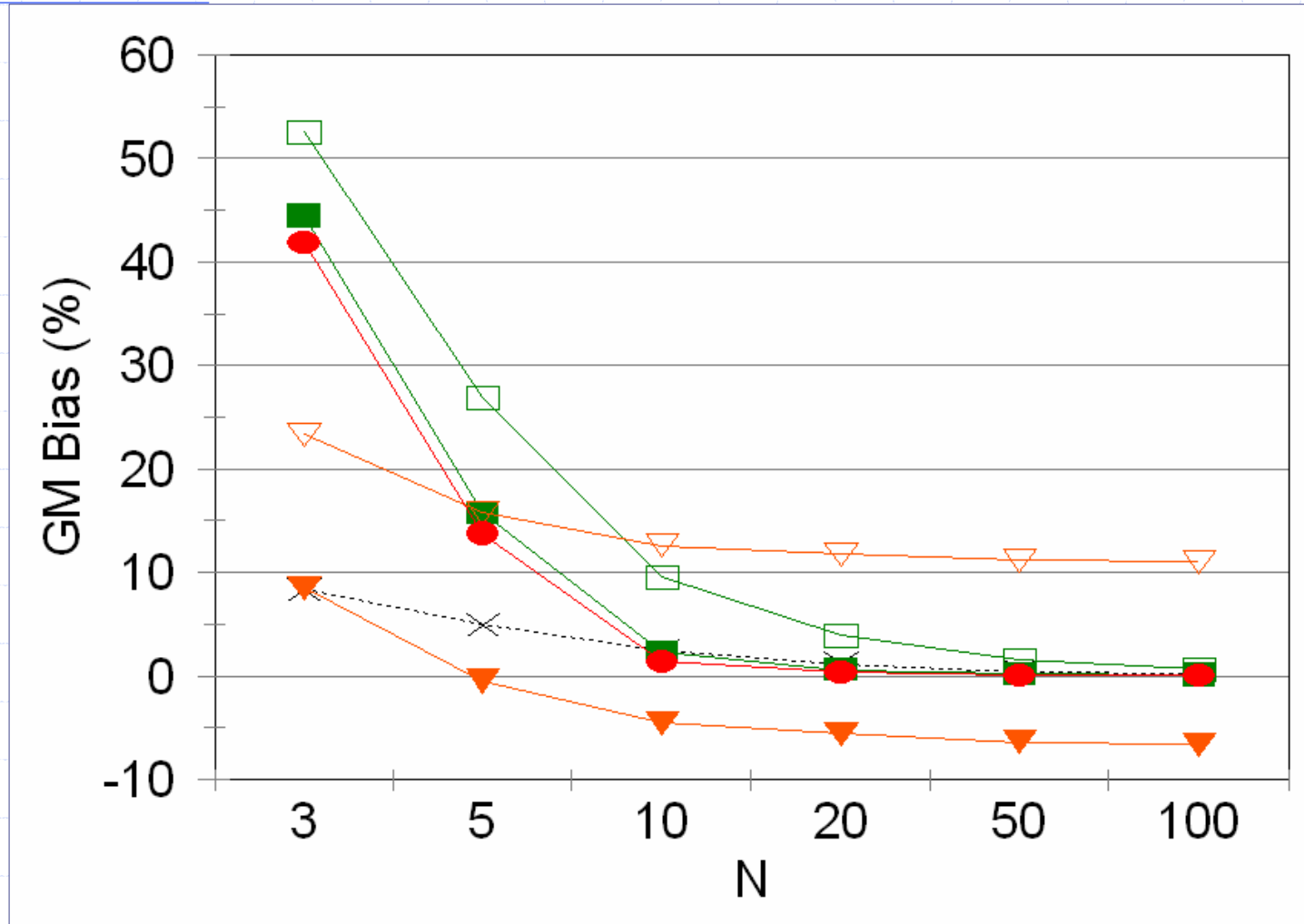
## ◆ Notes:

- Use the 0 %censored curve as the baseline for comparison.
- 32 charts were generated for each method
- ***For illustrative purposes only the results for GSD=2 and %censored=50% are shown.***
- Different methods were used to calculate the mean:
  - ◆ MVUE equation for the MLE and LPR methods
  - ◆ Arithmetic mean for the substitution methods

# GM Bias

(GSD=2 and %censored=50%)

(Legend: X 0%, # MLE, Q LPR,  $\nabla$  LOD/2,  $\square$  LOD/ $\sqrt{2}$ , !  $\beta$ -Sub)



$n = 3,$

true %censored = 50%

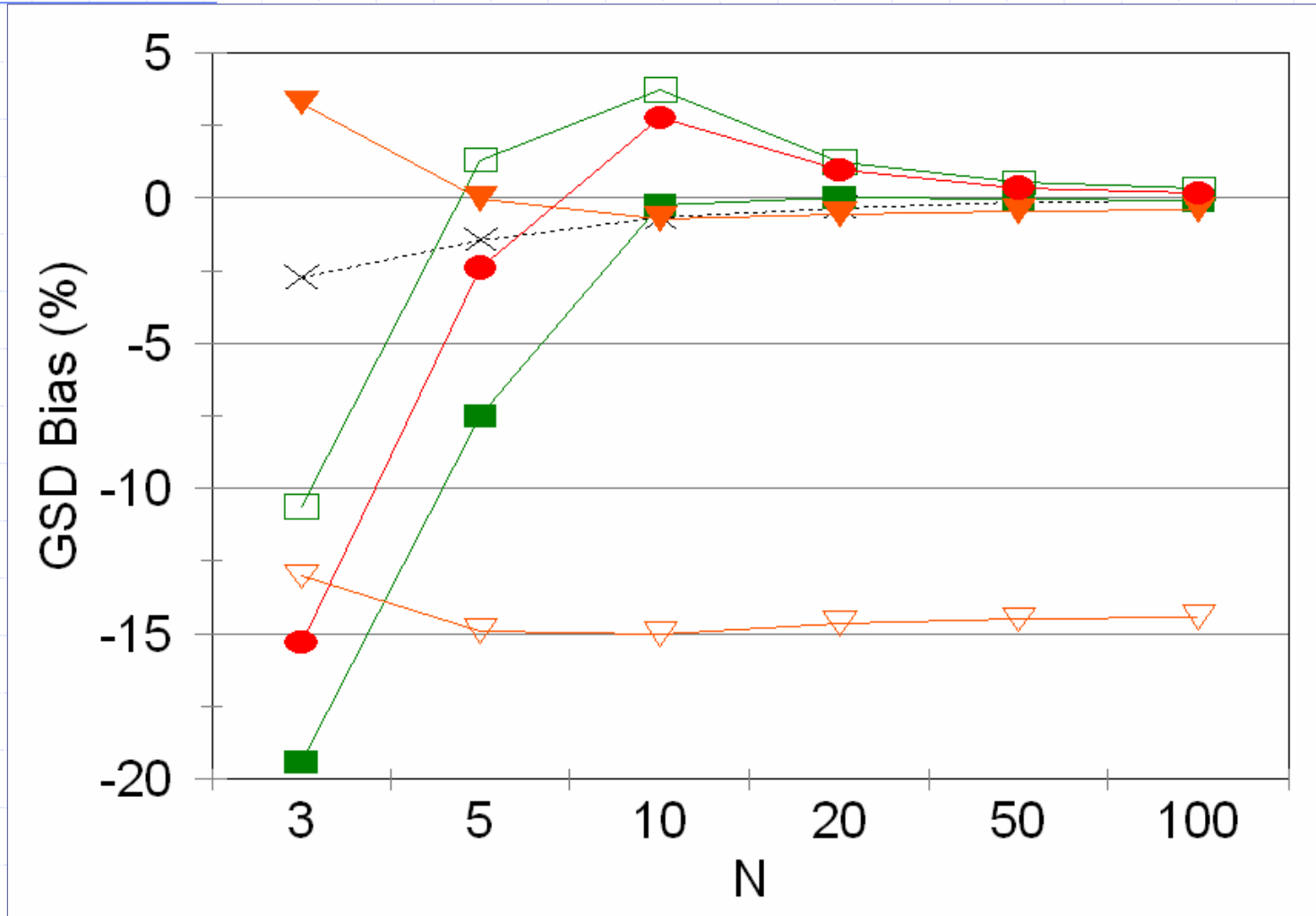
$m =$  (number of measurements  $<$  LOD)

Possible Outcomes			
$m=3$ (completely censored)	$m=2$	$m=1$	$m=0$ (uncensored)
↓ ↓ ↓	↑ ↓ ↓ ↓ ↑ ↓ ↓ ↓ ↑	↓ ↑ ↑ ↑ ↓ ↑ ↑ ↑ ↓	↑ ↑ ↑

# GSD Bias

(GSD=2 and %censored=50%)

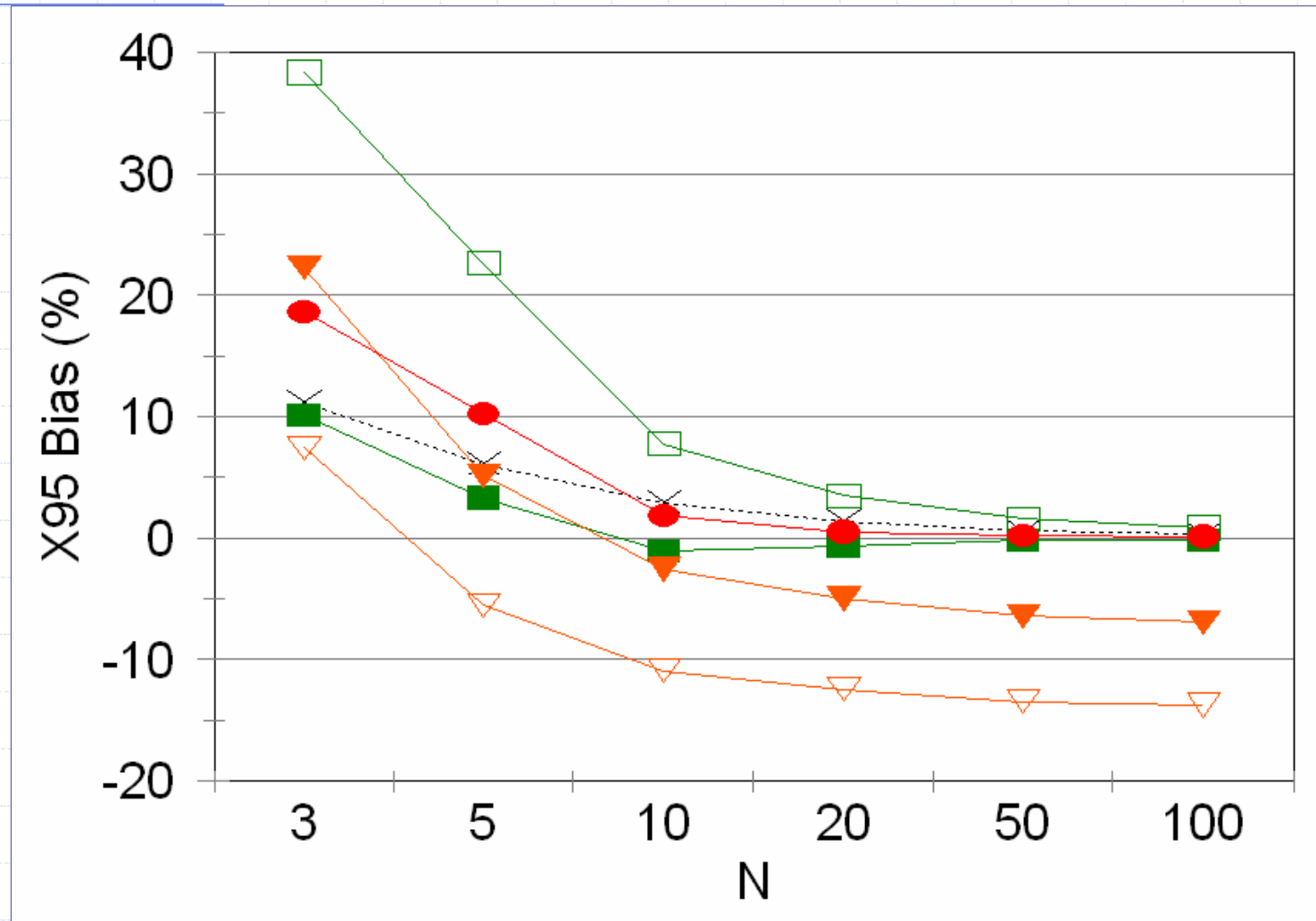
(Legend: X 0%, # MLE, Q LPR,  $\nabla$  LOD/2,  $\square$  LOD/ $\sqrt{2}$ , !  $\beta$ -Sub)



# $X_{0.95}$ Bias

(GSD=2 and %censored=50%)

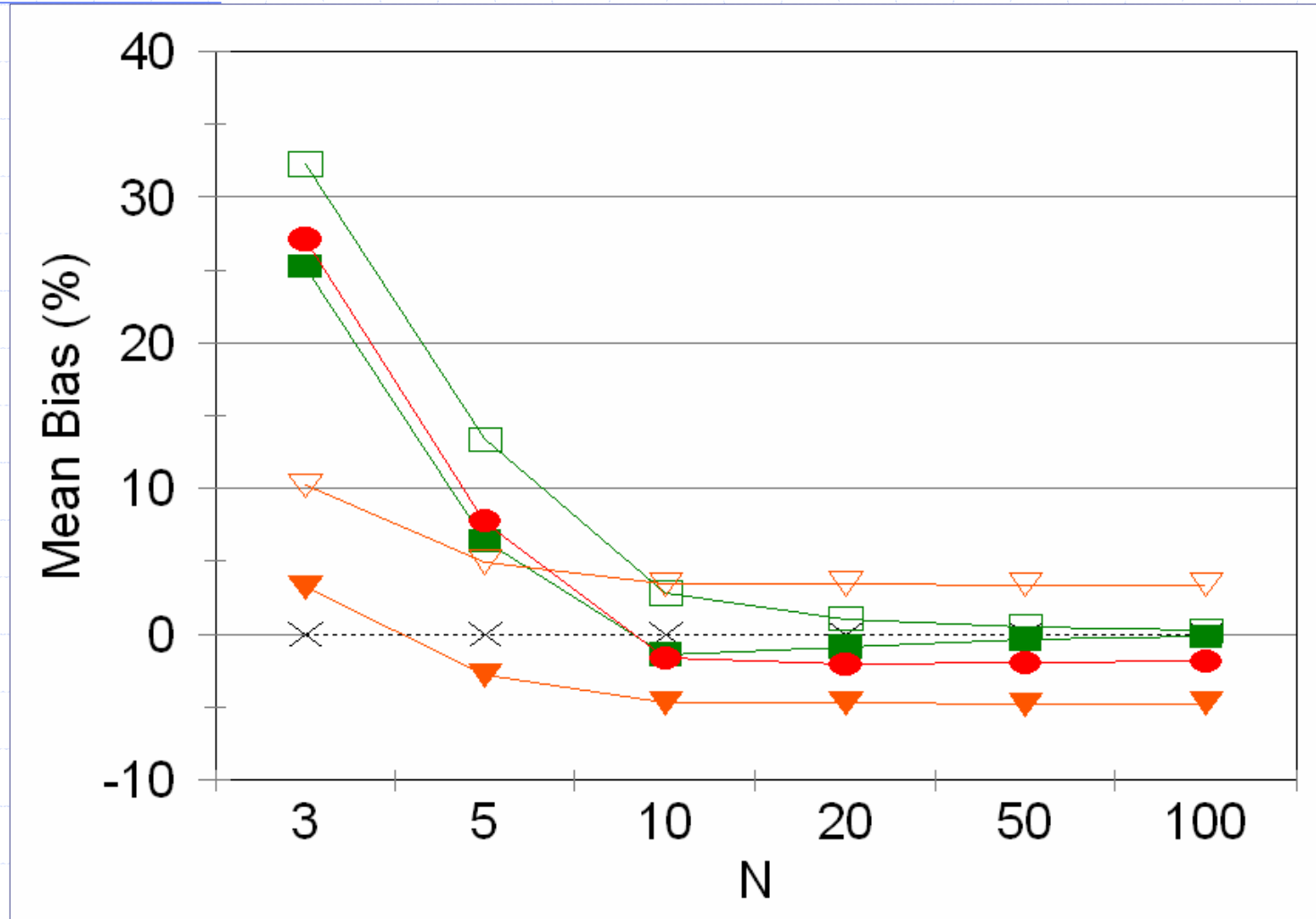
(Legend: X 0%, # MLE, Q LPR,  $\nabla$  LOD/2,  $\square$  LOD/ $\sqrt{2}$ , !  $\beta$ -Sub)



# Mean

(GSD=2 and %censored=50%)

(Legend: X 0%, # MLE, Q LPR,  $\nabla$  LOD/2,  $\square$  LOD/ $\sqrt{2}$ , !  $\beta$ -Sub)



## Comments (for Simulation 1; all combinations)

### ◆ As $n$ increases...

- MLE, LPR, and  $\beta$ -Sub approach zero bias
- LOD/2 and LOD/ $\sqrt{2}$  both approach a fixed bias
- By  $n=10$  MLE is reasonably close (i.e.,  $\pm 5\%$ ) to the *baseline* bias for all parameters

### ◆ For $n < 10$ ...

- All methods are biased
- LOD/2 and LOD/ $\sqrt{2}$  are less biased for the Mean

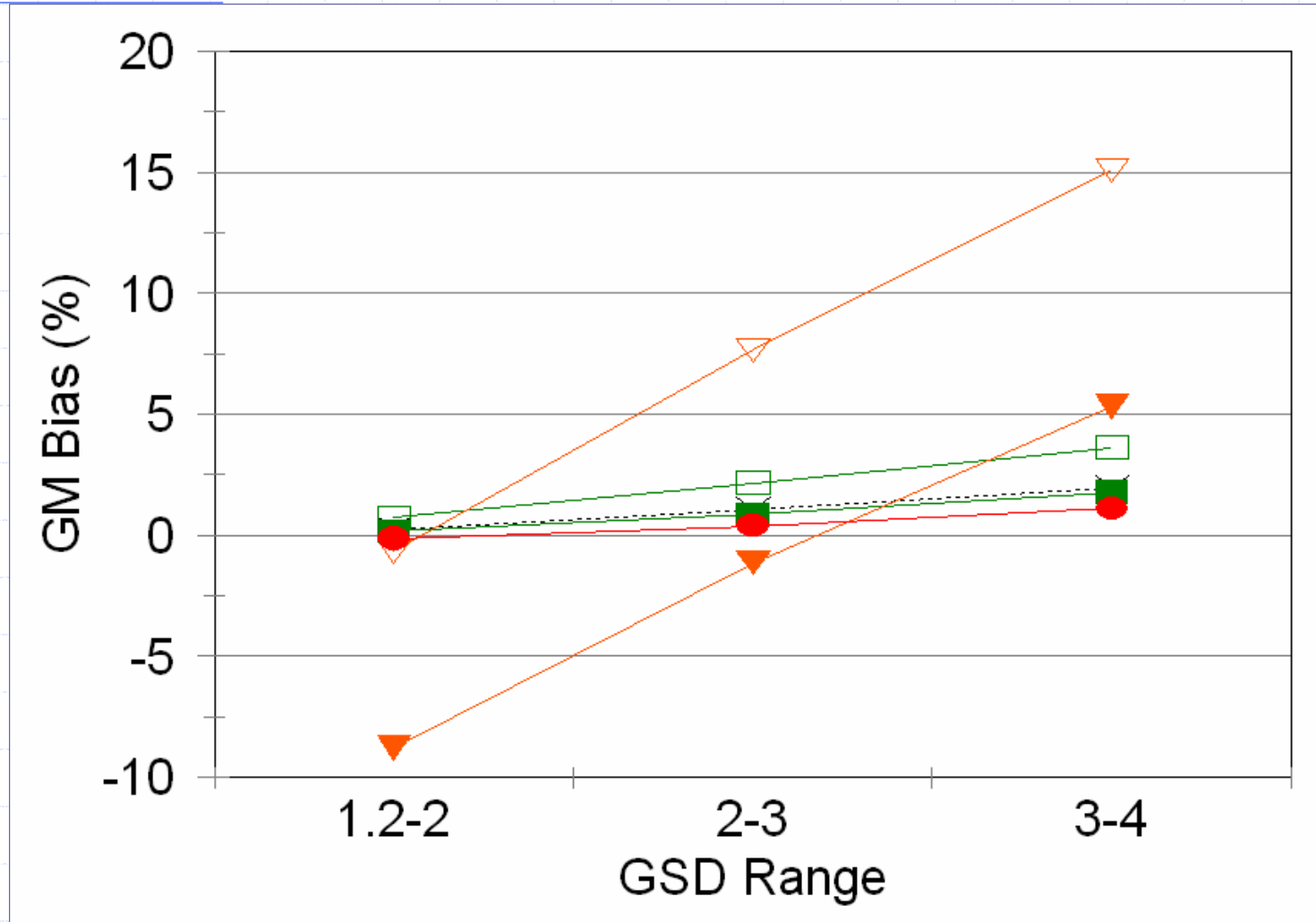
### ◆ $\beta$ -Sub bias closely tracks the MLE bias

# Results – Computer simulation 2

# GM Bias

(Composite datasets)

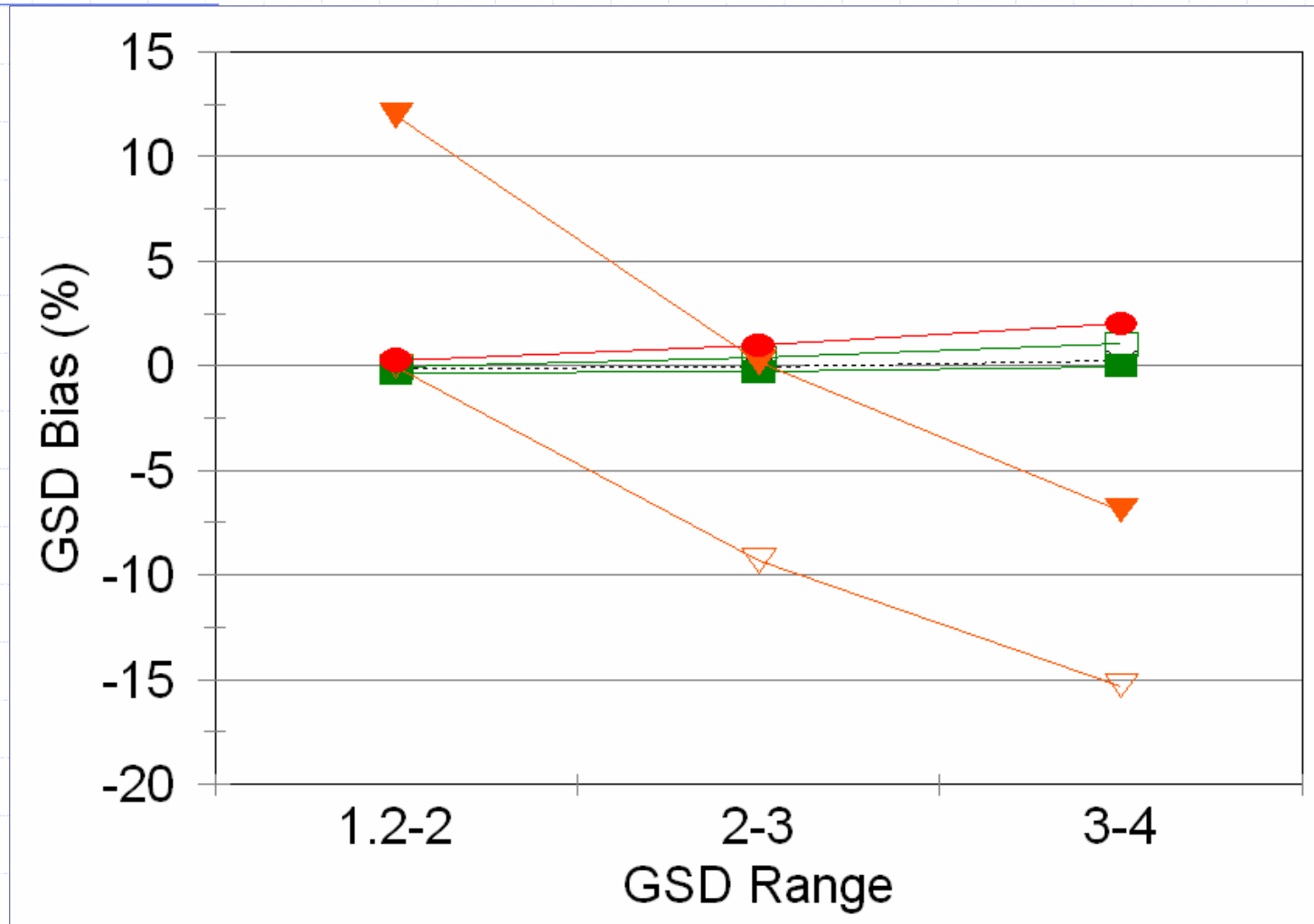
(Legend: X 0%, # MLE, Q LPR,  $\nabla$  LOD/2,  $\triangle$  LOD/ $\sqrt{2}$ , !  $\beta$ -Sub)



# GSD Bias

(Composite datasets)

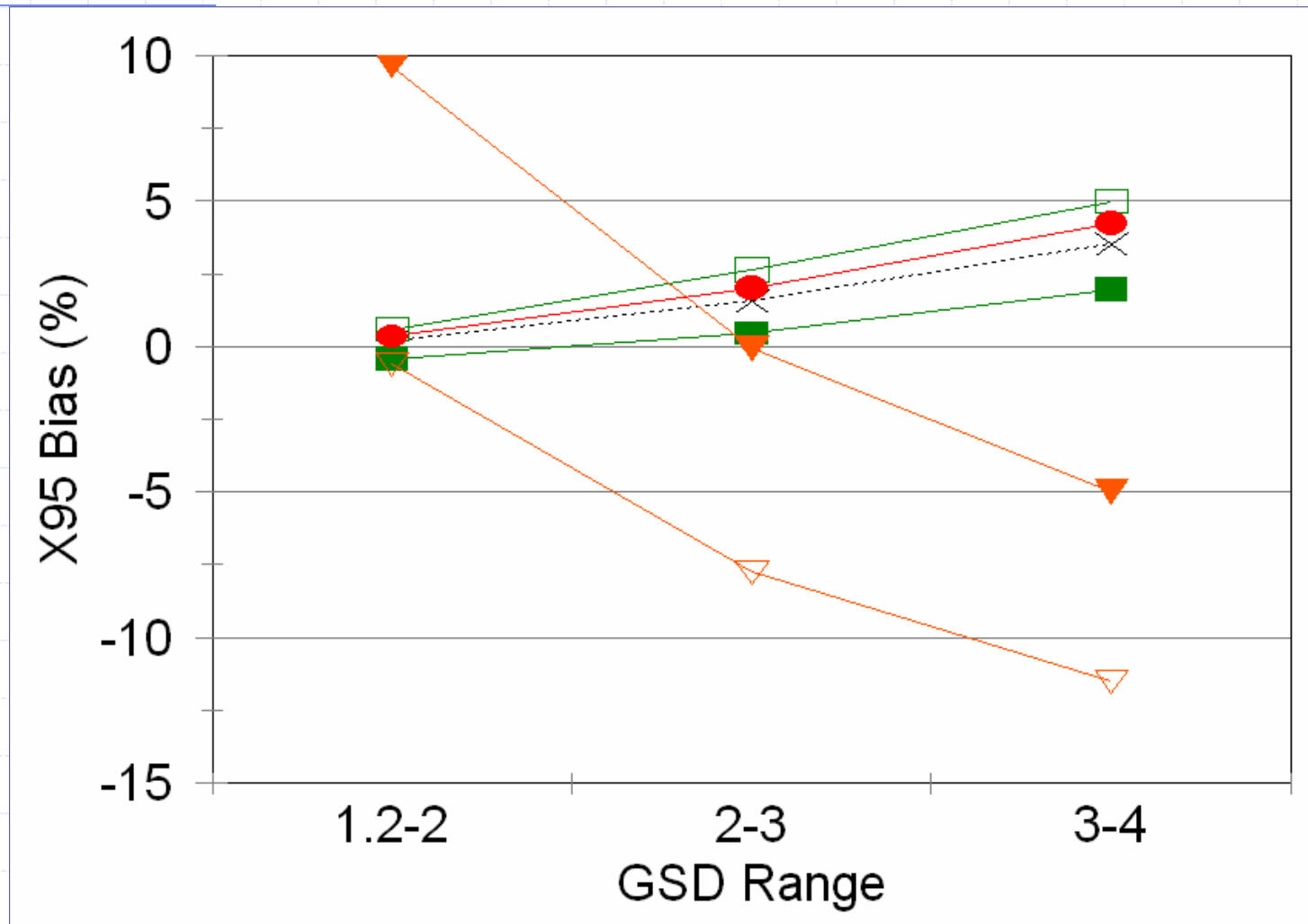
(Legend: X 0%, # MLE, Q LPR,  $\nabla$  LOD/2,  $\triangle$  LOD/ $\sqrt{2}$ , !  $\beta$ -Sub)



# X<sub>0.95</sub> Bias

(Composite datasets)

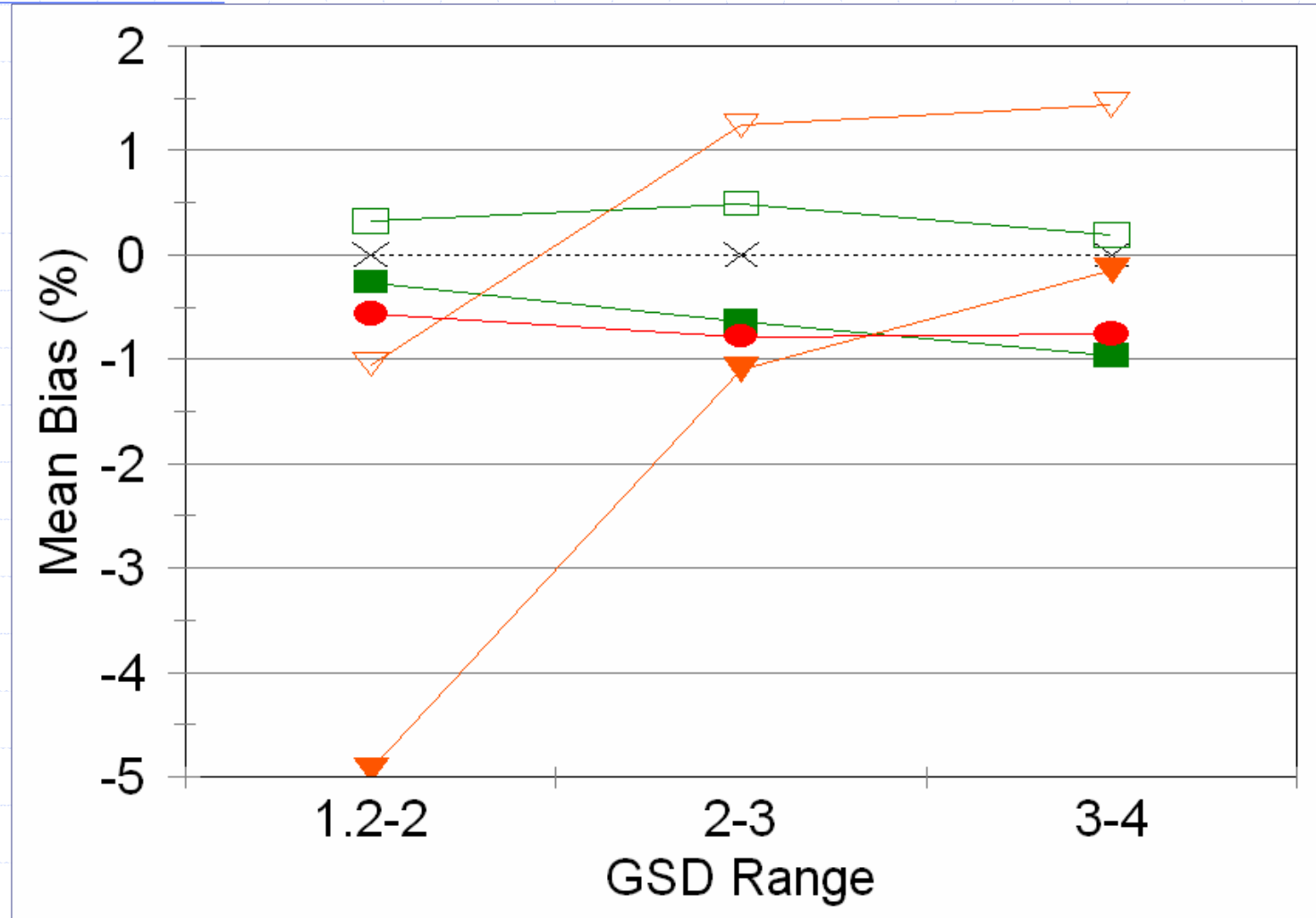
(Legend: X 0%, # MLE, Q LPR, ▾ LOD/2, ▽ LOD/√2, ! β-Sub)



# Mean Bias

(Composite datasets)

(Legend: X 0%, # MLE, Q LPR,  $\nabla$  LOD/2,  $\blacktriangledown$  LOD/ $\sqrt{2}$ , !  $\beta$ -Sub)



## Comments (for Simulation 2)

- ◆ MLE, LPR, and  $\beta$ -Sub tend to yield similar results that are close to the baseline.
- ◆ LOD/2 and LOD/ $\sqrt{2}$  tend to yield variable results that, in the long run, will be strongly biased – *except* when estimating the Mean.

## IV. Recommendations

- ◆ The MLE method appears to be “best” if  $n \geq 10$
- ◆  $\beta$ -Sub is a reasonable alternative to MLE
- ◆ For small sample sizes, say  $n < 10$ , all the methods are biased (+ or -) for the parameter estimates (GM and GSD), the 95th percentile, and the mean.
- ◆ The common substitution methods ...
  - tend to be strongly biased for the GM, GSD, and 95th percentile
  - are reasonably accurate when estimating the Mean
  - (tend to have comparable rMSE to the MLE method)

- ◆ Caution should be exercised when making important decisions based upon a highly censored and limited dataset.
  - Be aware of the direction and magnitude of the potential bias.
- ◆ Eliminate or reduce the need for Censored Data Analysis by reducing the LOD.
- ◆ *All of the above assumes that the appropriate model for occupational exposure data is the lognormal distribution.*

# What about highly censored datasets?

- ◆ For %censored up to 90% the computer simulation results were similar.
- ◆ Up to 80% censored use - with caution - MLE, LPR, or  $\beta$ -Sub if the sample size is 20 or greater.
  - Larger sample sizes are needed as the %censored increases.
- ◆ LOD/2 and LOD/ $\sqrt{2}$  should be avoided.
- ◆ Consider alternatives:
  - If LOD is  $\ll$  OEL, use Binomial Distribution calculations or Bayesian Decision Analysis to test hypotheses.

# What about severely censored datasets?

- ◆ Parameter estimation is not recommended or simply not possible.
- ◆ If LOD is  $\ll$  OEL, ...
  - use Binomial Distribution calculations or Bayesian Decision Analysis to test hypotheses.

# There are other CDA methods

## ◆ “Robust” variations

- robust LPR and robust MLE
- see Helsel (2005)

## ◆ Non-parametric statistics and methods

- Non-parametric percentiles and exceedance fractions
- Kaplan-Meier Method
  - ◆ Based upon “survival statistics”
  - ◆ See Helsel (2005)

## ◆ Decision making

- Test hypotheses using non-parametric methods or Bayesian Decision Analysis

## V. Research Opportunities

- ◆ Calculation of confidence limits
  - For each method, what sample size should be used to calculate confidence intervals, or should the confidence interval coefficient be adjusted?
- ◆ Analysis of complex censored datasets
  - Should the MLE method always be preferred, even for a highly complex censored dataset?
- ◆ Analysis of non-lognormal datasets
  - Which is preferred when the data are not well described by a *single* lognormal distribution?
  - How *robust* are the so-called robust methods?

# Contact Information

Paul Hewett PhD CIH

Exposure Assessment Solutions, Inc.

[phewett@oesh.com](mailto:phewett@oesh.com)

304.685.7050

# Abstract

- ◆ Exposure datasets often occur where one or more measurements are below the limit of detection (LOD). The purpose of this study was to test various standard censored data analysis methods when applied to low (i.e., <20% censored) and medium (20%-50%) censored datasets, and compare these results to those for beta-substitution, a newly developed substitution technique.
- ◆ Using computer simulation, the bias and root mean square error (rMSE) for the two commonly used substitution methods (i.e., LOD/2 and LOD/sqrt(2)), beta-substitution, log-probit regression (LPR), and maximum likelihood estimation (MLE) were determined for the scenario where there is a single LOD. The parameters estimated in the computer simulation were the distribution parameters (i.e., geometric mean and geometric standard deviation), the 95<sup>th</sup> percentile, and the mean.

- ◆ The MLE method has substantial bias for small  $n$ , but for the larger sample sizes is nearly unbiased. The bias for the common substitution methods can be positive or negative, depending upon the sample size, fraction censored, and geometric standard deviation. LPR is substantially biased for small sample sizes. The bias and rMSE for the beta-substitution method is nearly identical to that of the MLE method for all combinations of sample size, fraction censored, and GSD considered.
- ◆ The sample estimates for the parameters selected tend to be substantially biased for  $n < 10$ , so it is important to understand the direction and magnitude of the potential bias for each method. The common substitution methods are biased for all sample sizes and should be used with caution. LPR could be used whenever  $n \geq 20$ , but is clearly not suitable for small  $n$  unless only the mean is of interest. MLE is preferred above all others, but the beta-substitution method, which is easier to calculate, is nearly equal to the MLE method in terms of both bias and rMSE.