

Estimating Population Percentiles using the Turnbull Estimator when Some Data are Below the Limit of Detection

**Brenda Gillespie, Heidi Reichert,
Qixuan Chen, Al Franzblau, Jim Lepkowski,
Peter Adriaens, Avery Demond, William
Luksemburg and David Garabrant**

The University of Michigan Dioxin Exposure Study

The University of Michigan, Ann Arbor, MI, USA
School of Public Health
Institute for Social Research
College of Engineering
Center for Statistical Consultation & Research

August 24, 2009



Financial support for this study comes from The Dow Chemical Company through an unrestricted grant to the University of Michigan.

The University of Michigan has complete independence to design, carry out, and report the results of the study.

The investigators report to an independent Scientific Advisory Board (SAB).

Study Team

- **UM School of Public Health**
 - David Garabrant, MD, MPH**
 - Alfred Franzblau, MD
 - Olivier Jolliet, PhD
 - Lynn Zwica, MS
 - Kristine Knutson, MPH
 - Elizabeth Hedgeman, MS, MPH
 - Qixuan Chen, PhD**
 - Biling Hong, MA**
 - Shih-Yuan Lee, MS
 - Chiung-Wen Chang, MS
 - Xiaohui Jiang, MS**
 - Xiaobo Zhong, MS candidate
 - Jinkyung Ha
 - Camelia Sima, MS
 - Meghan Milbrath
 - Yvan Wenger, MS
 - Sharyn Vantine
- **UM Institute for Social Research**
 - James Lepkowski, PhD, MPH
 - Barbara Lohr Ward, MBA
 - Kathy Ladronka
 - Kristen Olson, PhD
 - Jennifer Sinibaldi, MS
- **UM Center for Statistical Consultation and Research**
 - Brenda W. Gillespie, PhD
 - Heidi Reichert, MA
 - Danielle Gwinn
 - Scott Swan, MS
- **UM College of Engineering**
 - Peter Adriaens, PhD, PE
 - Avery Demond, PhD, PE
 - Shu-Chi Chang, PhD
 - Hoa Trinh, MS
- **LimnoTech**
 - Tim Towey, MS
 - Noémi Barabas, PhD
- **Contractors**
 - Vista Analytical Laboratories
 - Environ
 - MidMichigan Medical Center
 - Mobile Medical Response
 - Foote Hospital
 - Lifespan Visiting Nurses
 - Regional Medical Labs

Acknowledgements

- The authors acknowledge:
 - For their assistance,
 - Dr. Donald Patterson Jr
 - Ms. Sharyn Vantine
 - for their guidance as members of our Scientific Advisory Board,
 - Dr. Linda Birnbaum
 - Dr. Ronald Hites
 - Dr. Paolo Boffetta
 - Dr. Marie Haring Sweeney

Background

- Data below a limit of detection (LOD) are termed non-detects, or **left-censored**.
- Common methods of handling:
 - Assign to zero
 - Assign to $LOD/2$ or $LOD/\sqrt{2}$
 - Maximum likelihood (ML) methods
- The method of handling non-detects should always be stated in papers or presentations.

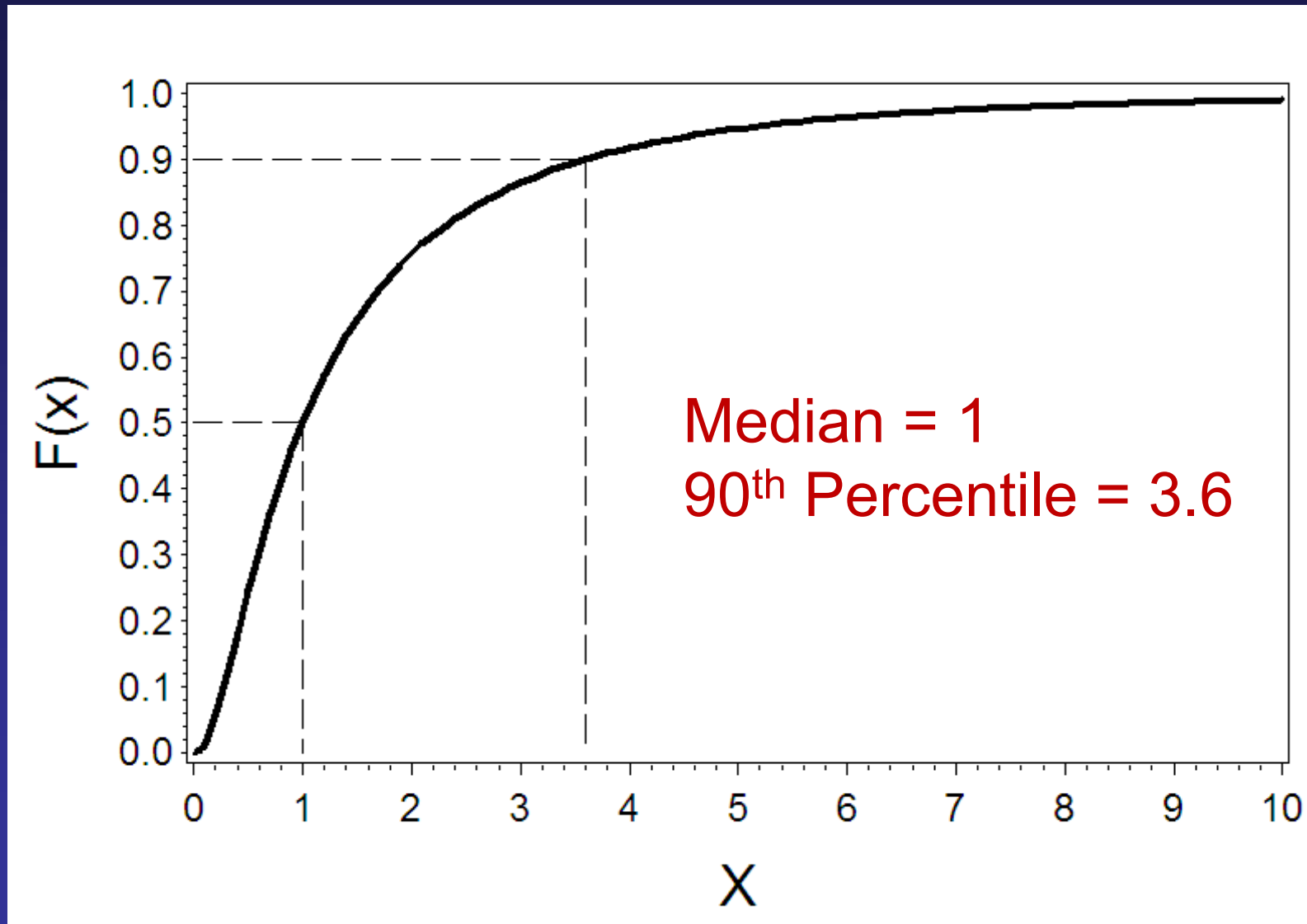
Background

- Estimating overall population percentiles is often a goal in reporting contamination levels in a population.
 - E.g., a population median or 95th percentile of a toxin level.

Outline of talk

- Introduce of the Turnbull estimator
 - Examples using serum dioxin data from
 - University of Michigan Dioxin Exposure Study (**UMDES**), $n = 251$ in control region
 - National Health and Nutrition Examination Survey (**NHANES**, 2003-2004), $n \sim 1790$
 - percents below LOD ranging from 12% to 97%.
- Give motivation for a nonparametric estimator of the population distribution
- Discuss software for calculating the Turnbull estimator

Example Cumulative Distribution Function



Estimating $F(x)$ with Complete Data

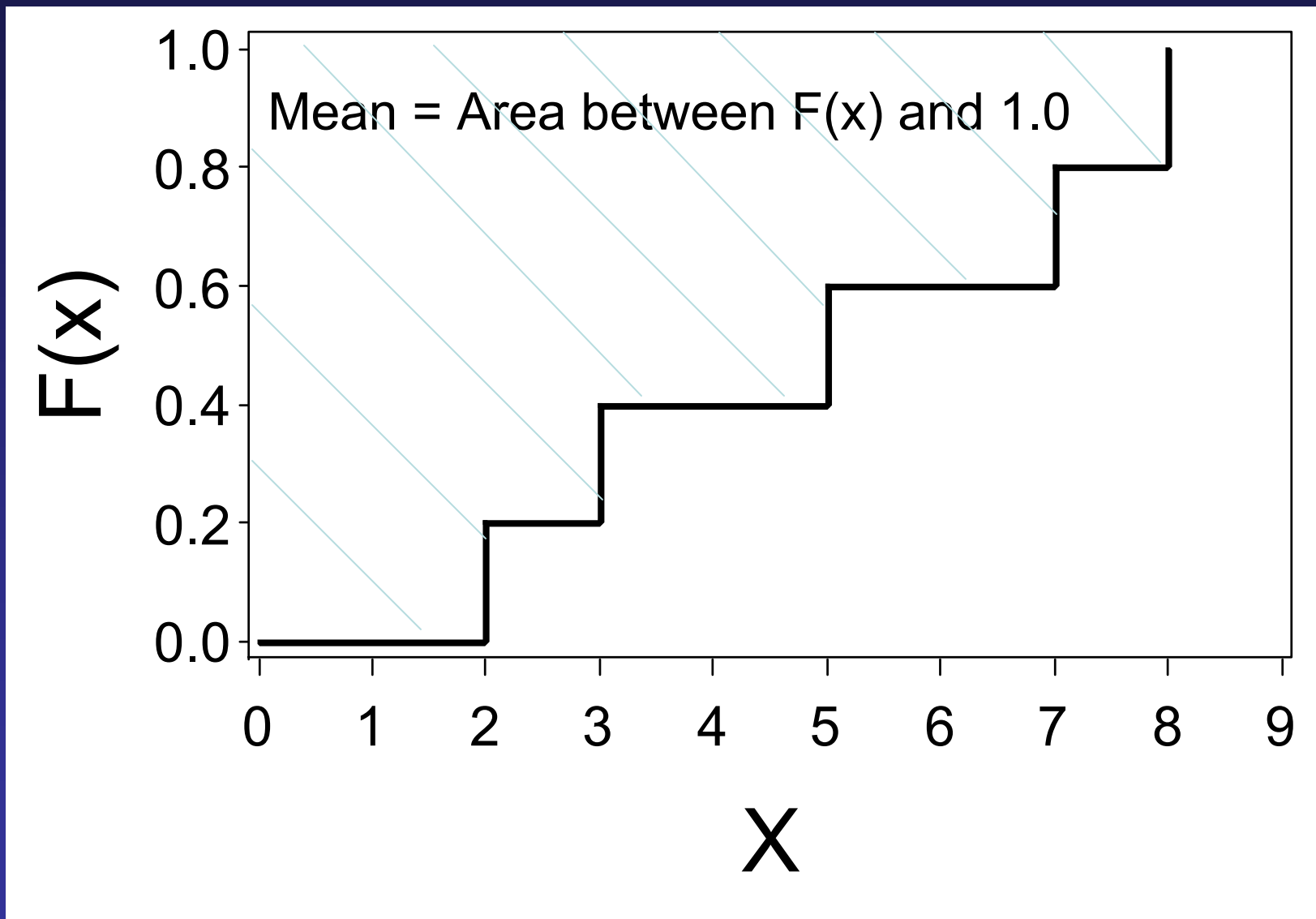
- $F(x)$ = probability that a concentration is less than or equal to x
- Example **without censoring**:
 - sample concentrations of 2, 3, 5, 7, 8

$$F(x) = \frac{\#(X \leq x)}{\text{total}}$$

$$F(4) = \frac{2}{5}$$

$F(4)$ = the proportion of sample values ≤ 4

F(x) for the sample data (2,3,5,7,8)



F(x) with Left-Censored Data

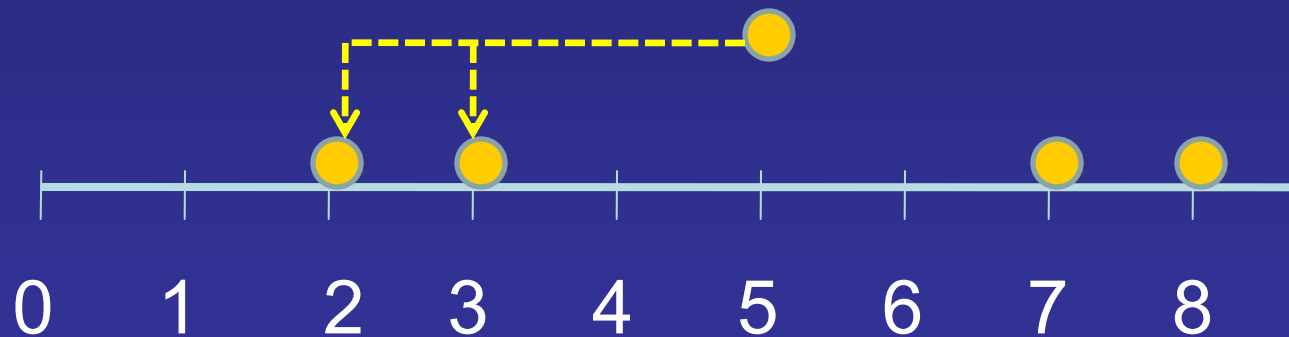
- Example with left censoring:
 - consider values of 2, 3, <5, 7, 8

$$F(4) = \frac{\#(X \leq 4)}{\text{total}} = ?$$

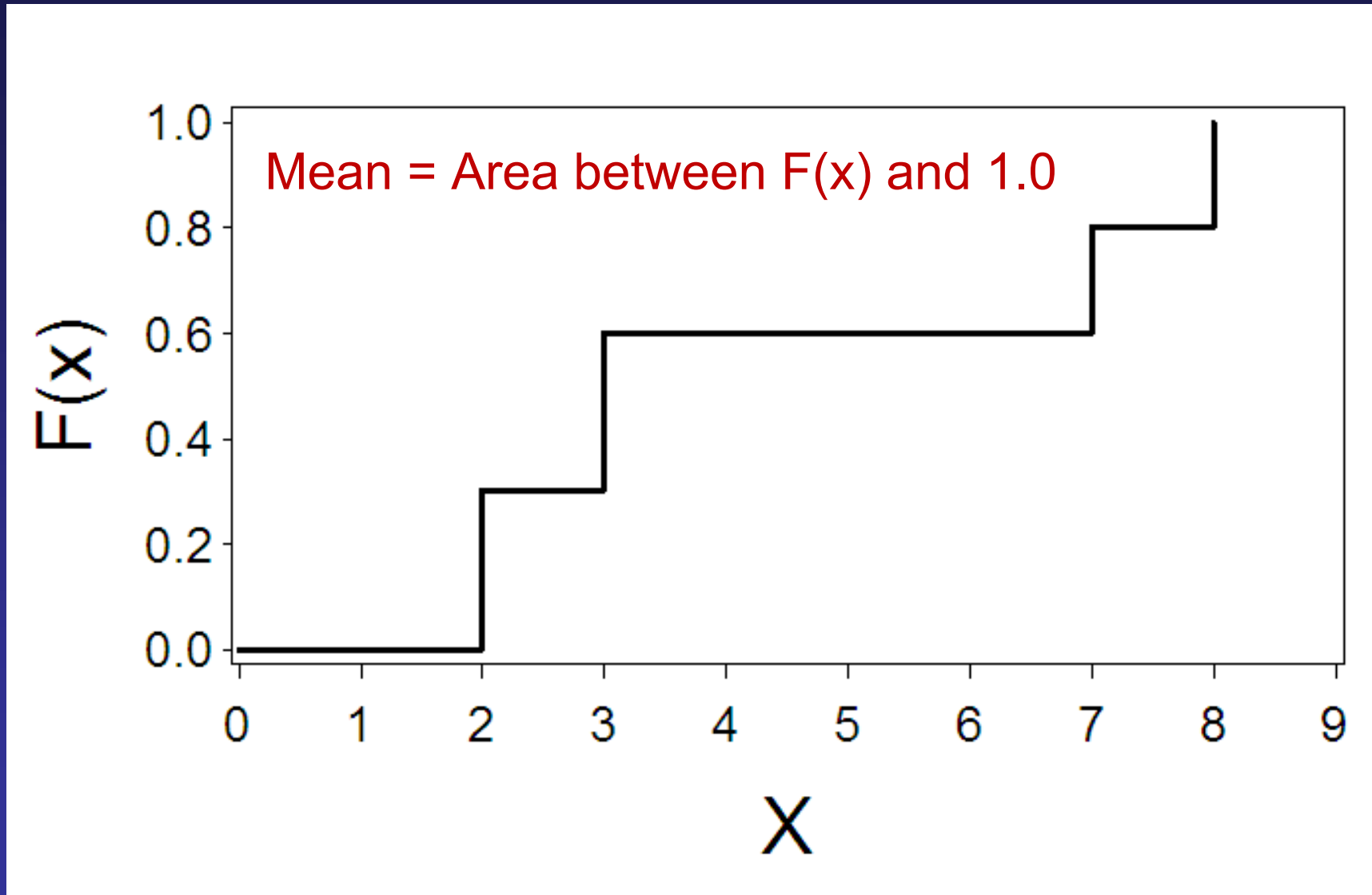
⇒ Use the Turnbull estimator

Options for “<5”:

- Assign 0
- Assign $5/\sqrt{2} \approx 3.54$
- Distribute the mass equally to all points <5 (nonparametric ML method, i.e., Turnbull)



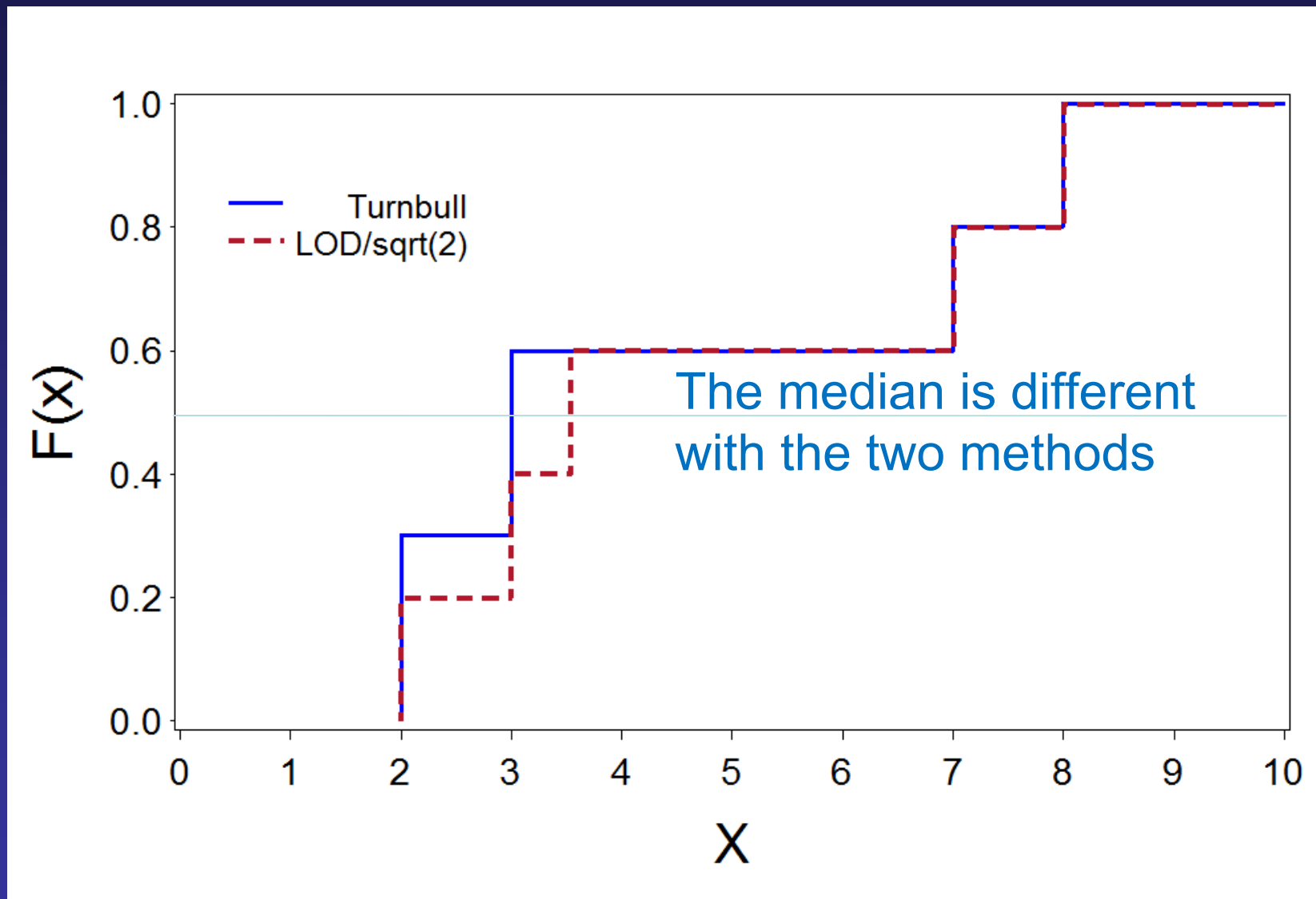
F(x) for the sample data (2,3,<5,7,8)



Turnbull : Redistribution to the Left

- The probability associated with each left-censored observation ($1/n$) is re-distributed equally to all observations to the left.
- This redistribution to the left assumes that the true value comes from the same distribution as the rest of the values to the left (i.e., no distributional assumptions – the data provide the distribution).

Comparison of Turnbull and LOD/ $\sqrt{2}$



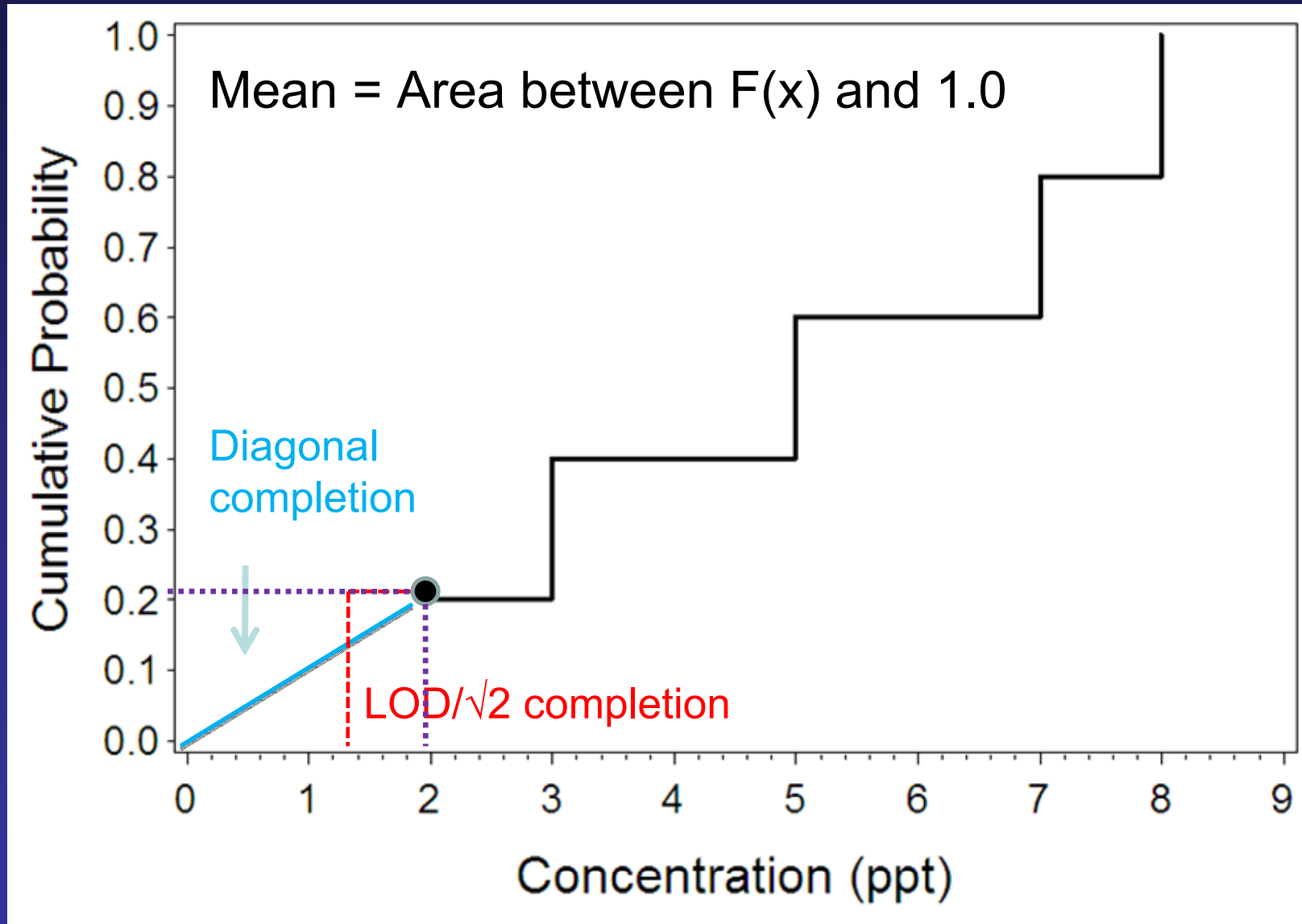
Comparison of Turnbull and $\text{LOD}/\sqrt{2}$

- Note that whereas the Turnbull estimator spreads the probability for a value below LOD evenly over the values to the left, using $\text{LOD}/\sqrt{2}$ deposits the probability at a single point.

F(x) with Left-Censored Data

- When the smallest value is $< \text{LOD}$, then there are no points to the left to redistribute onto.
- Example: $< 2, 3, 5, 7, 8$
- The Turnbull estimate is left “hanging” at this point. This appropriately reflects lack of knowledge in this region.
- “Completion” of $F(t)$ below the smallest LOD is arbitrary, but reasonable choices are possible.

F(x) for the sample data (<2,3, 5,7,8)



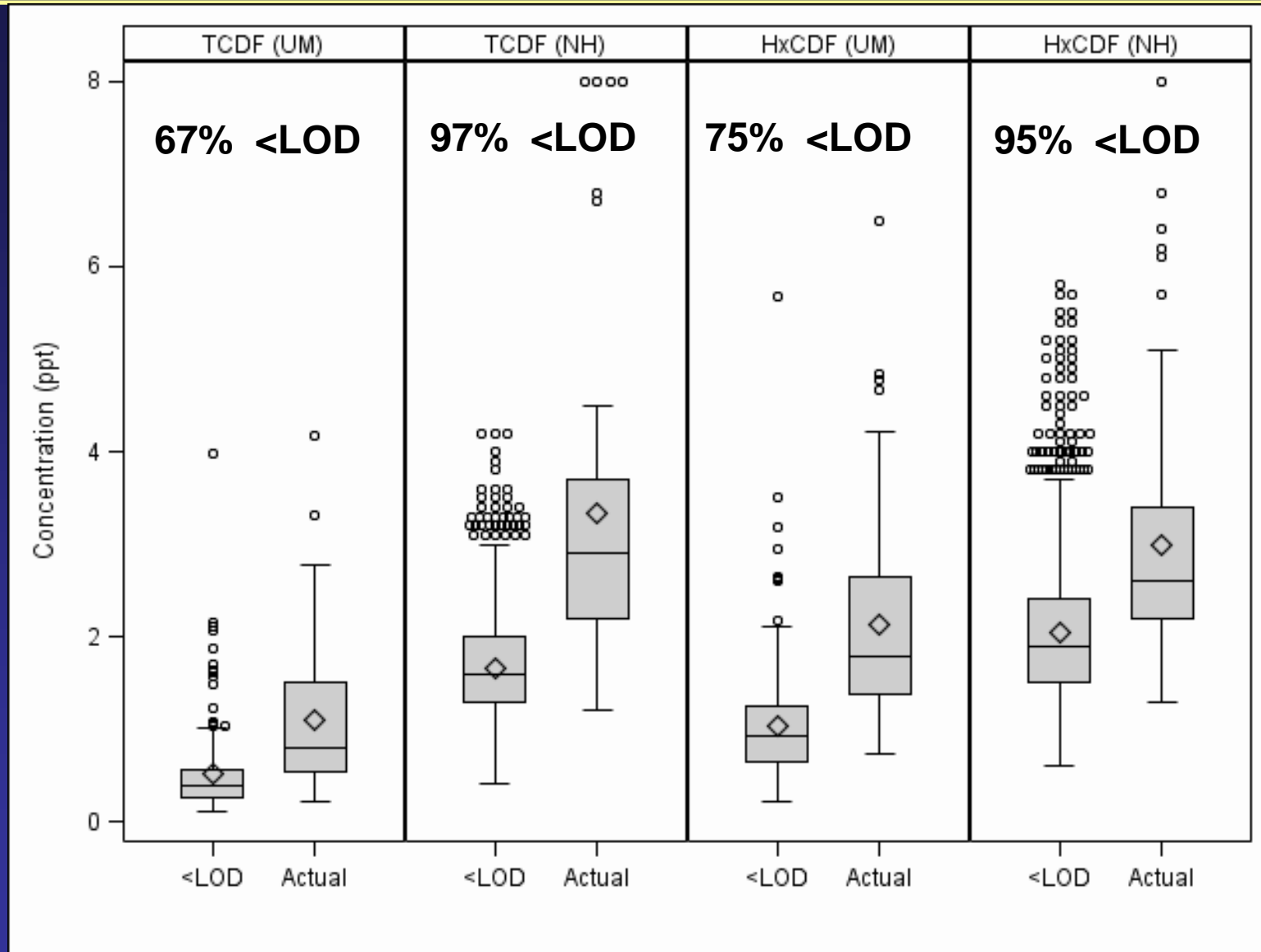
Examples from Serum Dioxin Data

Congener	Study	% Below LOD	Median LOD (ppt)
2,3,7,8 TCDD	UMDES	21%	0.5
1,2,3,4,7,8 HxCDD	UMDES	12%	2.6
TCDF	UMDES	67%	0.1
2,3,4,6,7,8 HxCDF	UMDES	75%	0.21
TCDF	NHANES	97%	0.4
2,3,4,6,7,8 HxCDF	NHANES	95%	0.60

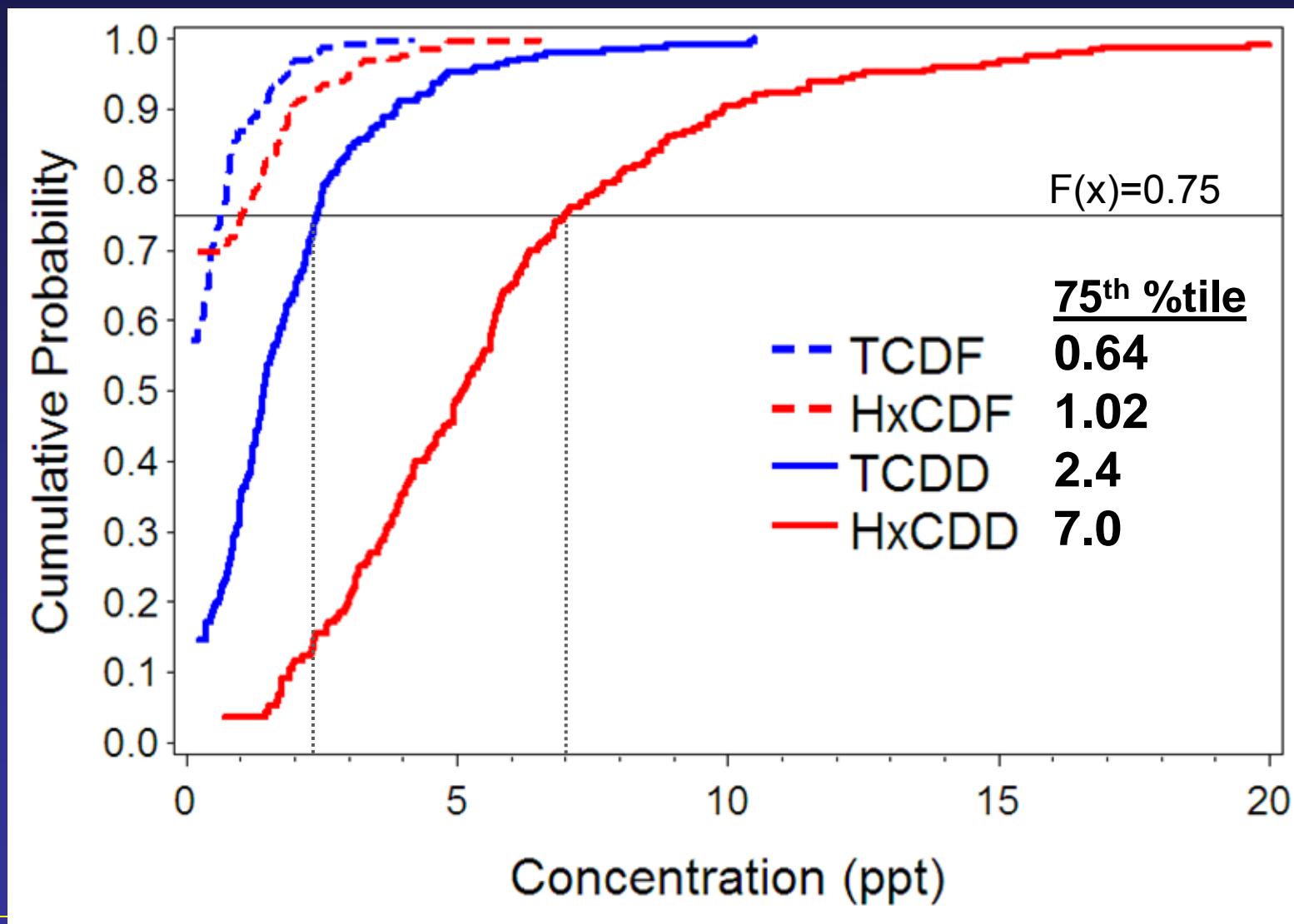
Distributions of LODs and Actual Values

- For serum, the LOD is a function of blood lipid weight, which can vary widely.
- The LOD distribution may be more variable than you think!

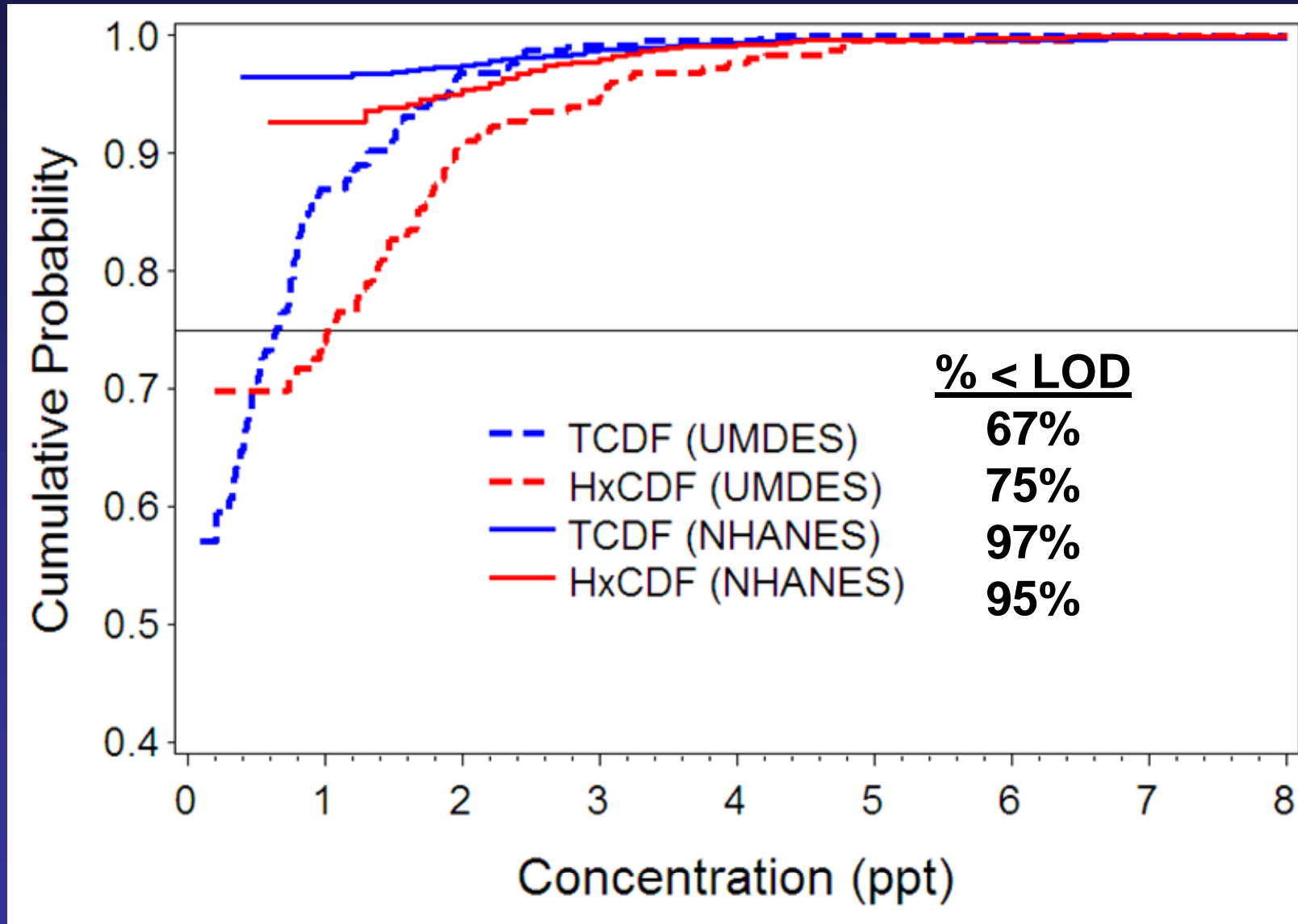
Boxplot Distributions of LODs and Observed Values



Turnbull Estimates for Four UMDES Congeners



Turnbull Estimates for TCDF and HxCDF



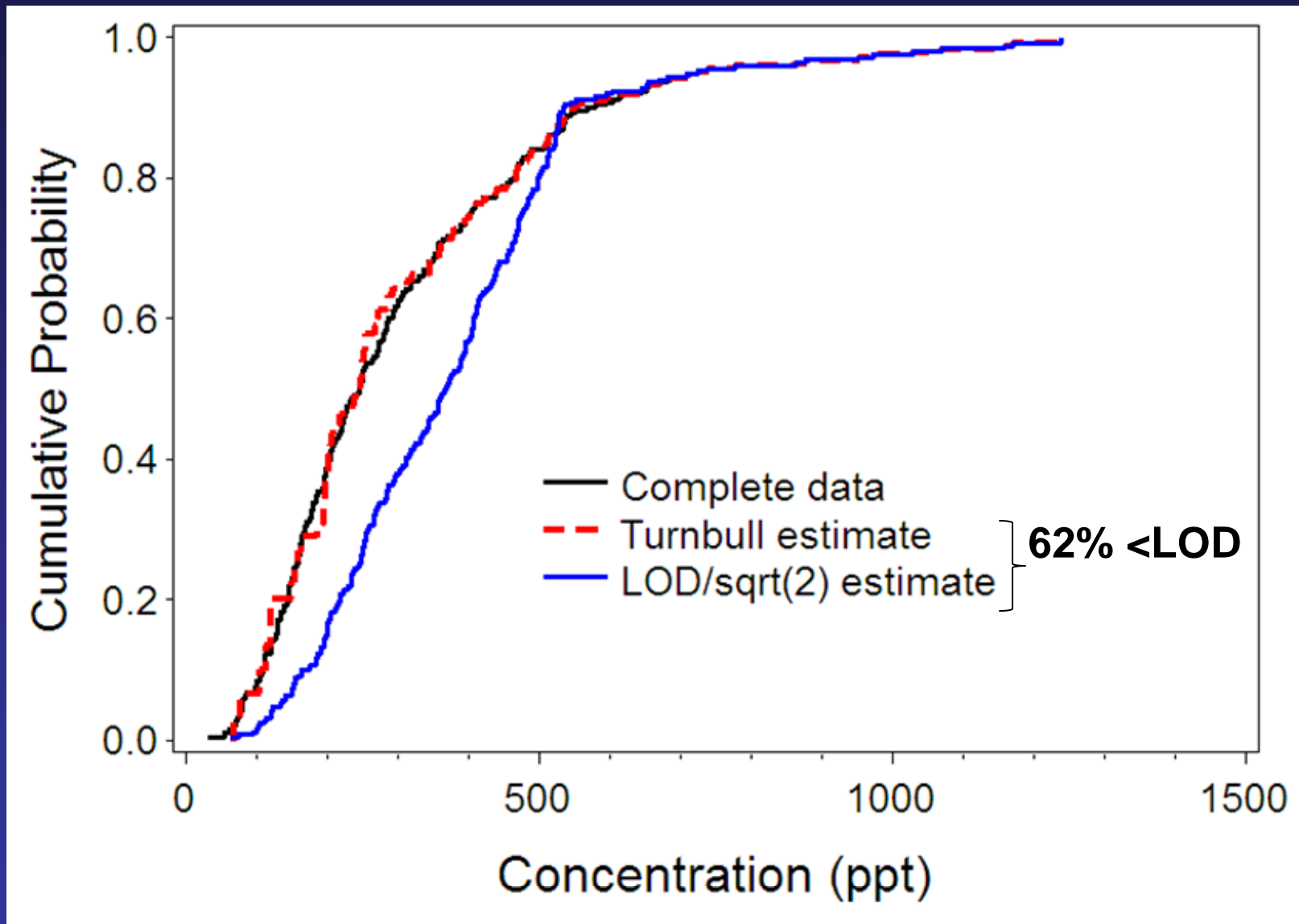
Median, 75th percentile, and mean using Turnbull and LOD/ $\sqrt{2}$

	Method	TCDF		HxCDF	
		UMDES	NHANES	UMDES	NHANES
Median	Turnbull	<0.1	<0.4	<0.2	<0.6
	LOD/ $\sqrt{2}$	0.4	1.1	0.8	1.3
75th Percentile	Turnbull	0.6	<0.4	1.0	<0.6
	LOD/ $\sqrt{2}$	0.7	1.4	1.3	1.8
Mean					
Lower bound	Turnbull	0.41	0.11	0.60	0.19
Upper bound	Turnbull	0.47	0.50	0.74	0.75
	LOD/ $\sqrt{2}$	0.61	1.24	1.09	1.52

Complete Data vs Turnbull vs LOD/ $\sqrt{2}$

- We started with OCDD data, which was complete.
- We randomly generated an LOD value for each concentration
- For values below their generated LOD, only the LOD was kept.
- This resulted in 62% of values below LOD.
- We calculated $F(x)$ for
 - The complete data
 - The censored data, using both the Turnbull estimator, and replacing true values with $\text{LOD}/\sqrt{2}$

The Turnbull estimate beats LOD/sqrt(2)

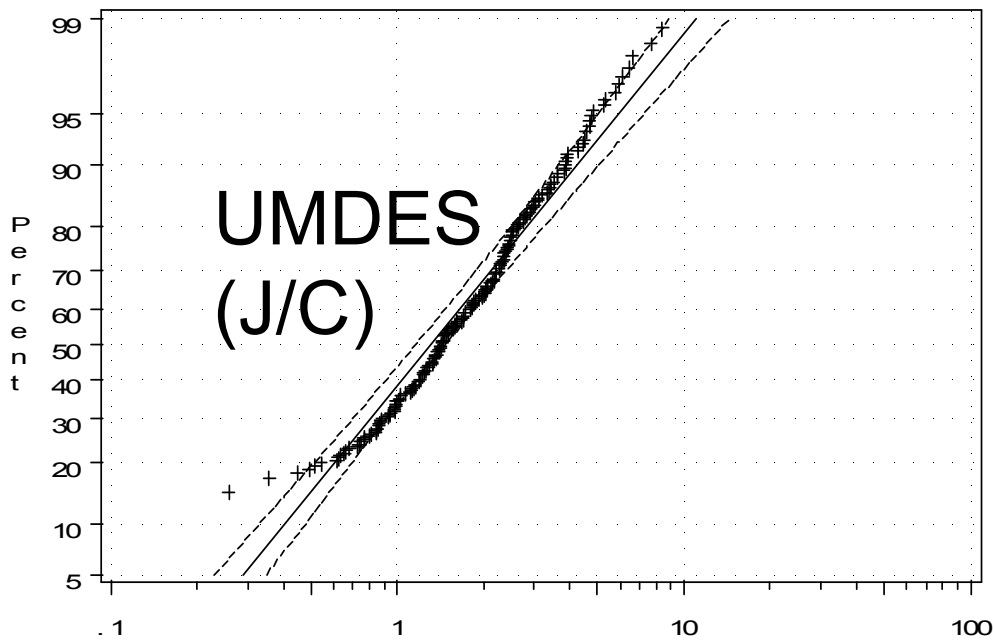


Why not Fit a Lognormal Distribution?

- For population distributions, the lognormal distribution may not fit well without adjusting for exposures.
- A *nonparametric* estimator makes no distributional assumptions.
- A Q-Q plot can be used to check the fit of a lognormal distribution.

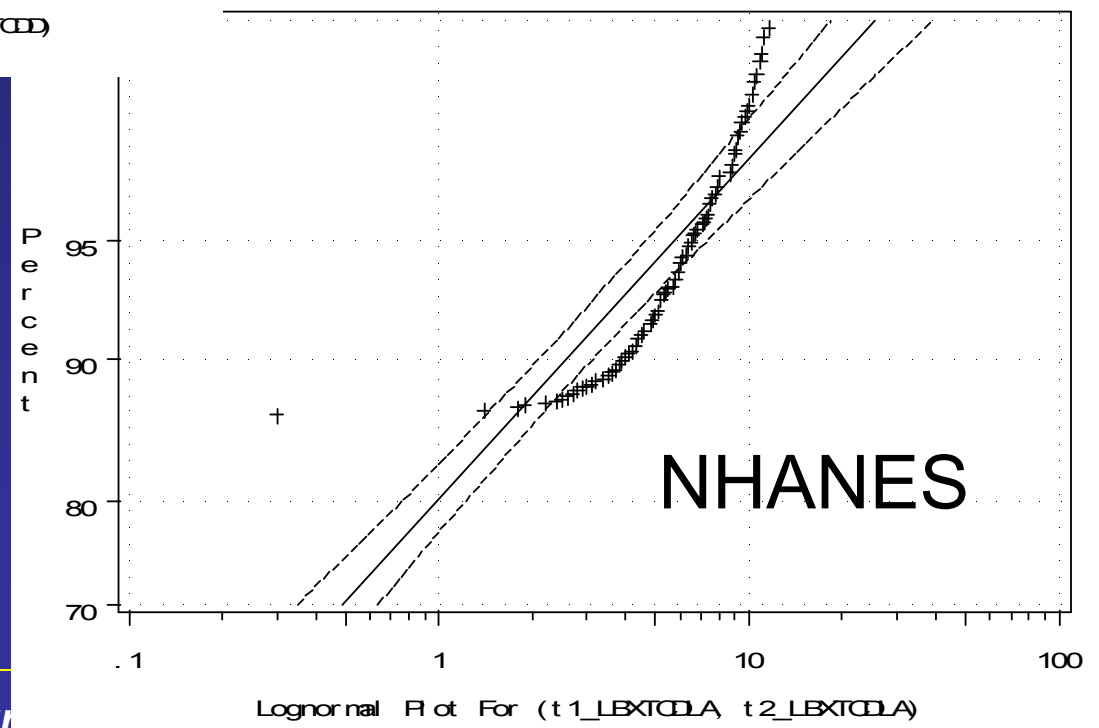
What is a Q-Q Plot?

- A Q-Q plot graphs the quantiles of the estimated lognormal distribution versus the quantiles of a nonparametric distribution estimator
- If the lognormal distribution fits, the points will fall on the diagonal line.
- The following slide gives Q-Q plots of the best fit lognormal distribution versus the nonparametric Turnbull estimate for TCDD.
 - Data are given for both UMDES (J/C) and NHANES



Q-Q Plots for TCDD:
Lognormal vs. Turnbull.
Lack of fit is seen, esp.
for NHANES.

University of Michigan Dioxin Exposure



Software

- JMP (SAS product)
- SAS (Proc Lifereg)
- R - Excel with Turnbull code
- Any software with a function for a Kaplan-Meier estimator (SAS, SPSS, Stata, Minitab)
 - Reverse the time scale (subtract all values from the maximum+1) (Left-censored values become right-censored)
 - Run the KM function or procedure on the reversed values
 - To plot $F(t)$, apply the survival estimates to the original values.
 - Be careful where the “steps” are (the point is right-justified on the step).

Conclusions (1)

- The Turnbull estimator is the appropriate nonparametric estimator for estimating population percentiles for data with values below LOD.
- It has excellent statistical credentials:
 - nonparametric maximum likelihood estimator
 - it is the estimator recommended in statistics texts

Conclusions (2)

- Its use has been constrained by
 - Limited software availability
 - Lack of understanding of its advantages
 - Lack of understanding of the “undefined” area, and use of less desirable completion methods
- We hope that this introduction will help increase use of the Turnbull estimator to estimate percentiles when some data are below the LOD.

Thank you!