

# Data Analysis Methods when Some Data are Below a Limit of Detection

**Brenda W Gillespie, Ph.D.**

Dept of Biostatistics, School of Public Health, and  
Center for Statistical Consultation & Research  
(CSCAR)

The University of Michigan, Ann Arbor, MI, USA

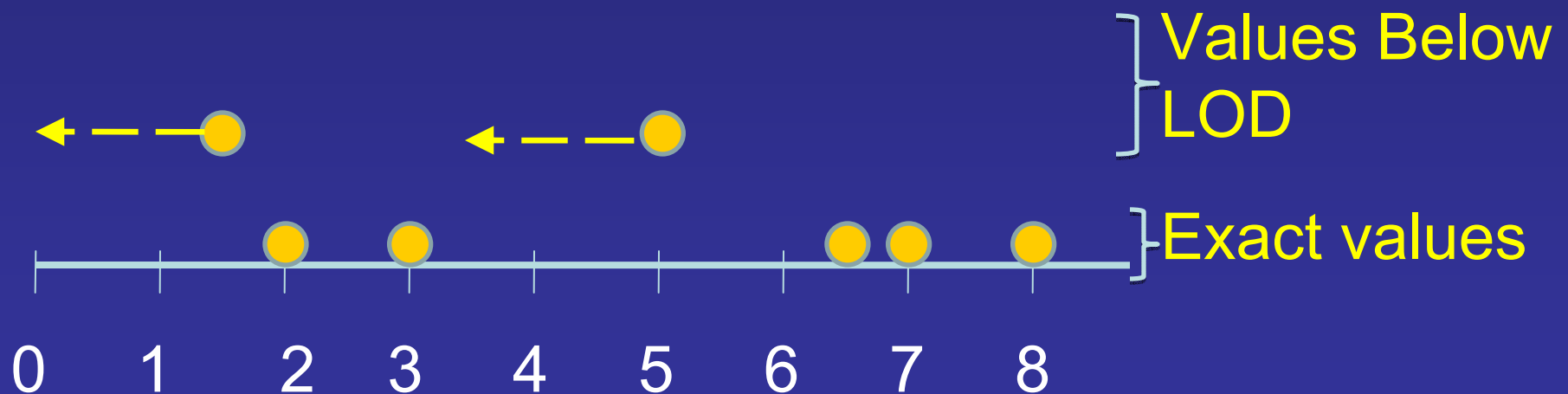
*May 19, 2011*

# Data Below the Limit of Detection (LOD)

## Terms:

- non-detects,
- values below LOD, or
- left-censored observations.

Data: <1.5, 2, 3, <5, 6.5, 7, 8



# Variability in LOD values

---

- In some datasets, there is a common LOD, such as the detection limit of the measuring instrument.
  - A bathroom scale that is not accurate below 10 lbs.
  - Measurements of blood lead in the NHANES\* 3 data
- Sometimes the LOD varies
  - as a function of the sample volume
  - as a function of adjustment values
    - Dioxin is almost exclusively found in lipids, so blood samples must be adjusted for blood lipid weight (e.g., in the NHANES 3 blood dioxin data)
- \*NHANES = U.S. National Health and Nutrition Examination Survey

# Right and Left Censoring

---

- In biostatistics, right and left censoring commonly occur.
- Right censoring: The true value is larger than some known value.
  - E.g., time to death after disease diagnosis. If a person is still alive after 3.4 years, they are right-censored at that time.
- Left censoring: The true value is smaller than some known value.
  - E.g., values below a limit of detection (dioxin, HCV infection levels, age at learning to read, time to HIV infection).

# Contrasting Right and Left Censoring

- Right censoring (e.g., time to death):



- Left censoring (e.g., lab value or time to HIV):



# Dealing with the LOD

---

- Common methods of handling values below LOD:
  - Ad hoc methods:
    - Assign to zero, LOD/2 or  $\text{LOD}/\sqrt{2}$
  - Methods with statistical justification:
    - Maximum likelihood (ML) methods (e.g., lognormal)
    - Nonparametric methods (e.g., Turnbull)

# Reporting LOD Information

---

In papers or presentations, the following should always be given (but rarely are):

- The number (%) of values below LOD
- The range and the mean or median of LOD values
- The method of handling values below LOD in data analyses

# Outline of topics covered

---

- 1) Below LOD data as left-censored data, a special case of censored data [done]
- 2) Estimating the distribution function (and population percentiles) with left-censored data
- 3) Two-sample tests with left-censored data
- 4) Regression with left-censored data
  - 1) Parametric (e.g., Weibull, lognormal)
  - 2) Semi-parametric (e.g., reverse Cox) (not covered)

# Software for left-censored data

---

- Estimating the distribution function
  - SAS Proc Lifereg (Turnbull estimate)
  - SAS Proc Lifetest with reversed data + adjustment
  - R – reverse KM procedure (NADA library)
- Two-sample or k-sample tests
  - SAS Proc Lifetest (with reversed data)
  - R (NADA, or logrank and Wilcoxon tests with reversed data)

# Software for left-censored data

---

- Regression with left-censored data
  - SAS Proc Lifereg (parametric regression for left-cens data)
  - SAS Proc phreg (Cox regression with reversed data)
  
  - R - Parametric regression for left-censored data (NADA)
  - R - Cox regression with reversed data



Estimating the  
cumulative distribution function  
with left-censored data  
using the  
Reverse Kaplan-Meier  
(or Turnbull) Estimator

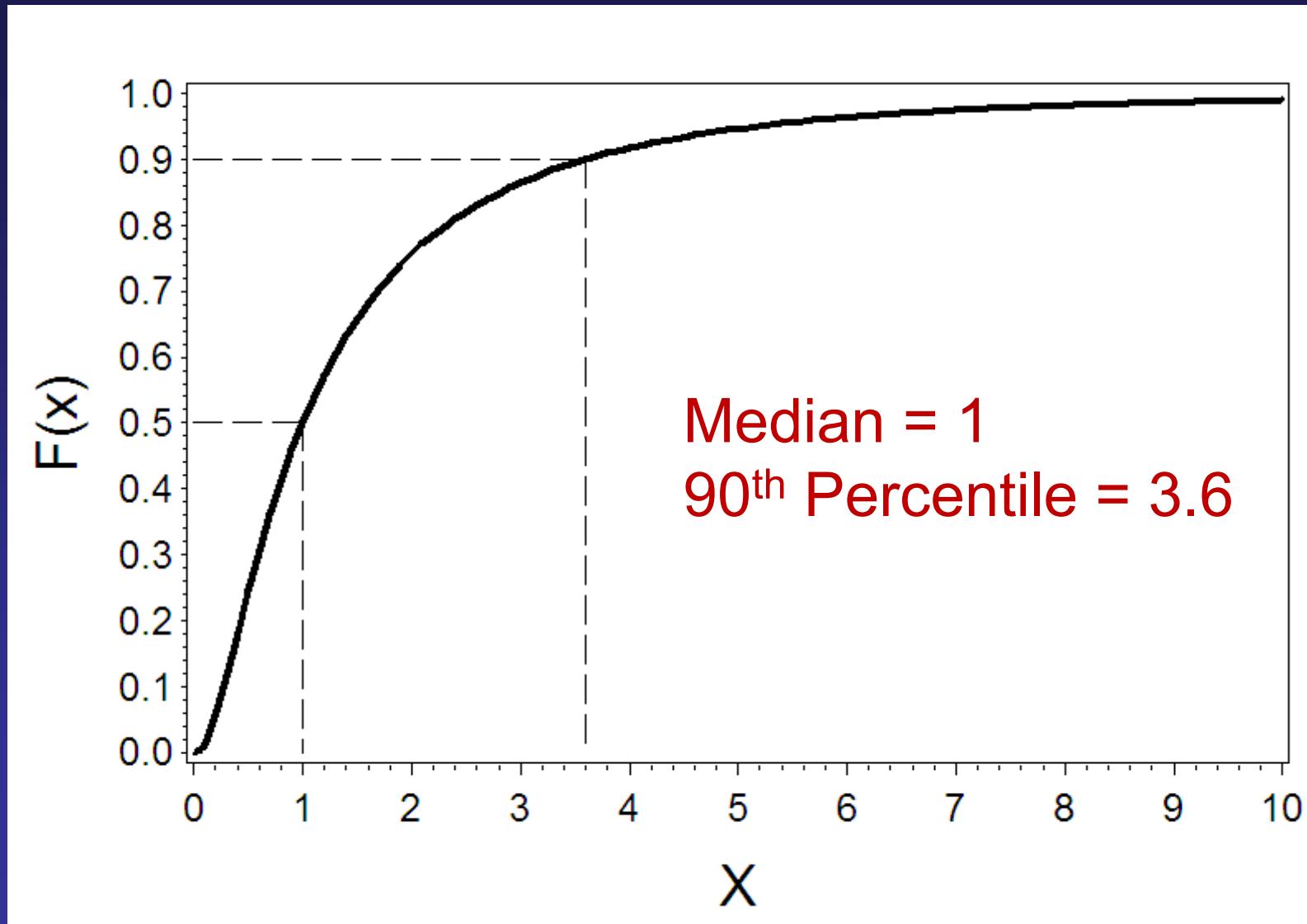
# Background

---

Estimating population percentiles is often a goal in reporting contamination levels in a population.

- E.g., a population median or 95<sup>th</sup> percentile of a toxin level.

# Example Cumulative Distribution Function



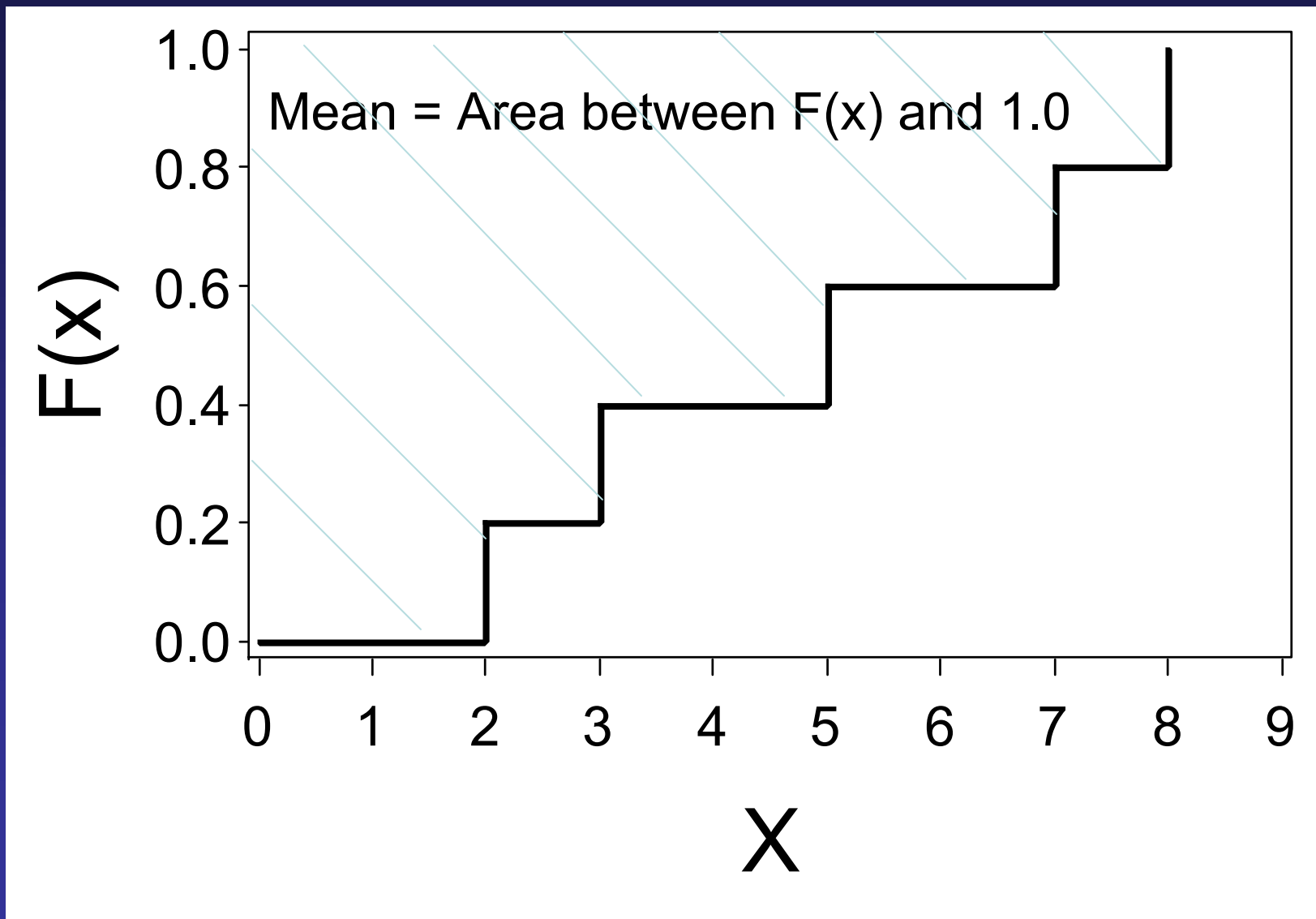
# Estimating F(x) with Complete Data

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

$$F(x) = \frac{\#(X \leq x)}{n}$$
$$F(4) = \frac{2}{5}$$

$F(4)$  = the proportion of sample values  $\leq 4$

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



# Estimating $F(x)$ with **left-censored data**

- Introduce 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

# F(x) with Left-Censored Data

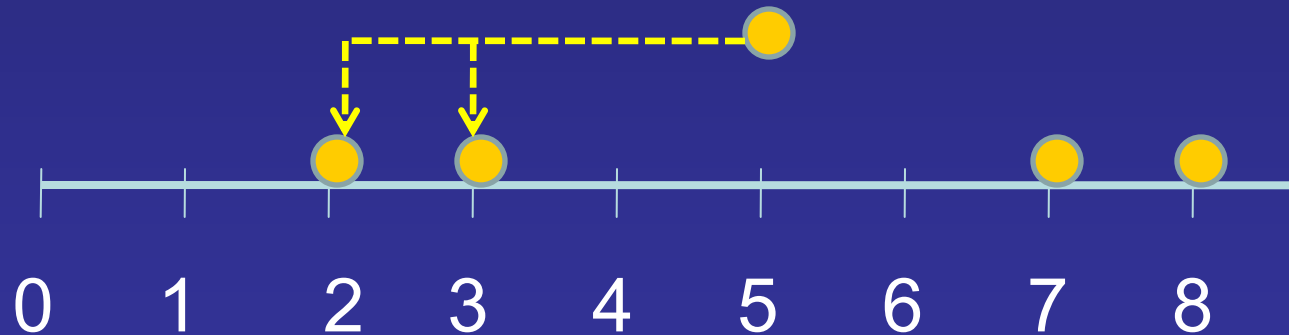
- Example with **left censoring** (n=5):
  - consider values of 2, 3, <5, 7, 8

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

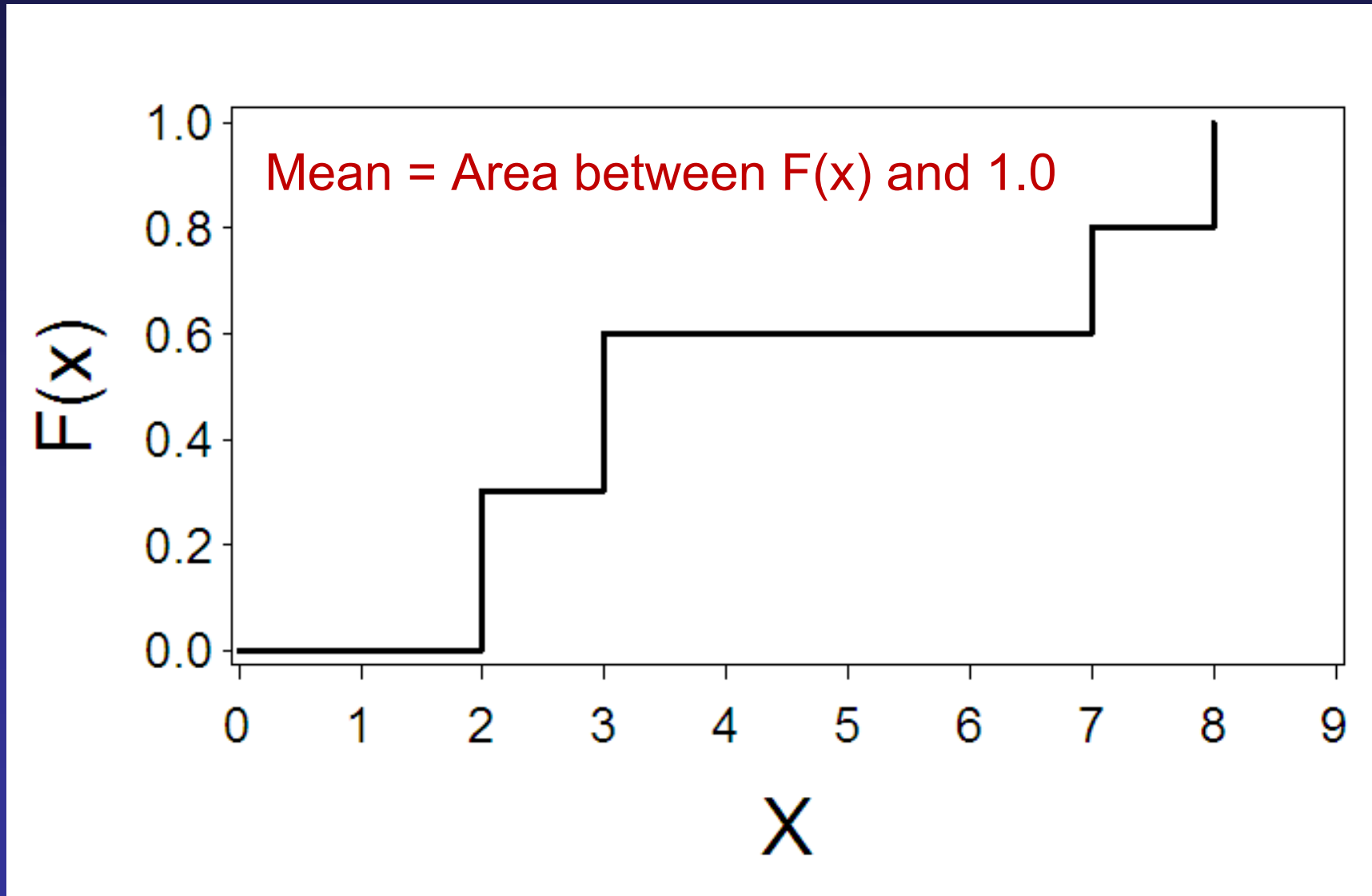
⇒ Use the Turnbull (reverse Kaplan-Meier) estimator

# Options for “<5” in data (2, 3, <5, 7, 8 ):

- 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)

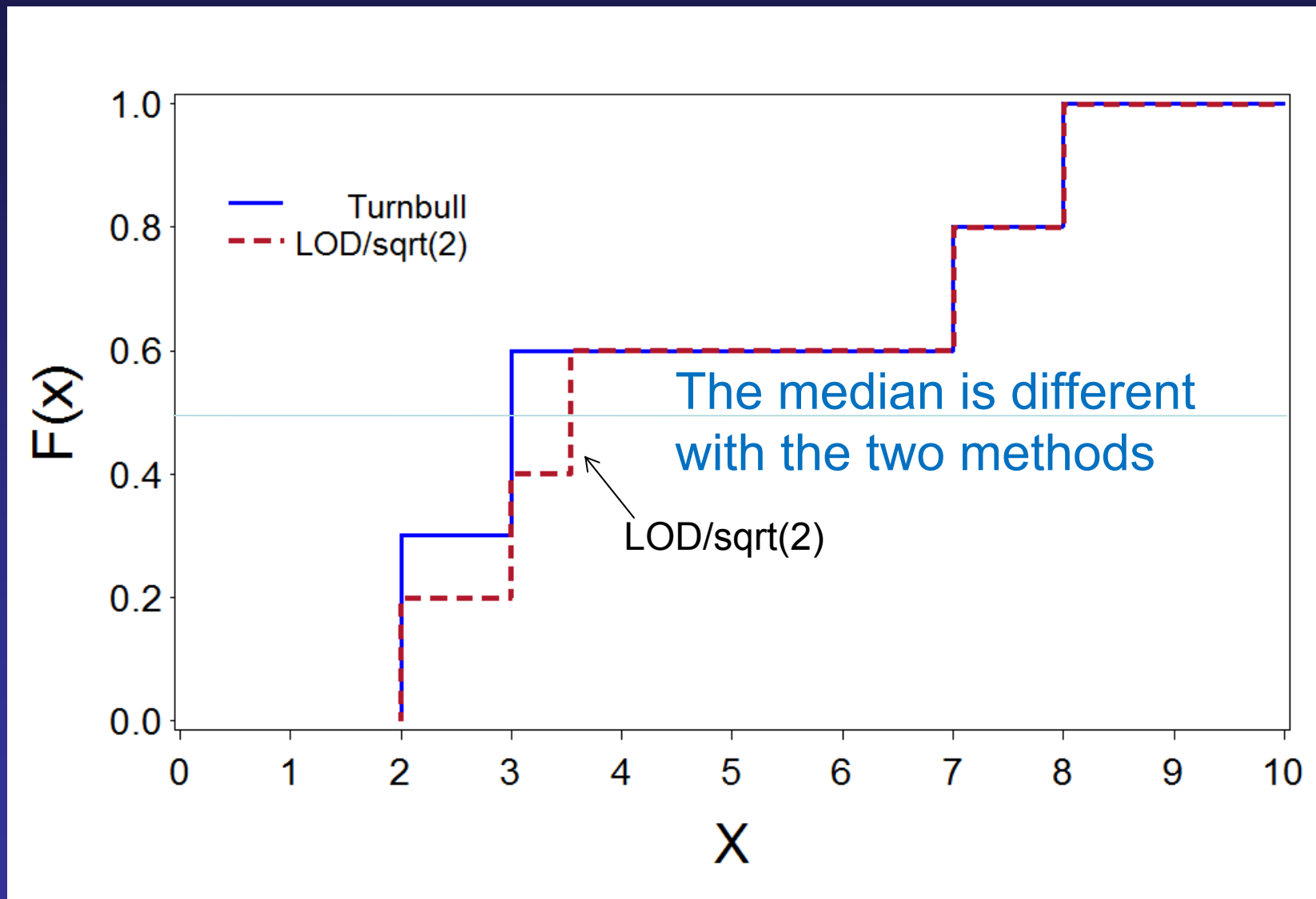


# Turnbull : Redistribution to the Left

---

- 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 $\text{LOD}/\sqrt{2}$

---

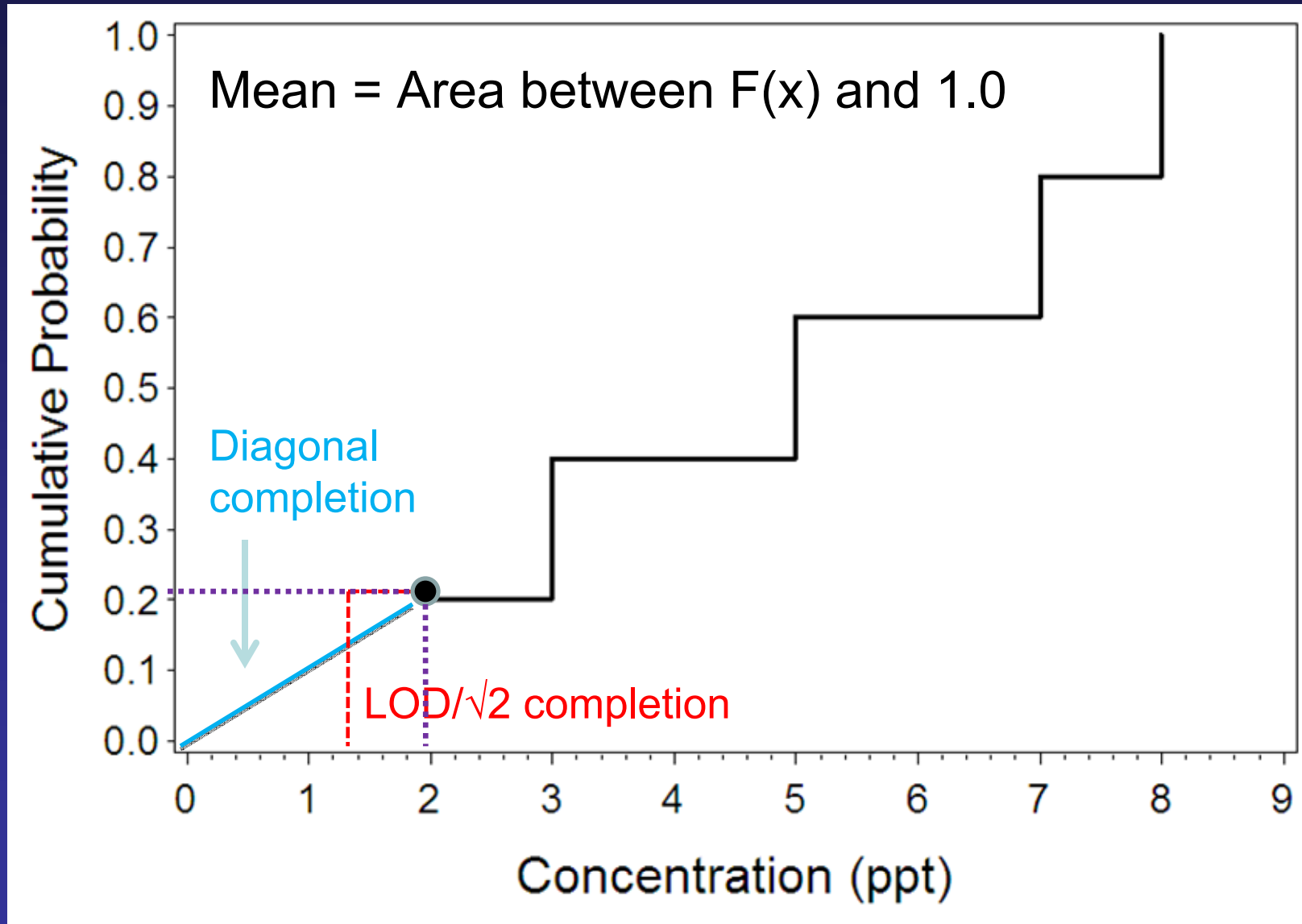
- Note that whereas the Turnbull estimator spreads the probability for a value below LOD evenly over the values to the left, using  $\text{LOD}/\sqrt{2}$  deposits the probability at a single point.

# 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

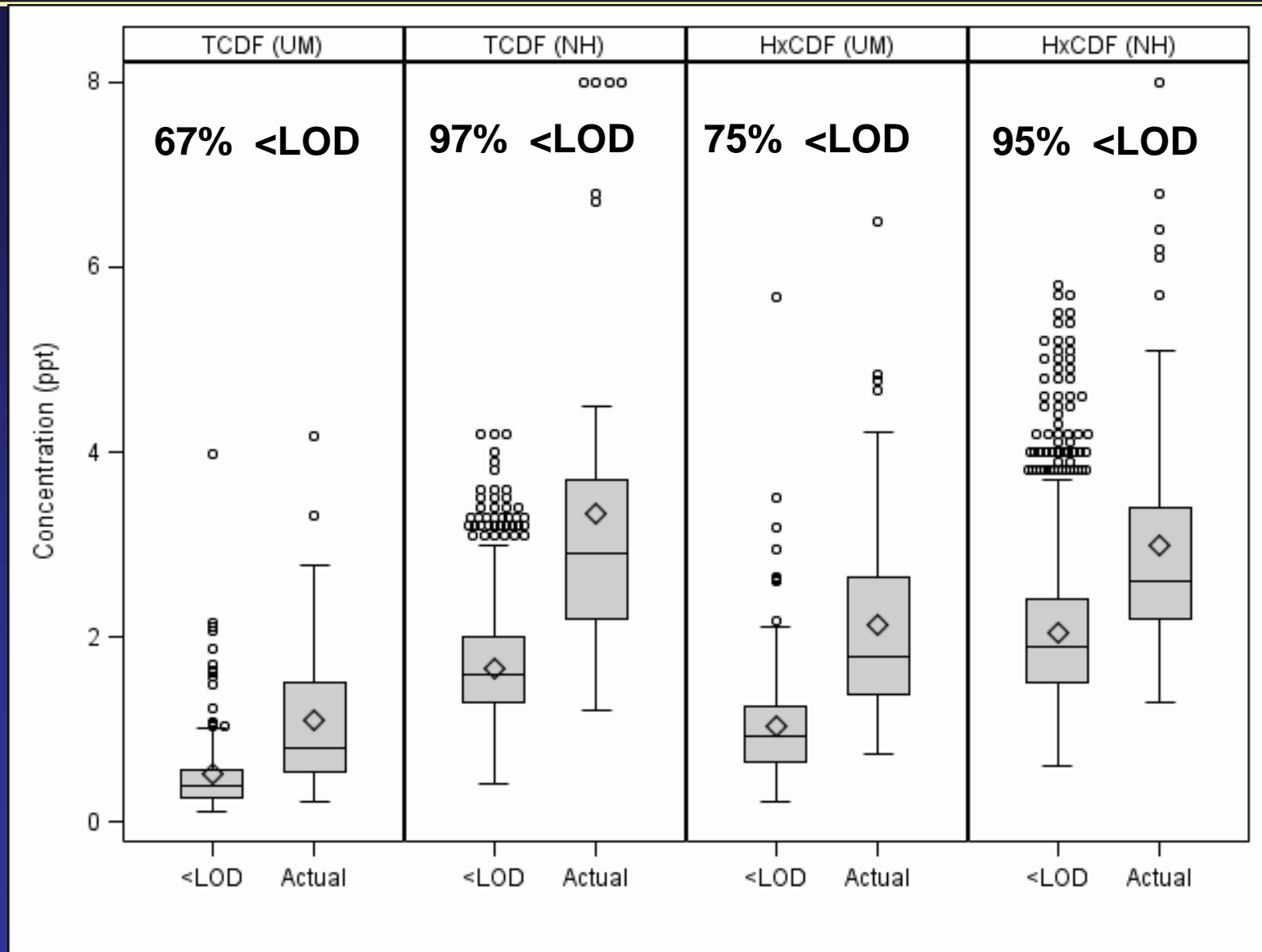
Congener	Study	% Below LOD	Median LOD (ppt)
<b>2,3,7,8 TCDD</b>	UMDES	21%	0.5
<b>1,2,3,4,7,8 HxCDD</b>	UMDES	12%	2.6
<b>TCDF</b>	UMDES	67%	0.1
<b>2,3,4,6,7,8 HxCDF</b>	UMDES	75%	0.21
<b>TCDF</b>	NHANES	97%	0.4
<b>2,3,4,6,7,8 HxCDF</b>	NHANES	95%	0.60

# Distributions of LODs and Actual Values

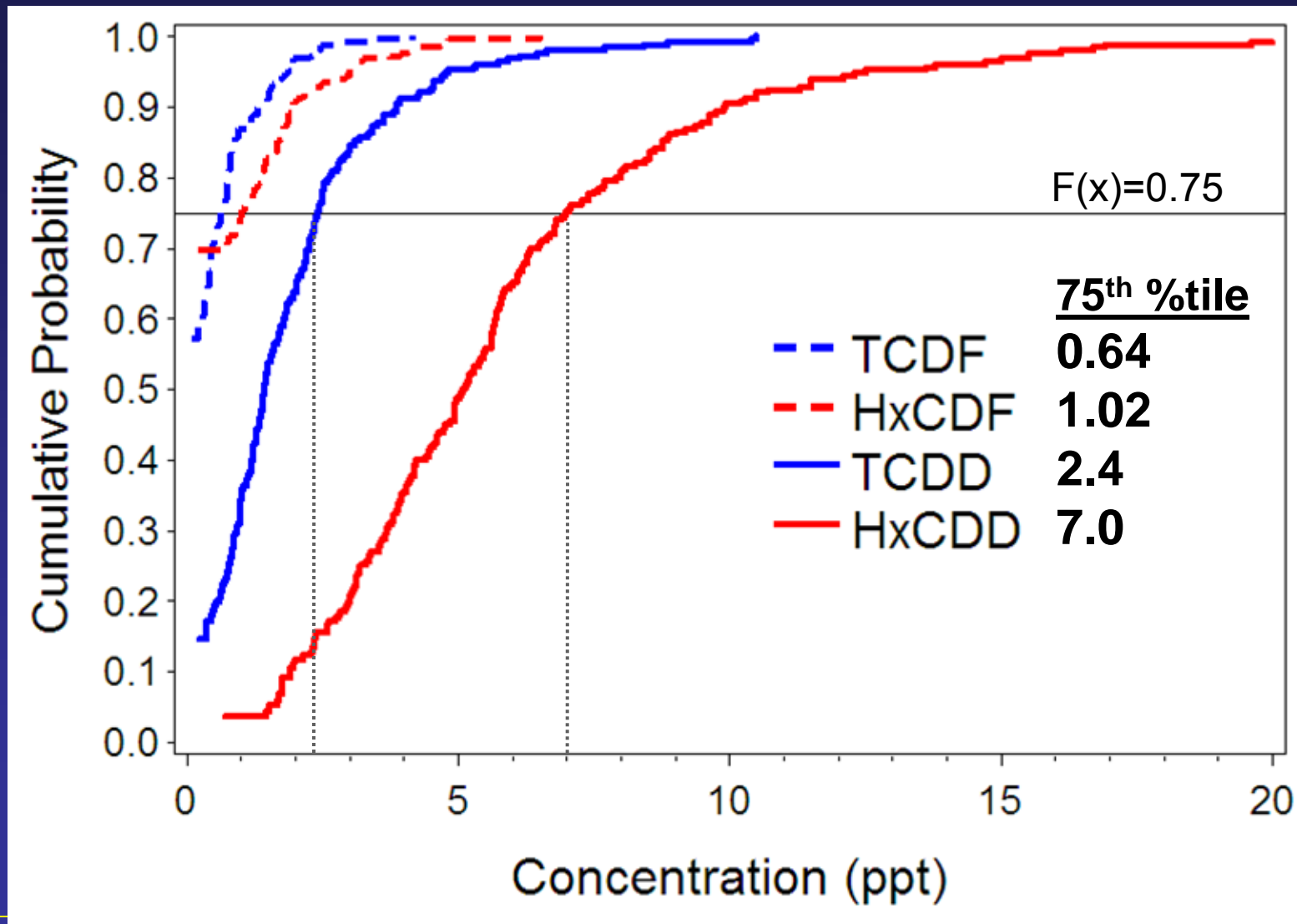
---

- 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!

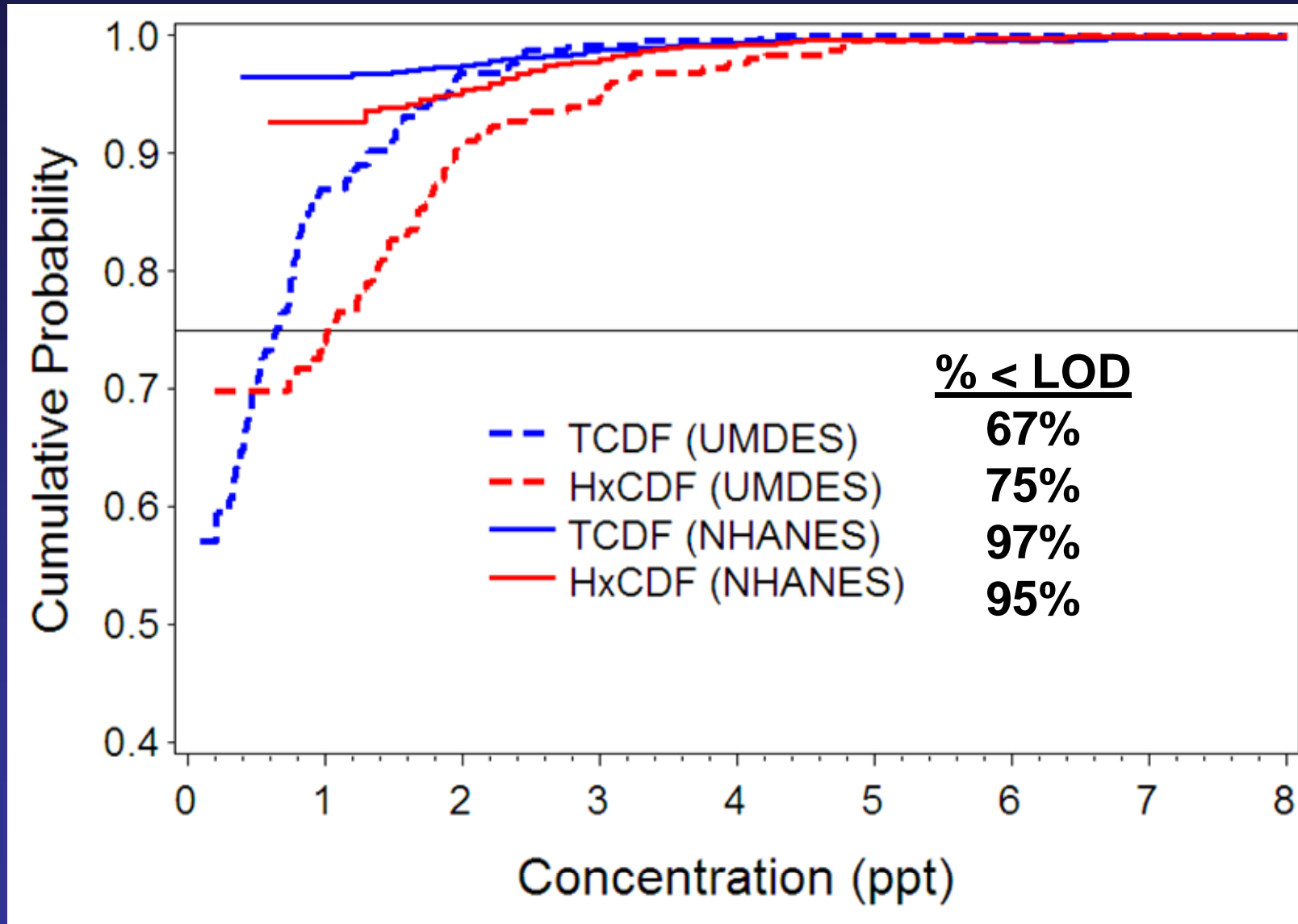
# Boxplot Distributions of LODs and Observed Values



# Turnbull Estimates for Four UMDES Congeners



# Turnbull Estimates for TCDF and HxCDF



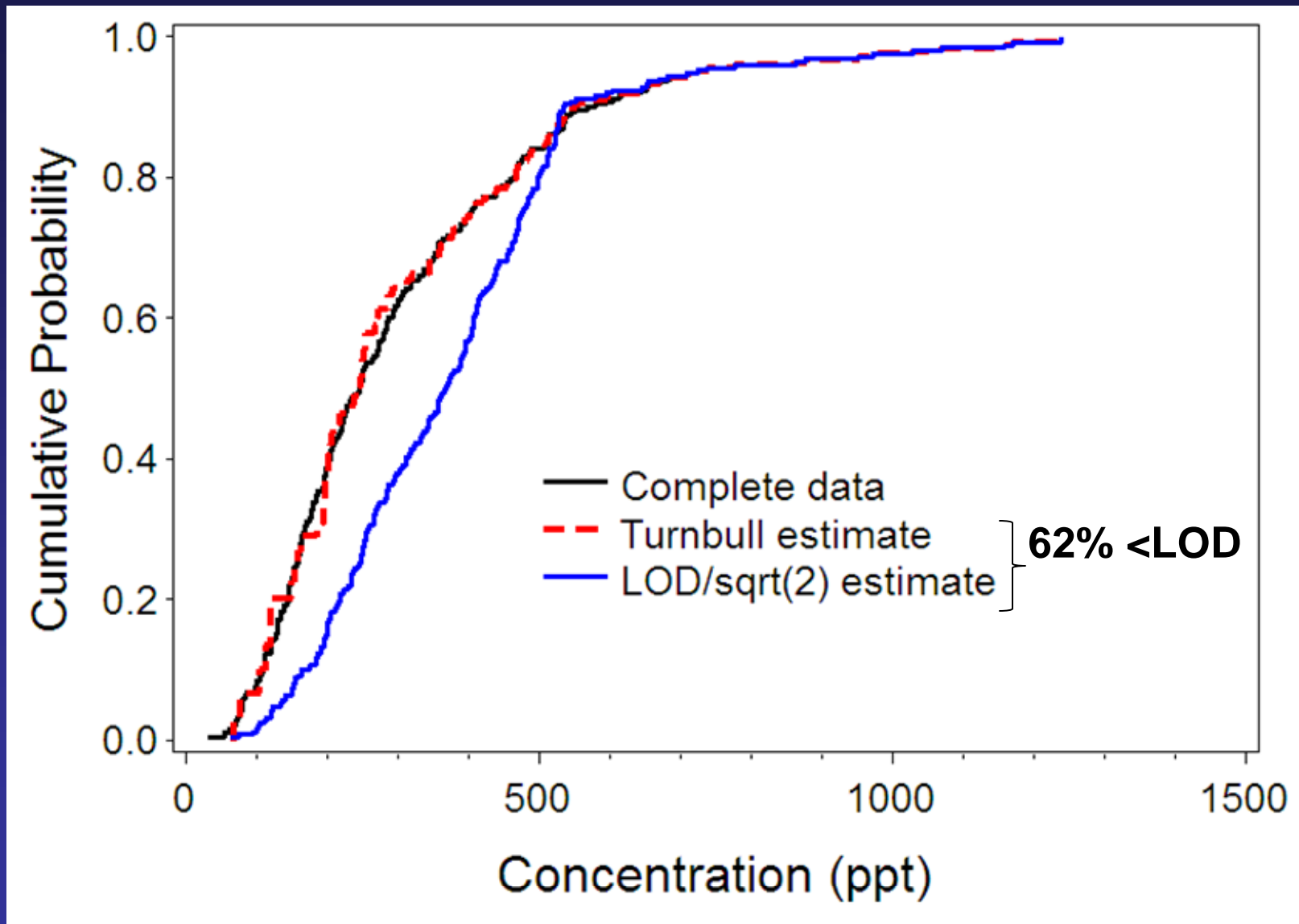
# Median, 75th percentile, and mean using Turnbull and LOD/ $\sqrt{2}$

	Method	TCDF		HxCDF	
		UMDES	NHANES	UMDES	NHANES
<b>Median</b>	Turnbull	<0.1	<0.4	<0.2	<0.6
	LOD/ $\sqrt{2}$	0.4	1.1	0.8	1.3
<b>75th Percentile</b>	Turnbull	0.6	<0.4	1.0	<0.6
	LOD/ $\sqrt{2}$	0.7	1.4	1.3	1.8
<b>Mean</b>					
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
    - the Turnbull estimator
    - replacing true values with  $\text{LOD}/\sqrt{2}$

# The Turnbull estimate beats LOD/sqrt(2)



## Why not Fit a Lognormal Distribution?

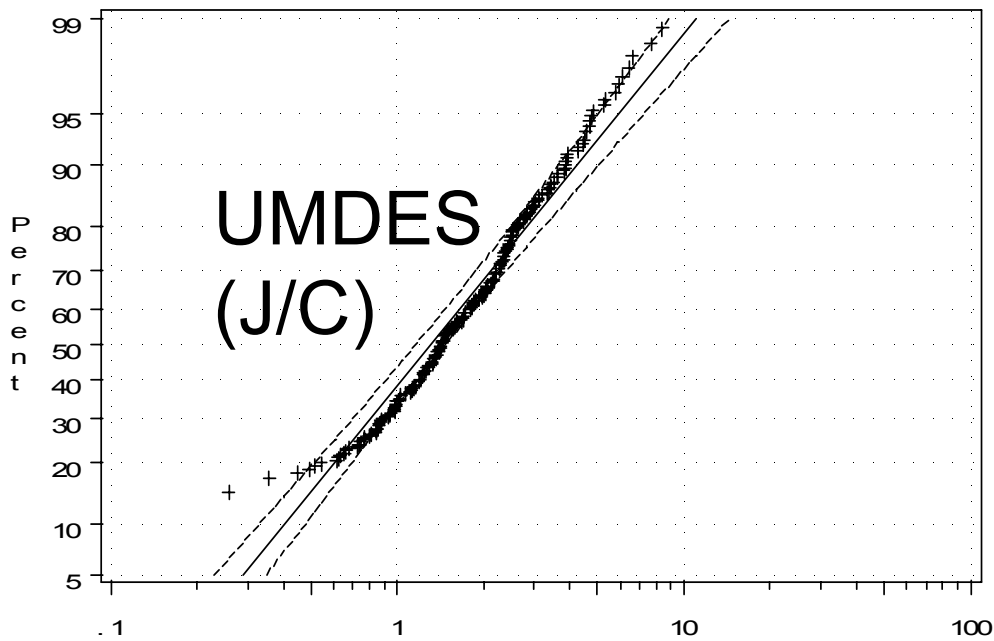
---

- For population distributions, the lognormal distribution may not fit well without adjusting for exposure variables (e.g., age, or diet).
- A *nonparametric* estimator makes no distributional assumptions.
- A Q-Q plot can be used to check the fit of a lognormal distribution.

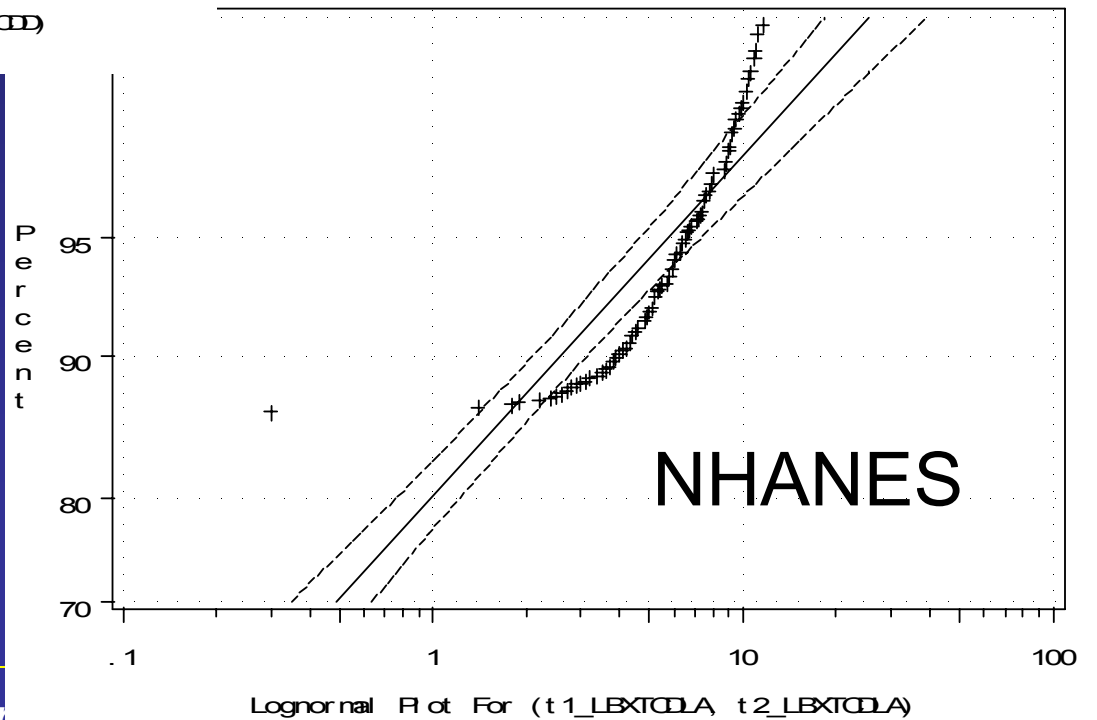
# What is a Q-Q Plot?

---

- 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.



# Software

---

- JMP (SAS product)
- SAS (Proc Lifereg)
- R – NADA package
- Any software with a function for a Kaplan-Meier estimator (SAS Proc Lifetest, 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). Can be tricky to plot correctly.

# Conclusions (1)

---

- The Reverse KM (or 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 Reverse-KM (Turnbull) estimator to estimate percentiles when some data are below the LOD.

# What if the concentration is a covariate?

- If the concentration is a covariate in a regression model, but some values are below LOD:
  - Substitute  $\text{LOD}/\sqrt{2}$  for values below LOD, or
  - Use multiple imputation (e.g., draw 5 values for each value below LOD):
    - Draw random values, using the estimated (Turnbull) distribution, conditional on being below the LOD for that observation.
    - If the smallest value is below LOD, a completion method must be specified.
  - All subsequent analyses must be performed for each imputation set, and the results pooled.

---

# Two-Sample or k-Sample Tests

# Two-Sample or k-Sample Tests

---

STATISTICS IN MEDICINE

*Statist. Med.* 2009; **28:700–715**

Nonparametric methods for measurements  
below detection limit

Donghui Zhang, *Chunpeng Fan, Juan Zhang and  
Cun-Hui Zhang*

*(see excerpt on next slide)*

## 2.3. Nonparametric methods (p. 703)

Traditional nonparametric methods for censored data such as the logrank, the Gehan, and the Peto-Peto tests are applicable for right-censored data. We can use the so-called 'flipping' technique to transform the left-censored  $Y_i$  into right-censored data. To do so, we first find a constant  $M > \max_i \{Y_i\}$ . The flipped observations are then constructed using the transformation

$$\{(f(Y_i), \delta_i): f(Y_i) = M - Y_i, i = 1, \dots, n\}.$$

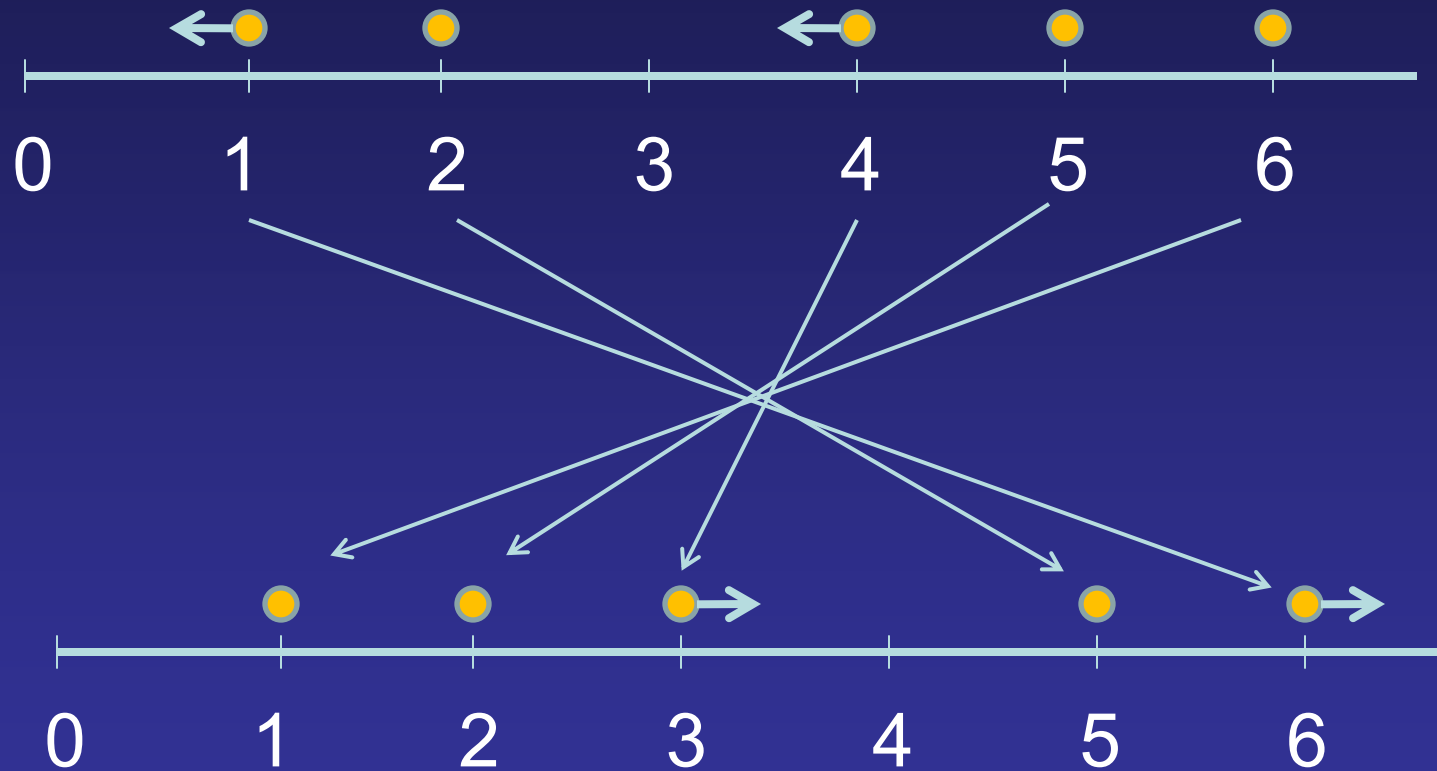
For values  $< LOD$ ,  $\{f(Y_i)\}$  is now right-censored at  $C_i = M - LOD_i$ .

Results from the rank-based nonparametric methods based on flipped data can be related to un-flipped data as follows:

- (1) *the  $p$ -value for testing the difference of the two groups will not change; ...*

# 'Flipping' the Data

Original data,  $X$  (left-censored):



Flipped data,  $Y = 7 - X$  (right-censored):

# Comparing Survival Curves

---

---

- Two Samples – many possible nonparametric tests

$$H_0: F_1(t) = F_2(t) \quad \text{for all } t$$

1. Logrank test (based on Mantel-Haenszel test)
  2. Peto-pike logrank test
  3. Wilcoxon test (Gehan or **Peto-Peto** versions)
  4. Family of rank tests (weighted logrank)
- Extensions of Two-sample Tests
    - Stratified logrank (or Wilcoxon) test
    - k-sample logrank (or Wilcoxon) test
    - Trend test

# 1. Mantel-Haenszel Logrank Test

For  $x_1 < x_2 < \dots < x_r$  distinct ordered uncensored values

at $x_j$	value at $x_j$	no value at $x_j$	
group 1	$d_{1j}$		$n_{1j}$
group 2	$d_{2j}$		$n_{2j}$
	$d_j$		$n_j$

Logrank test compares observed and expected number observed at each value and combines them over time - essentially M-H test for  $r$  2x2 tables

# Mantel-Haenszel Logrank Test – cont'd

$$U = \sum_{j=1}^r (d_{1j} - e_{1j})$$

where  $r$  = # of unique uncensored values in both groups,

$d_{1j}$  = observed # uncensored in group 1 at  $x_j$ ,

$e_{1j}$  = expected # uncensored in group 1 at  $x_j$

=  $n_{1j} (d_j/n_j)$

= (# values in group 1 that are  $\geq x_j$ )  $\times$

(# uncensored at  $x_j$  in both groups / # that are  $\geq x_j$ )

Then,  $U^2 / \text{var}(U)$  has a  $\chi^2_{(1)}$  distribution under  $H_0$ .

### 3. Wilcoxon Test

$$W = \sum_{j=1}^r n_j (d_{1j} - e_{1j})$$

- **W = Weighted** sum of (observed minus expected number of values).
- $W^2 / \text{var}(W)$  has a  $\chi^2_{(1)}$  distribution under  $H_0$
- Wilcoxon gives more weight to early values, and therefore, it is less sensitive to differences between 2 groups occurring at later values. (For left censoring, this is flipped – more weight to later values)

## 4. Family of Logrank Tests

$$\sum_{j=1}^r w_j (d_{1j} - e_{1j})$$

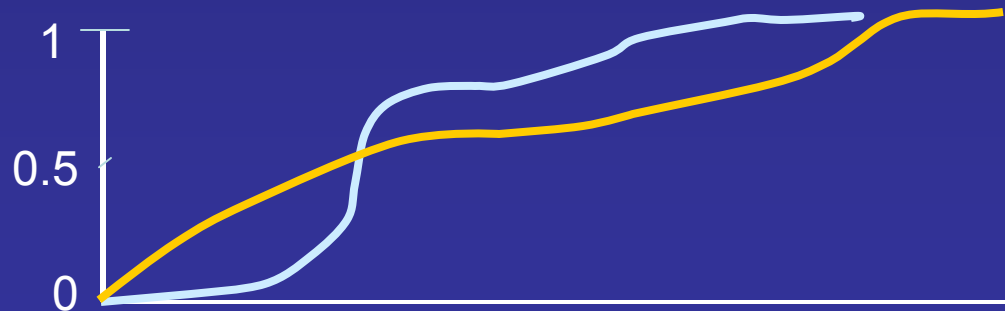
- $W_j = 1 \quad \Rightarrow$  logrank test
- $W_j = n_j \quad \Rightarrow$  Wilcoxon test
- $W_j = \hat{S}(t_{j-1}) \quad \Rightarrow$  **Peto-Peto-Wilcoxon**  
where  $\hat{S}(t_{j-1}) =$  KM estimate from pooled sample
- $W_j = \sqrt{n_j} \quad \Rightarrow$  Tarone-Ware Test

# Which Test should I use?

- Advice for right-censored data (Have to reconsider for left-censored data):
- Use the **logrank** or **Wilcoxon** test as the “default”. Usually equal weighting along the survival curve is reasonable.
- Weights of  $\hat{S}(t_j)$  or  $\sqrt{n_j}$  are also reasonable and may become more popular as software develops.
- When early values are of most interest [later values for left-censoring], use the **Wilcoxon** test. (An *a priori* decision.)
- SAS Proc Lifetest also provides a test labeled “-2 Log (LR)”, which is an LR test of equality of exponential parameters and assumes that both groups follow an exponential distribution. Use only after assumption check. [**Not likely to be useful for left-censored data**]

# Which Test should I use (cont'd)?

- If  $F(x)$  curves cross, then the positive and negative deviations will average out and result in a non-significant logrank test, even if a strong pattern is evident (see figure).
  - Always graph distribution functions for the groups you want to compare.
  - Sometimes it is not clear whether  $F(x)$  curves are truly different and crossing, or whether the groups have similar shapes and are crossing due to random variation.
  - With crossing curves, try a Kolmogorov-Smirnov or other similar test.



---

# Regression Models Assuming a Parametric (e.g., Lognormal or Weibull) Distribution

# Parametric Regression Models

---

- Parametric methods: fitting a specific distribution to the data such as **lognormal**, Weibull, or log-logistic.
- Why regression models?
  - To test the effect (predictive value) of covariates on the concentration

# Parametric Regression Models – cont'd

---

- Recall ordinary linear regression  
 $Y = X\beta + \varepsilon$ , where  $\varepsilon \sim N(0, \sigma^2)$
- Rewrite as  $Y = X\beta + \sigma \varepsilon$ , where  $\varepsilon \sim N(0, 1^2)$
- Consider  $Y = \log(T)$
- Let  $\varepsilon$  have a distribution such as **normal**, extreme value, logistic, or log-gamma.

## Parametric Regression Models – cont'd

---

- Cannot use ordinary least squares in non-normal situation. Use maximum likelihood to estimate the  $\beta$ 's.
- Write the likelihood function to incorporate censored data (right, left, or interval).  
--> parametric models for censored data.
- They are ordinary linear models in  $\log(T)$  with possibly non-normal errors.
- Modeling in  $\log(T)$  also gives the Accelerated Failure Time (AFT) Property.

# Accelerated Failure Time (AFT) Model

$$S_i(t | \mathbf{x}) = S_0(t \cdot \varphi_i(\mathbf{x})) \quad [ = 1 - F_i(t | \mathbf{x}) ]$$

where  $S_0(t) = \text{pr}(T > t | \mathbf{x} = \mathbf{0})$

= baseline survivor function

–  $\varphi(\mathbf{x})$  is usually chosen so that  $\varphi(\mathbf{0})=1$

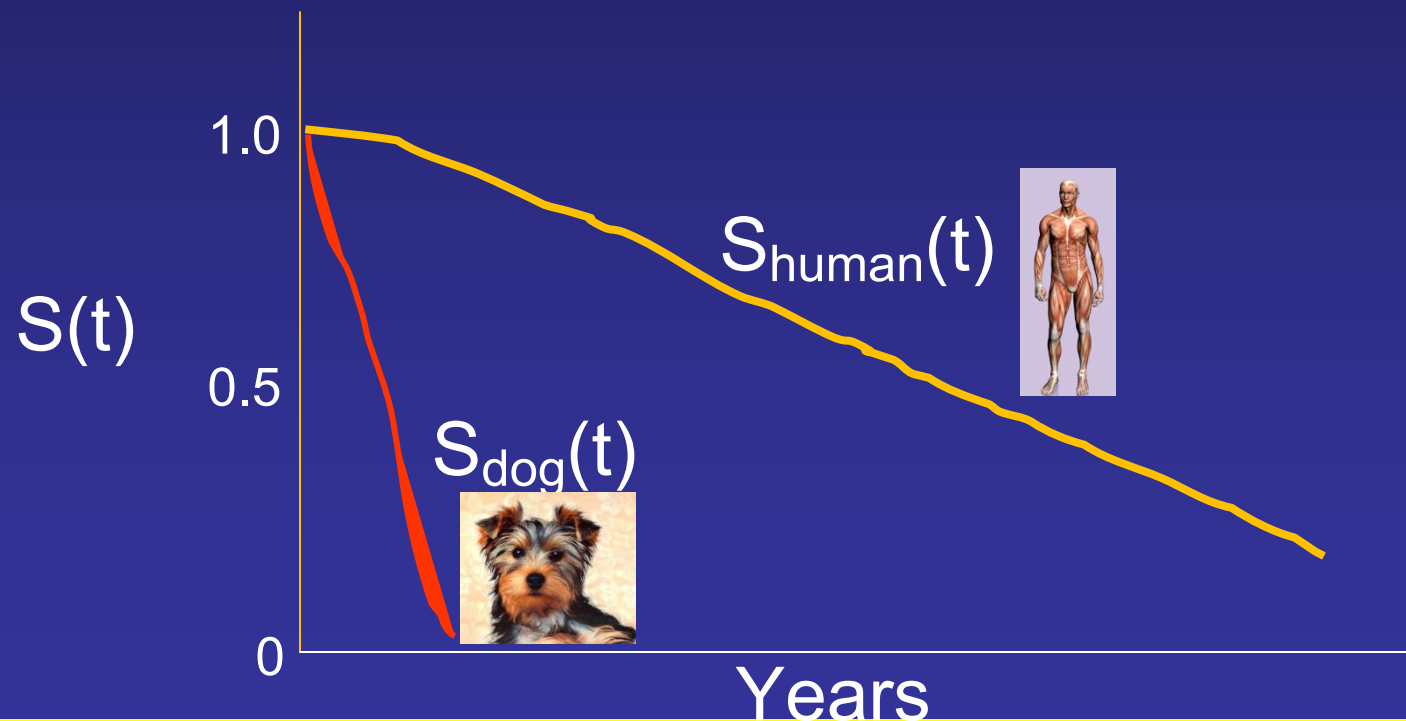
–  $\varphi(\mathbf{x}) = \exp\{\mathbf{x}'\beta\}$

–  $\mathbf{x}$  shrinks or stretches the time scale via  $\varphi(\mathbf{x})$ ,  
that is, **covariates alter the speed at which a  
subject proceeds through time**

# AFT Model Example

A year for a dog is equivalent to 7 years for a human.

$$S_{\text{dog}}(t) = S_{\text{human}}(7t) \quad \text{or} \quad S_{\text{human}}(t) = S_{\text{dog}}(t/7)$$



# AFT Model - continued

$$\log T_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \sigma \varepsilon_i$$

$$T_i = \exp[ \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \sigma \varepsilon_i ]$$

$\beta_0$  = intercept,  $\sigma$  = scale,  $\varepsilon_i$  = random error

- The log transformation of T ensures predicted time to be positive.
- **SAS fits four different AFT models:**
  - Weibull (including exponential)
  - Log-logistic
  - Lognormal, and
  - Gamma (3 parameter).

# SAS Code for Parametric Regression

*In SAS, use the Lifereg procedure:*

```
proc lifereg;
```

```
    model (t1, t2)= sex age
```

```
    / dist = lnormal;
```

(or dist = exponential, weibull, llogistic, gamma)

If left-censored:      t1=. ;      t2=LOD ;

If exact (say, =x):    t1=x ;      t2=x ;

*In R, use the NADA cenreg() function.*

# Reverse Cox Regression

---

Similar to the logrank or Wilcoxon tests, we can use the “flipped” data and fit a Cox regression Model.

Interpretation of covariate effects may be difficult because:

- the scale is reversed
- the concept of “proportional hazards”, and even the hazard function itself, are not so intuitive when the scale is not “time”.

# What we have covered

---

- 1) Below LOD data as left-censored data
- 2) Estimating the distribution function (and population percentiles) with left-censored data  
(reverse KM or Turnbull)
- 3) Two-sample tests with left-censored data  
(logrank or **Wilcoxon**)
- 4) Regression with left-censored data
  - 1) Parametric (e.g., Weibull, lognormal)
  - 2) Semi-parametric (e.g., reverse Cox)

# Conclusions

---

- Statistical methods are well developed for the left-censored case.
- These methods should be used in preference to *ad hoc* methods.
- Use of these methods would be facilitated by further software development.