Water-Quality in Everglades National Park: Statistical Methods for Censored Data

SFNRC
NPS

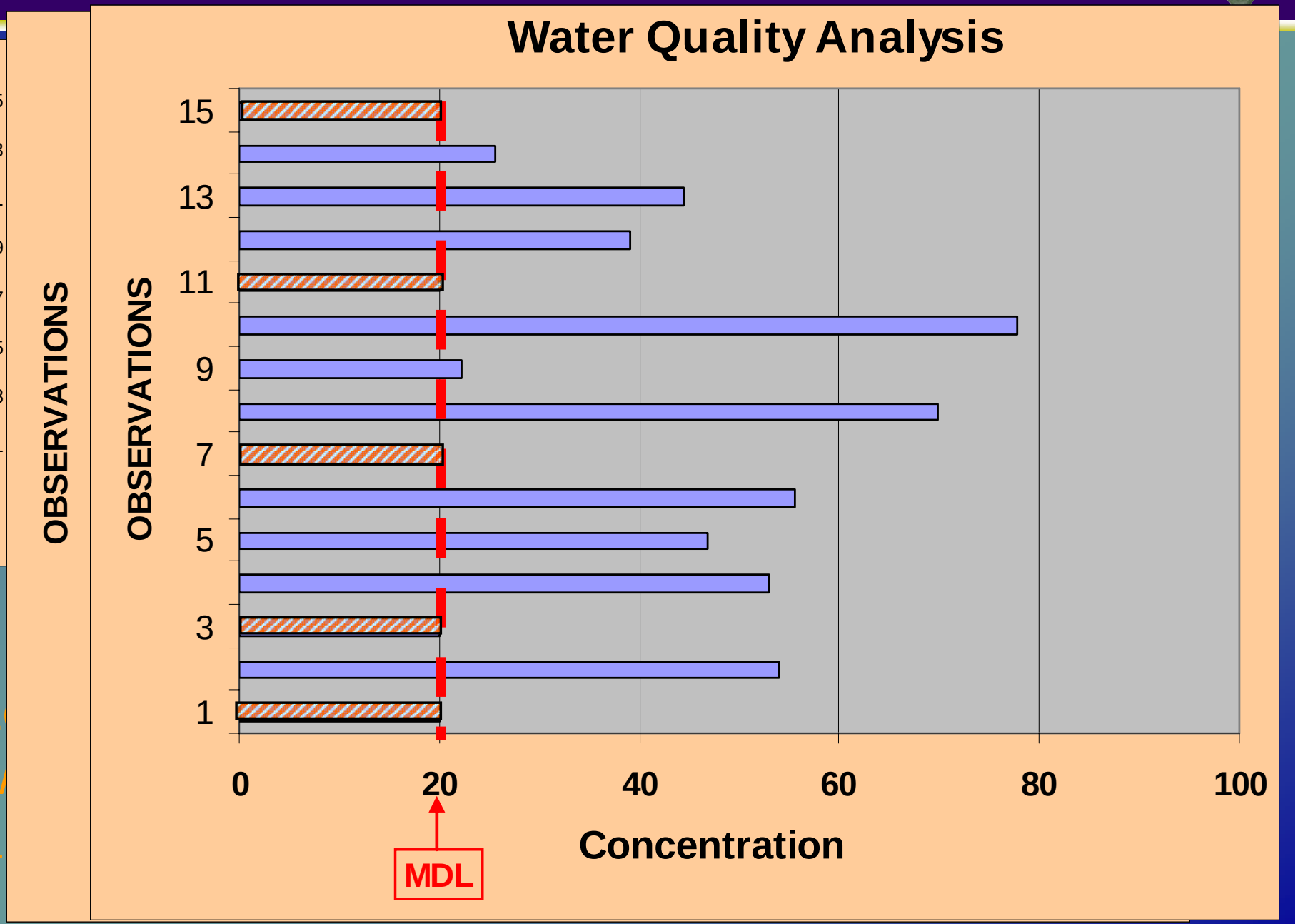
Joffre Castro

July, 2008

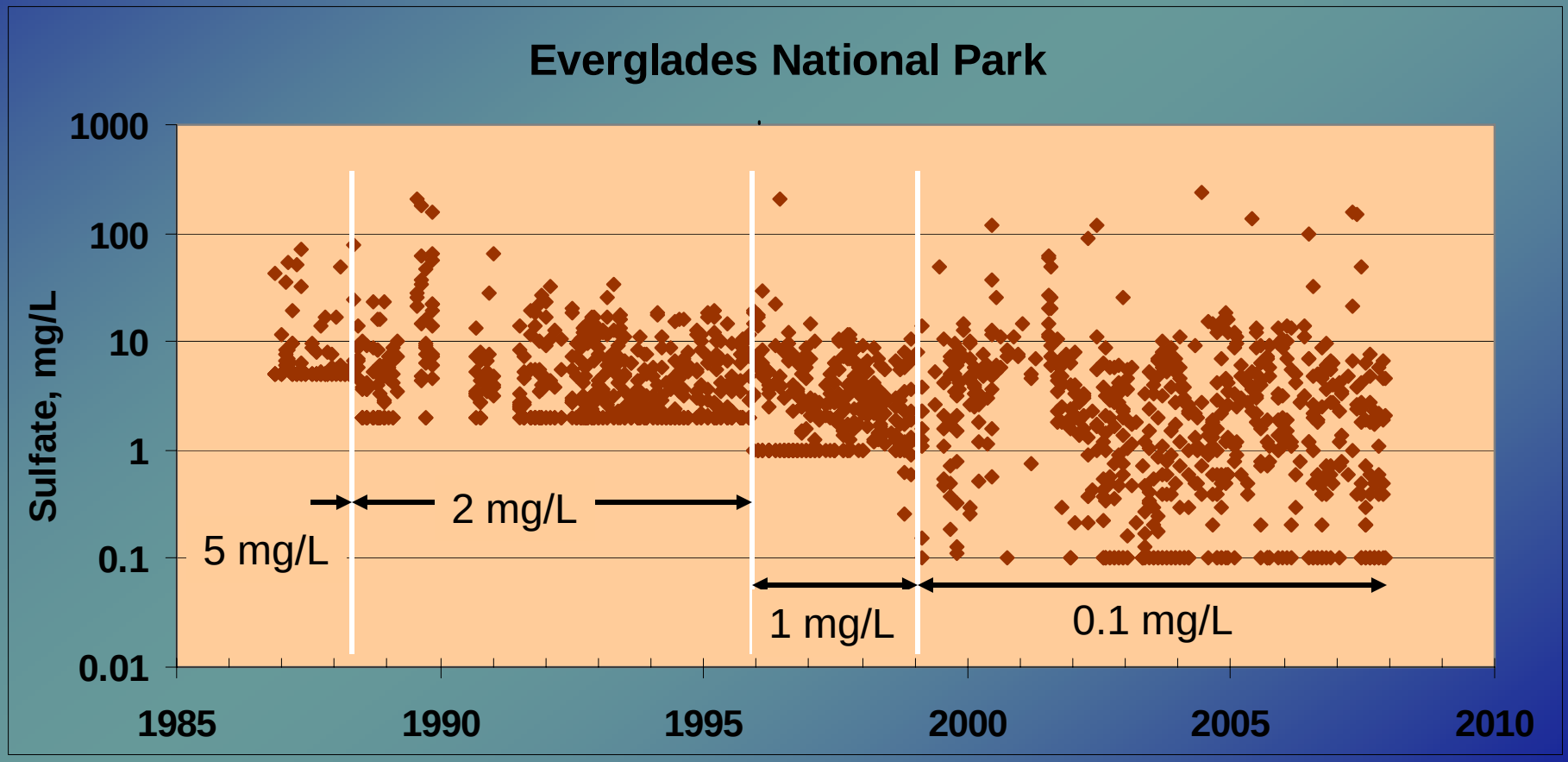
CENSORED DATA

- **Data below a limit**
- **Limits:**
  - Single value, range
  - Types:
    - Practical Quantitation
      - Method Reporting
        - Method Detection
          - Reliable Detection
- **Assigned by laboratory**
- **Qualifies the data**
- **“noise” in analytical process**

Censored Data



MULTIPLE MDLs



Everglades National Park WQ Data

No- MDLs

Alkalinity
Calcium
Chloride
Color
DO
Hardness
Magnesium
pH
Potassium
Sodium
SpC
Turbidity

	EP	NE1	P35	P36	P37	TSB	Total
			OF MDLs				
Ammonia-N	33	16	40	14	37	48	30
Arsenic	79	40	58	49	77	91	63
Cadmium	89	73	81	80	82	90	84
Copper	64	61	58	59	66	71	63
Iron	20			1	1	4	3
Kjeldahl-N	7		1	1	11	9	3
Lead	89	81	78	83	85	87	83
Nox-N	23	21	10	16	12	26	20
Sulfate		28	22	18	64	25	25
Total-P	37	9	11	11	46	28	22
TSS	75	57	66	60	72	77	68
Zinc	85	88	78	90	87	88	86

Black numbers are less than 50 perc

ers are between 50 and 80 percent nondetects

ers are greater than 80 percent nondetects

METHODS

- Substitution (arbitrary)
- Robust Methods
  - Parametric methods

    (assumes a distribution)
  - Non-parametric methods

    (distribution free)

METHODS (continued)

Simple Substitutions

- Zero
- MDL/SQRT(2)
- ½- MDL
- 0.75-MDL
- 1-MDL

Recommended Methods

- **KM (Kaplan-Meier)[1]**
- ROS (regression on order statistics)
- **MR (multiple-limit regression) [1]**
- MLE (maximum likelihood estimate)
- AMLE (adjusted maximum likelihood estimate)

[1] none parametric

Simple Substitution

STAID	DATES	Arsenic
P34	6/9/1986	1.40
P34	7/25/1986	1.30
P34	8/18/1986	-0.80
P34	9/15/1986	1.40
P34	10/27/1986	1.20
P34	11/19/1986	-0.80
P34	12/30/1986	1.60
P34	1/26/1987	2.03
P34	2/9/1987	1.36
P34	3/9/1987	-0.80

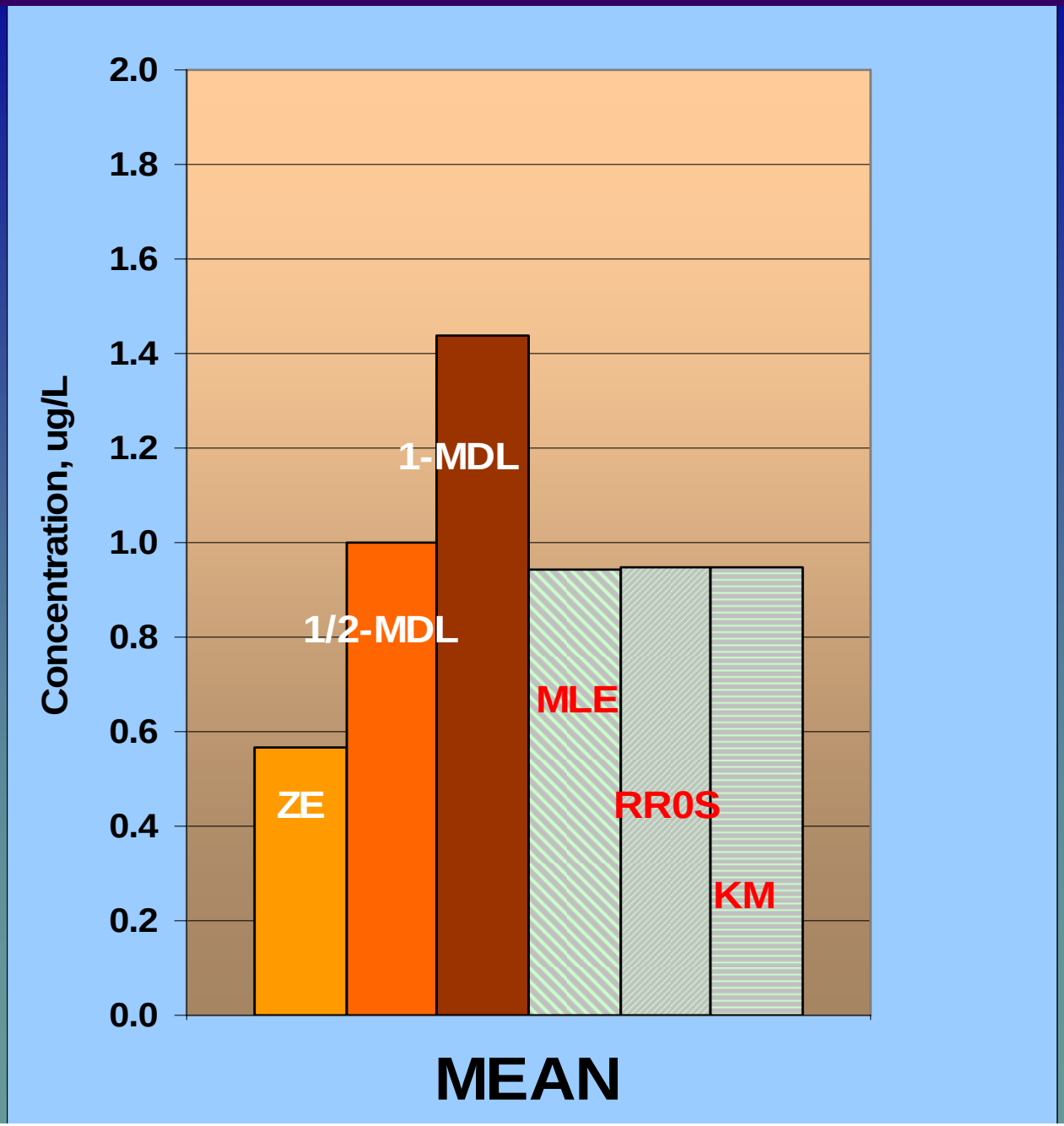
Values below MDL

ZE-→ 0

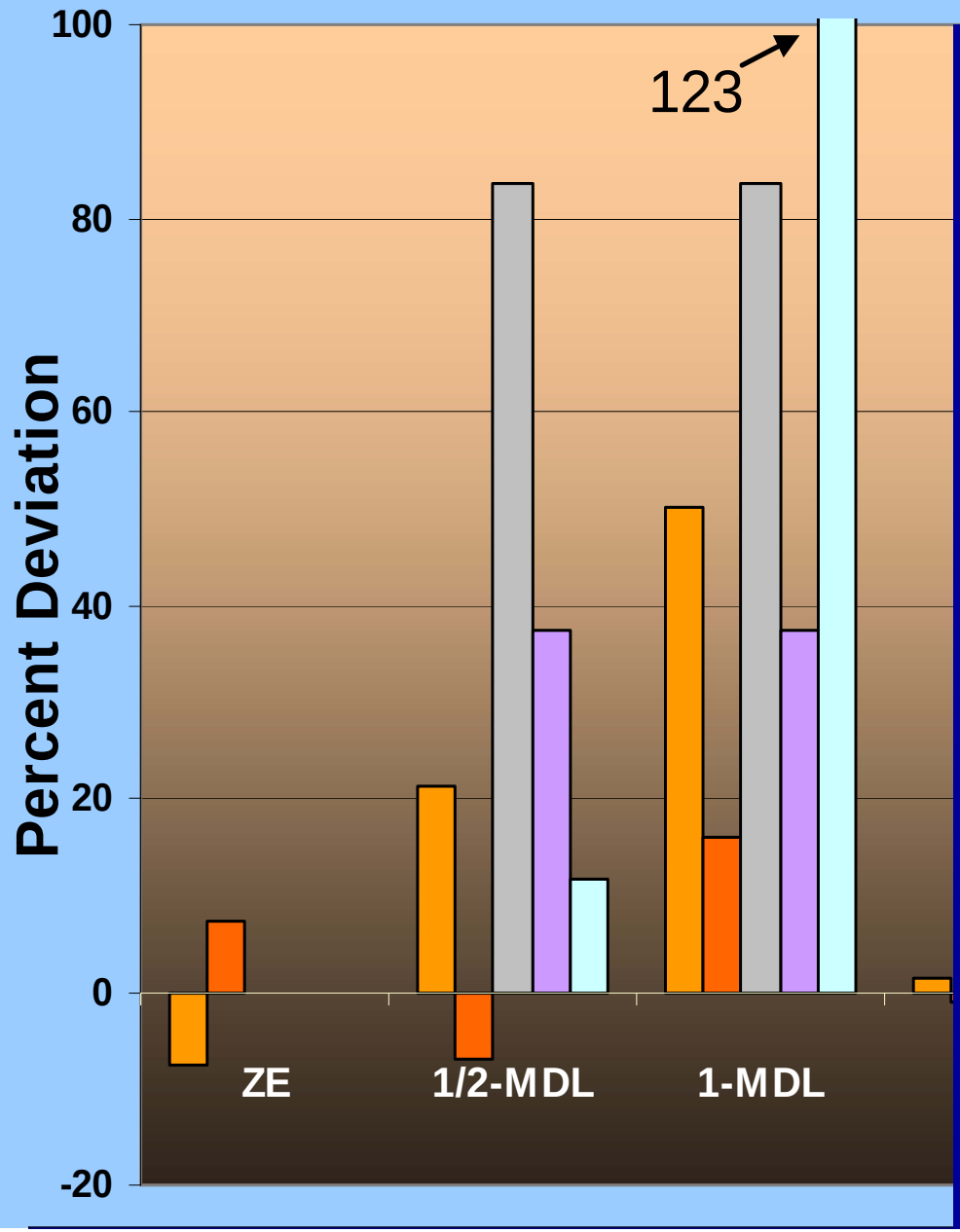
½-MDL→ 0.4

1-MDL→ 0.8

Comparison of Results

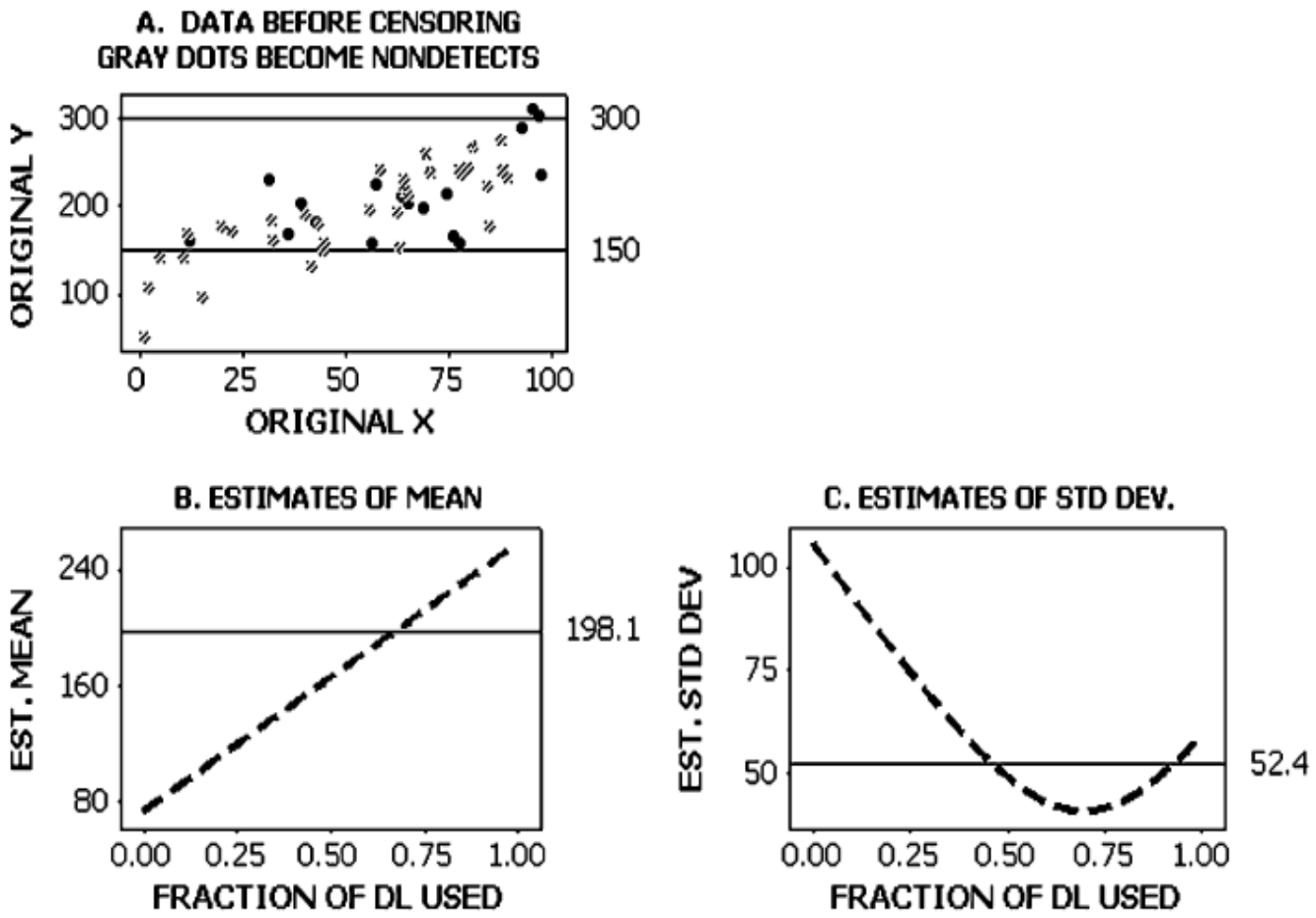


Comparison of Results (continued)



Simple Substitution Results

Helsel, 2006



From H
Chemosphere 65, 2434–2439)

Simple Substitution (continued)

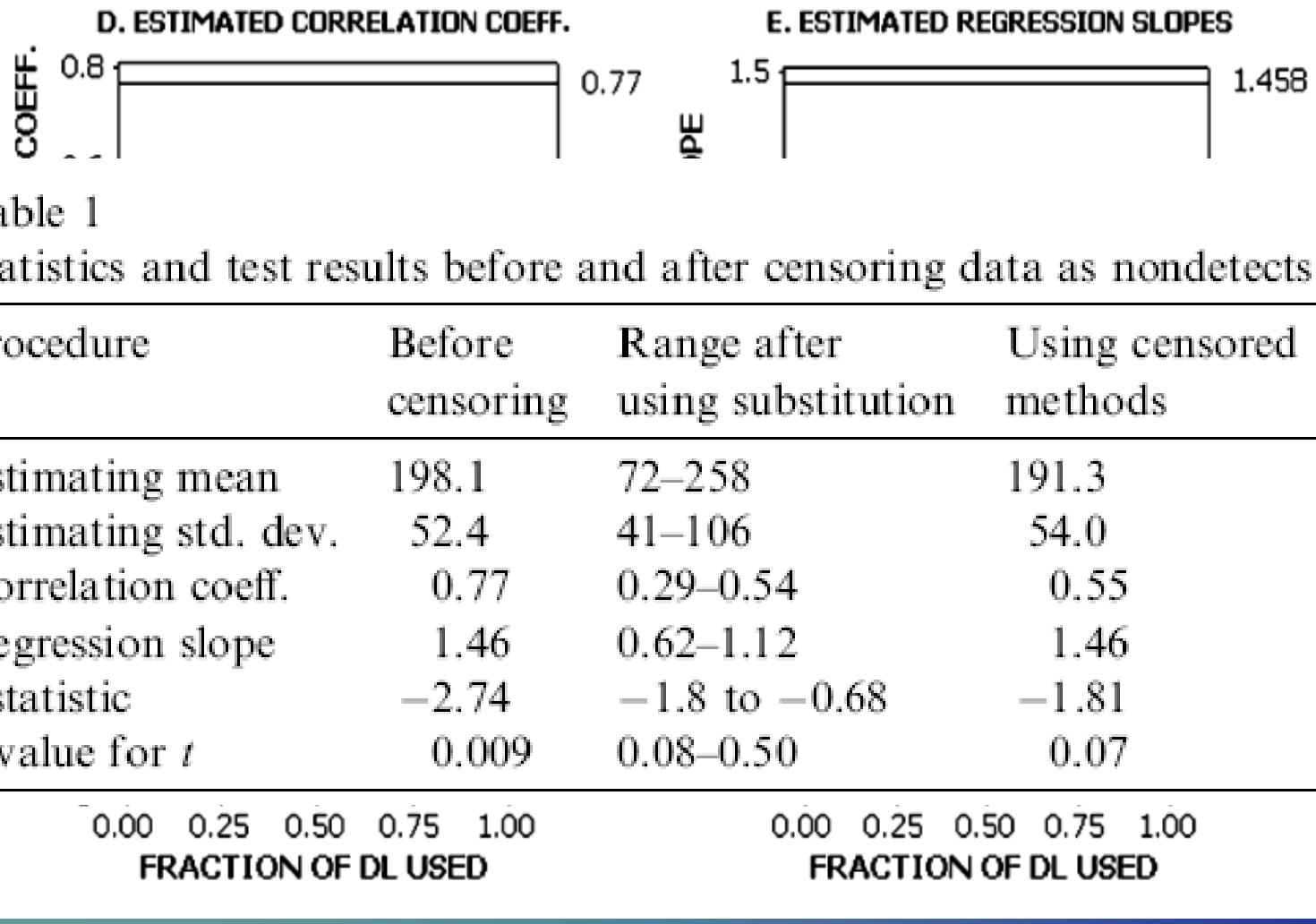


Table 1
Statistics and test results before and after censoring data as nondetects

Procedure	Before censoring	Range after using substitution	Using censored methods
Estimating mean	198.1	72–258	191.3
Estimating std. dev.	52.4	41–106	54.0
Correlation coeff.	0.77	0.29–0.54	0.55
Regression slope	1.46	0.62–1.12	1.46
t-statistic	−2.74	−1.8 to −0.68	−1.81
p-value for *t*	0.009	0.08–0.50	0.07

Recommendations

RECOMMENDED METHODS FOR SUMMARY STATISTICS	
Percent Censored	**>50 Observations**
<50% nondetects	Kaplan-Meier
50-80% nondetects	Maximum Likelihood
>80% nondetects	May report high sample percentile 90th and 95th

Helsel, 2005

Statistical Tests

	STATISTICAL METHODS	
	UNCENSORED	CENSORED
Summary statistics	Descriptive	K-M, MLE, ROS
Comparing two groups	t-test	Censored regression, 0/1 indicator
	Wilcoxon rank-sum	Generalized Wilcoxon test
	Paired t-test	Censored CI on differences
	sign/signed rank test	PPW or Akritas tests
Comparing three or more groups	ANOVA	Censored regression, 0/1 indicator
	Kruskal-Wallis test	Generalized Wilcoxon test
Correlation	Pearson's r	Likelihood r
	Kendall's tau	Kendall's tau b
Linear regression	Regression	Censored regression
		Logistic regression
	Robust regression	Proportional hazards (Cox) regression
	Theil-Sen median line	Akritas-Theil-Sen median line

From Helsel, 2005

Conclusions

- **Substitution methods:**

  arbitrary and inaccurate

- **Robust methods:**
  - Statistically sound
    - ✓ Parametric: MLE, AMLE, ROS
    - ✓ Non-parametric: KM, MR, bootstrapping
  - Available in most stats. packages
  - Suitable for most tests (trend analysis)
  - Easy to use