Nondetects And Data Analysis:
Interval Estimates with NDs

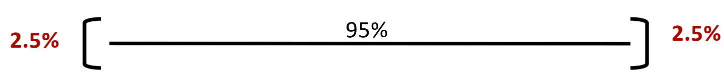
Dennis R. Helsel, Ph.D

Practical Stats

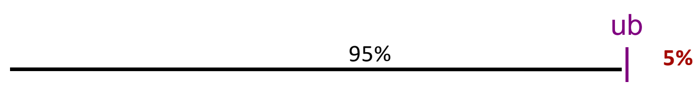
© PracticalStats.com

Interval Estimates

Two-sided intervals: Where might the true value be -- with (1-α)% probability)? For α = 0.05:



One-sided upper bounds: The true value is no more than ub -- with (1-α)% probability). For α = 0.05:



2

Three types of Intervals

1. Confidence intervals: range of possible values for the true population parameter (mean, etc)
2. Prediction intervals: range of possible values for one or more future observations from the same population
3. Tolerance intervals: limit of possible values for a population percentile

3

1. Confidence Intervals

Use the same methods as for estimating summary statistics:

Parametric method

- MLE software

Semi-parametric method

- Robust ROS

Nonparametric method

- Kaplan-Meier

4

Older t-confidence Interval on Mean

$$\bar{x} - t_{(1-\alpha/2),n-1} \bullet s/\sqrt{n}\,, \qquad \bar{x} + t_{(1-\alpha/2),n-1} \bullet s/\sqrt{n}$$

Parametric Interval, two-sided

- Assumes data follow normal distribution, or that the variation in estimates of mean do ("Central Limit Theorem", requires 50+ obs.)

If used with small skewed data sets, probability of including true mean within interval may be much smaller than the expected (1-α) -- can compute a '95% CI' but get a ~60% CI

With nondetects, estimate parameters with MLE, K-M or robust ROS

Cannot take logs, compute means, sd and compute CI in log space, then transform back. (This can be done for prediction and tolerance intervals, but not confidence intervals). Instead, use equations & functions for those distributions!

5

CI for Mean of Lognormal Data

There are several methods, all have problems except bootstrap

- Commercial software often uses a normal CI in log units (Cox method)

$$\exp(\underbrace{\boxed{\bar{y} + s_y^{\,2}/2}}_{\text{lognormal mean}} \pm z_{\alpha/2} \bullet \gamma)$$

- where (requires a good estimate of s, the std dev)

$$\gamma = \frac{s^2{}_y}{n} + \frac{s^4{}_y/2}{n+1}$$

Best alternative is a nonparametric approach, bootstrapping

6

Confidence Interval on the Mean - Lognormal

```
elnormAltCensored(Pyrene, PyreneCen, ci=TRUE, ci.method = "bootstrap", n.bootstraps = 5000)
Results of Distribution Parameter Estimation Based on Type I Censored Data
--------------------------------------------
Assumed Distribution:           Lognormal            parameters fit by MLE;  CI by bootstrap
Censoring Side:                 left
Estimated Parameter(s):         mean = 133.914189
Estimation Method:              MLE
Confidence Interval Method:     Bootstrap
Number of Bootstraps:           5000
Confidence Interval Type:       two-sided
Confidence Level:               95%
Confidence Interval:            Pct.LCL = 100.1207       BCa.LCL =  98.3675
                                Pct.UCL = 189.0668       BCa.UCL = 184.7112
                                percentile bootstrap     Bias corrected bootstrap
```

7

Confidence Interval on the Kaplan-Meier Mean

```
enparCensored(Pyrene,PyreneCen, ci=TRUE, ci.method="bootstrap", n.bootstraps = 5000)
Results of Distribution Parameter Estimation Based on Type I Censored Data
--------------------------------------------
Assumed Distribution:           None              parameters fit by K-M;  CI by bootstrap
Censoring Side:                 left
Estimated Parameter(s):         mean    = 164.09450
                                sd      = 389.41379
                                se.mean =  49.75292
Estimation Method:              Kaplan-Meier
Confidence Interval Method:     5000  Bootstraps
Confidence Interval Type:       two-sided
Confidence Level:               95%
Confidence Interval:            Pct.LCL = 100.10254      BCa.LCL =  98.68195
                                Pct.UCL = 264.47772      BCa.UCL = 261.92596
                                percentile bootstrap     Bias corrected bootstrap
```

8

Confidence Interval on the rROS Mean

```
elnormAltCensored(Pyrene, PyreneCen, method = "rROS", ci = TRUE, ci.method = "bootstrap", n.bootstraps =
     5000)
Results of Distribution Parameter Estimation Based on Type I Censored Data
--------------------------------------------
Assumed Distribution:            Lognormal
Censoring Side:                  left
Estimated Parameter(s):          mean = 163.371129
Estimation Method:               Imputation with Q-Q Regression (rROS)
Confidence Interval Method:      5000 Bootstraps
Confidence Interval Type:        two-sided
Confidence Level:                95%
Confidence Interval:             Pct.LCL = 100.94089      BCa.LCL =  97.22056
                                 Pct.UCL = 264.69006      BCa.UCL = 255.91613
```

parameters fit by rROS; CI by bootstrap

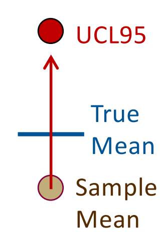
percentile bootstrap — Bias corrected bootstrap

9

Upper 1-sided Confidence Limit on Mean

$$\bar{x} + t_{(1-\alpha),n-1} * {}^{s}\!/\!\sqrt{n}$$

- Is a ‘protective’ estimate of the mean
- The sample mean may be lower or higher than the true population mean
- The true population mean has only a 5% probability to exceed the UCL95.
- The sample mean is the best estimate of the population mean, given the observed data



10

Bootstrapping

1. Sample with replacement a temporary set of n observations
   (see next slide)
2. Compute an estimate of the mean of the temporary set of obs.
3. Save the estimate of the mean and repeat the process many times with a new temporary set of data. Ten thousand is a commonly-used number of replicates. All of these are equally plausible to the observed sample mean, given the observed data
4. Locate the 2.5th and 97.5th percentiles of the estimates of the mean. These are the endpoints of the two-sided 95% "percentile bootstrap" confidence interval for the mean; the 95th percentile of the estimates is the 1-sided percentile bootstrap UCL95.
5. The BCA bootstrap corrects the percentile bootstrap for skewness of the data. However, it incorrectly estimates the correction with nondetects. Do not use if the % censoring is greater than ~40% - use the percentile bootstrap instead.

11

Bootstrapping

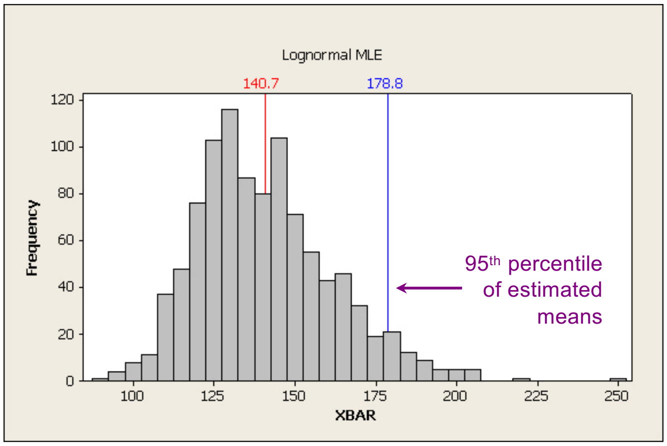
Sampling with replacement

Original		Temp 1	Temp 2	Temp 3		
3		3	10	13		
5	⇒	21	34	34	⇒	and 5 to 10 K more
10		34	10	34		
13		21	3	5		
21		21	21	21		
34		21	5	21		
$\overline{x}$ = 14.33		20.17	13.83	21.33		

12

Histogram of Bootstrap Results

Histogram of bootstrapped means using MLE:



Mean
140.7

UCL95
178.8

Unlike a t-interval, the bootstrap interval is not required to be symmetric

13

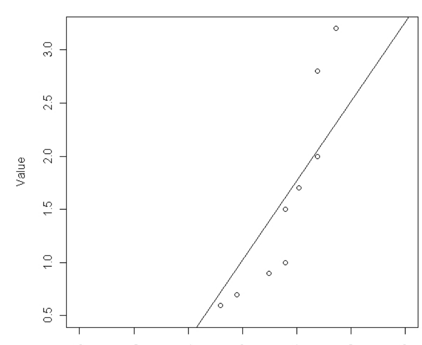
95% CI for Arsenic by MLE: NADA package

(a t-interval)

```
> omle=cenmle(As, AsCen, dist="gaussian")
> omle
        n     n.cen     median       mean         sd
24.0000000 13.0000000  1.0200176  1.0200176  0.7451676
> mean(omle)
     mean         se    0.95LCL    0.95UCL
1.0200176 0.1684655 0.6898313 1.3502039

> plot(omle)
```

(data don't fit a normal distribution well)



14

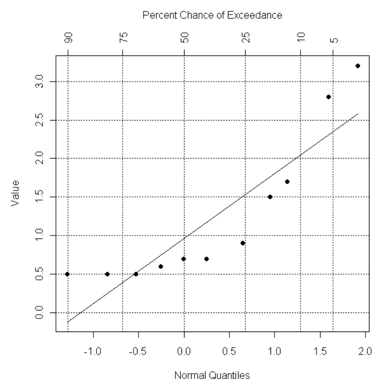
© PracticalStats.com

95% CI for Arsenic by ROS: NADA package

(a t-interval)

```
> oros=cenros(As,AsCen, forwardT=NULL)
> oros
         n      n.cen     median       mean         sd
24.0000000 13.0000000  0.7317261  0.9923255  0.7775574
```

Normal Distribution



Compute by hand using t interval formula (no bootstrap in NADA package):

```
95%L    Mean     95%U
0.67    0.992    1.32
```

(data don't fit a normal distribution well)

15

© PracticalStats.com

Evaluation of Substitution for computing the UCL95

Singh et al (2006), developers of the ProUCL software for USEPA, determined that substituting ½ DL "does not provide adequate coverage [UCL95 is not high enough] …even for censoring levels as low as 10%"

Singh et al (2010) "...strongly recommends avoiding the use of the DL/2 method even when the percentage of NDs is as low as 5%-10%."

16

Summary UCL95 for data with nondetects

Bootstrapping K-M provides a UCL95 with the fewest assumptions. Requires 20 or so observations to cover the likely range of values. Available in enparCensored Works well unless a large % of data are below the lowest (or only) detection limit.

With larger datasets (n ≧ 50), can use MLE with bootstrapped CI. elnormCensored, etc.

With smaller datasets (n<20), rROS recommended because its not as sensitive as MLE to the distribution assumption (which is difficult to get correct with few data). elnormCensored with method = rROS option.

Not likely to get a good result with any method when n<8

Substitution does not work well

17

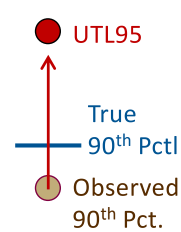
Prediction and Tolerance Intervals

These can be done with the EnvStats package. See Millard, S.P. 2013, *EnvStats: An R Package for Environmental Statistics*. (Springer).
-- or the free pdf guide to the R package on CRAN.

Prediction Intervals -- an interval with defined probability that the next 1 (or several) individual observations will be within it, if the distribution has not changed. Length increases as # of new observations increases

Tolerance intervals – an interval with defined probability that no more than (1-p)% of observations will exceed it, with (1-α)% probability.
For p = 0.9 -- "what value will at least 90% of observations lie below, and no more than 10% of observations exceed it, with 95% probability?"

p = coverage α = confidence coefficient



18

Prediction Intervals with Nondetects

```
> cenPredInt (Pyrene, PyreneCen, newobs = 1)
Lognormal 95% Prediction Limits
      LPL       UPL
 15.75406 533.15646

 Normal 95% Prediction Limits
      LPL       UPL
-783.7555  992.1820

 Approx. Gamma 95% Prediction Limits
        LPL         UPL
  0.7231388 581.0615117
```

"Within what range should I expect 1 or more new observations to fall when there is no change from this dataset, with 95% probability"?

- NADAscript that runs several EnvStats commands at once
- Uses MLE by default; can change to rROS
- Can compute either 2-sided intervals (default) or 1-sided upper or lower limits
- Default is for 1 new observation
- Accounts for censored observations by MLE or rROS estimates of mean and sd of logs (lognormal), of the data (normal), and of cube roots (gamma), computing normal PIs on those, and retransforming back the log and cube root interval endpoints.

19

Upper Prediction Limit with Nondetects

```
> cenPredInt (Pyrene, PyreneCen, newobs =2,
    pi.type="upper", method = "rROS")
Lognormal 95% Prediction Limit
     UPL
516.7649

 Normal 95% Prediction Limit
     UPL
969.0796

 Approx. Gamma 95% Prediction Limit
     UPL
572.7311
```

"What is a value that when exceeded by one or two higher new observations indicates that concentrations have changed (no longer background)"?

- Changed to 1-sided upper limit (UPL) with the pi.type option
- Changed to rROS with the method option
- Changed to 2 new observations with the newobs option

20

Tolerance Intervals with Nondetects

(EnvStats package)

```
> eqlnormCensored (Pyrene, PyreneCen, p=0.9,
    ci=TRUE, ci.type = "upper")

Results of Distribution Parameter Estimation
Based on Type I Censored Data
--------------------------------------------
Assumed Distribution:             Lognormal
Censoring Side:                   left
Censoring Level(s):               28  35  58  86
                                 117 122 163 174
Estimation Method:                MLE
Estimated Quantile(s):            90'th %ile
                                   = 279.7995
Confidence Interval:              LCL =   0.0000
                                  UCL = 376.4538
```

"What is the highest value I would expect the 90th percentile to be, with 95% probability (UTL95)? or

"What value should 90% of all new observations fall below, with 95% probability"?

- EnvStats functions for lognormal, normal and gamma distributions. Solve by MLE or rROS (for 1st two)
- Must specify the coverage -- the probability p for the chosen percentile
- Only 1-sided upper limit is of use in environmental applications
- Accounts for censored observations by MLE or rROS. Estimates percentile of logs, of the data, and of cube roots, computing normal TIs on those, and retransforming back the log and cube root interval endpoints.

21

Tolerance Intervals with Nondetects

```
cenTolInt(Pyrene, PyreneCen, p=0.9)
> cenTolInt (Pyrene, PyreneCen, cover = 85)
Lognormal 85th Pctl      95% Upper Tol Limit
           226.0141       295.7412

 Normal 85th Pctl       95% Upper Tol Limit
        559.3986         694.9937

 ~Gamma 85th Pctl       95% Upper Tol Limit
        278.4247         357.6741
```

- For an easy way to get the gamma results, take cube roots and use the eqnormCensored function. Then cube the results
- Done for you with the `cenTolInt` function in NADAscripts.R to get the results for all 3 distributions
- I don't guarantee that the formatting in the `cenTolInt` function is correct yet, but the numbers are

22

Methods for censored data

Method	Parametric	Nonparametric
Descriptive stats	MLE ROS	Kaplan-Meier
Intervals	Bootstrapping MLE ROS	Bootstrapping K-M
Paired Data	CI on differences by MLE	PPW
2 Indep Groups	MLE Regression on 0/1 factor	Peto-Peto
3+ Indep Groups	MLE Regression on 0/1 factor	Peto-Peto
Correlation	Likelihood R by MLE	Kendall's tau
Regression	MLE Regression	ATS line

23