Nondetects And Data Analysis:
Compare Data with Nondetects to Standards

Dennis R. Helsel, Ph.D
Practical Stats

1

How would you compare data without nondetects to standards?

1. If the sample mean is higher than the standard, find noncompliance?

   This doesn't account for the difference between your sample mean and the true mean out in the field (population). The probability of stating the exceedance or non-exceedance will be incorrect. See Smith et al (2001)*

2. Use the t-test to see if the mean is higher than the standard?

   Requires a normal distribution. If used on skewed data with n < 70-100 observations, will have low power to see exceedances.

3. Use the t-test on logs to fix skewness and see if the mean log of data is higher than the log of the standard?

   This doesn't test for whether the mean exceeds the standard, but whether the geometric mean (median) exceeds the standard. Its not how the regulation was probably written.

* Smith EP, Ye K, Hughes C, Shabman L. 2001. *Statistical assessment of violations of water quality standards under Section 303(d) of the Clean Water Act* Environmental Science and Technology 35: 606–612

2

2

Comparing Data with Nondetects to a Standard is No Different

Solution? Fit a skewed distribution to the data and test whether the mean exceeds the standard or not. Tests like a t-test compare confidence interval endpoints to the standard.

1. For n ≧ 20 observations, use a one-sample bootstrap test -- compute a bootstrap UCL95 or LCL95 and compare that to the standard.
2. For small sample sizes (n<20) assume the best-fit distribution and compute the UCL or LCL. Lognormal and gamma distributions are commonly used. The best-fit distribution is rarely to never the normal.
3. If there are 70 or more observations you can use t-interval UCL95 or LCL95 (a "t-test"). But there's no benefit to the t-interval over using a bootstrap UCL / LCL for larger n.

3

3

One-sample tests are just confidence intervals
1. H_0: Assume compliance

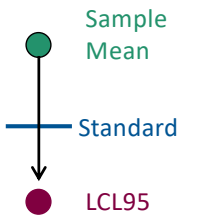
Assume compliance (the population mean is at or below the standard).

To reject the null hypothesis at $\alpha = 0.05$, the 95% one-sided LCL on the mean must be above the standard.

Example: The sample mean is above the standard, but this doesn't take into account the uncertainty that the population mean might be as low as the LCL.

The 95% one-sided lower confidence limit on the mean is computed.

If the LCL95 is not above the standard, then it is plausible at the 95% level that the population mean may be at or below the standard.



Do Not Reject H_0.
In compliance

4

4

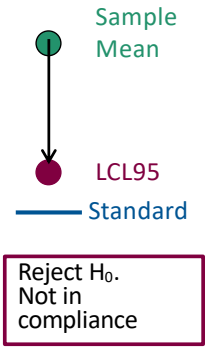
One-sample tests are just confidence intervals
1. H_0: Assume compliance

Assume compliance (the population mean is at or below the standard).

To reject the null hypothesis at $\alpha = 0.05$, the 95% LCL on the mean must be above the standard.

Example: The sample mean is above the standard, and the LCL95 is also above the standard.

It is not plausible at the 95% level that the population mean could be at or below the standard. Therefore, reject that the population mean is in compliance.



Reject H_0. Not in compliance

5

5

One-sample tests are just confidence intervals
2. H_0: Assume Non-compliance

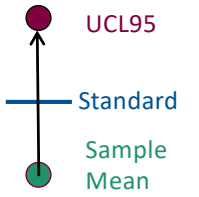
Assume non-compliance (that the population mean is above the standard).

To reject the null hypothesis at $\alpha = 0.05$ and find compliance, the 95% UCL must be below the standard.

Example: The sample mean is below the standard, but the UCL95 is above the standard.

It is plausible at the 95% level that the population mean could be above the standard. Therefore, do not reject that the population mean is out of compliance.

This is a strong burden of proof that the mean "in the field" must meet. And it assumes guilt until proven innocent.



Do Not Reject H_0. Not in compliance

6

6

One-sample tests are just confidence intervals
2. H_0: Assume Non-compliance

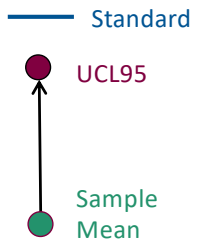
Assume non-compliance (the population mean is above the standard).

To reject the null hypothesis at $\alpha = 0.05$, the 95% UCL must be below the standard.

Example: The sample mean is below the standard, and the UCL95 is below the standard.

It is not plausible at the 95% level that the population mean has a value above the standard. Therefore, reject that the population mean is out of compliance.

This is a strong burden of proof that the mean "in the field" must meet. And it assumes guilt until innocence is proven.



Reject H_0.
In compliance

7

7

Flowchart for comparing data with nondetects to standards

Step 1. Are there at least 20 observations?

Yes.

Compute the UCL / LCL with a bootstrap. Could use KM or best fitting distribution.

Go to Step 3.

No.

Determine and use the best fitting distribution to compute UCL / LCL.

Go to Step 2.

8

8

Flowchart for comparing data with nondetects to standards

Step 2. Distributional Method
Decide which distribution best fits the data.

- Use the cenCompareCdfs command and choose the distribution with the lowest BIC.
- Use the best fit distribution to compute the LCL or UCL.

```
> cenCompareCdfs (data, data_Cen)
> elnormAltCensored(data, data_Cen, ci = TRUE, ci.type = "lower")
```

(to assume compliance)

9

9

BIC or PPCC or W to find the best fitting distribution?

- Choose the distribution with the lowest BIC (error) statistic output by the cenCompareCdfs function.
- If your software computes the PPCC or Shapiro-Wilk statistic instead (as does the cenCompareQQ function), choose the distribution with the highest PPCC or Shapiro-Wilk W. However, these two are less able to pick up the fitted normal distribution going below 0 than is BIC. If using PPCC or W, visually make sure that the lower percentiles are not negative numbers.
- If negative numbers are part of the fitted distribution, use the 2nd highest PPCC or W distribution instead.

10

10

Example: Distributional Method

Benz.rda dataset 269 observations of Benzene 4% NDs
9 detection limits

Yes it's a large dataset but I'll use it anyway to illustrate the process.

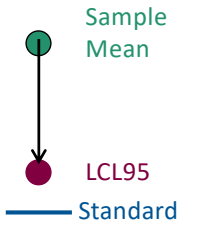
Water quality standard: 0.5 ppb

Null hypothesis: Assume compliance (innocence) until it is shown that the mean of observations exceeds the standard.

Test statistic -- compute the LCL.

Decision -- If the LCL exceeds the standard, non-compliance is evident.

If the LCL does not exceed the standard, compliance is not disproven.



Reject H_0. Not in compliance

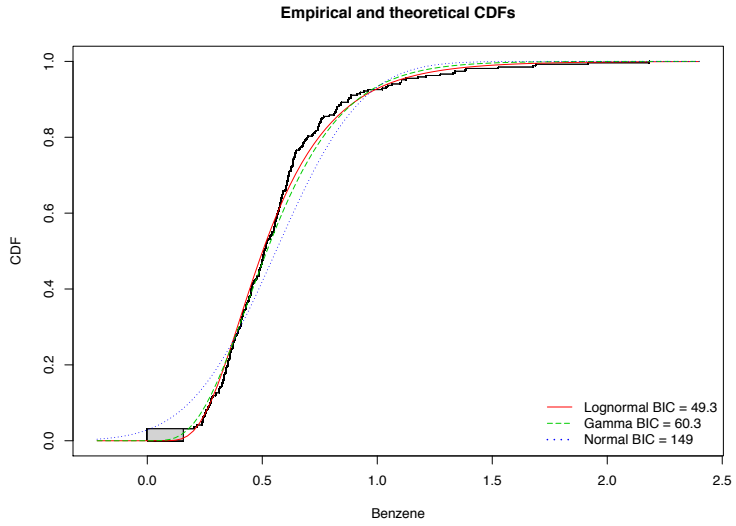
11

11

Step 2. Distributional Method

```
> attach(Benz)
> cenCompareCdfs (Benzene, BenzCen)
```

Lowest BIC distribution is lognormal



12

12

Step 2. Distributional Method
Computing the LCL using MLE

```
> elnormAltCensored(Benzene, BenzCen, ci = TRUE, ci.type = "lower")

Results of Distribution Parameter Estimation
Based on Type I Censored Data
--------------------------------------------
Assumed Distribution:           Lognormal
Censoring Side:                 left
Censoring Level(s):             0.1565 0.1693 0.1757 0.1853 0.1949 0.2045 0.2268 0.2428 0.2620
Estimated Parameter(s):         mean = 0.5606636
                                cv   = 0.4987887
Estimation Method:              MLE                    Confidence Level:   95%
Data:                           Benzene                Confidence Interval for:  mean
Censoring Variable:             BenzCen                Conf Int Method: Profile Likelihood
Sample Size:                    269                    Confidence Interval:  LCL = 0.5339487
Percent Censored:               3.717472%                                    UCL =       Inf
Confidence Interval Type:       lower
```

Exceeds std of 0.50. Out of compliance

13

13

Step 2. Distributional Method
Computing the LCL using rROS

```
> elnormAltCensored(Benzene, BenzCen, ci = TRUE, ci.type = "lower", method = "rROS", ci.method =
    "bootstrap")

Results of Distribution Parameter Estimation
Based on Type I Censored Data
--------------------------------------------
Assumed Distribution:           Lognormal
Censoring Side:                 left
Censoring Level(s):             0.1565 0.1693 0.1757 0.1853 0.1949 0.2045 0.2268 0.2428 0.2620
Estimated Parameter(s):         mean = 0.5623029
                                cv   = 0.5165693
Estimation Method:              rROS                   Confidence Level:   95%
Data:                           Benzene                Confidence Interval for:  mean
Censoring Variable:             BenzCen                Conf Int Method: Bootstrap
Sample Size:                    269                    Confidence Interval: BCa.LCL = 0.5355690
Percent Censored:               3.717472%
Confidence Interval Type:       lower
```

Exceeds std of 0.50. Out of compliance

14

14

Commands to compute mean and CI with the best distribution

`elnormAltCensored`	lognormal	method="rROS"
`egammaAltCensored`	gamma	
`enormCensored`	normal	

15

15

Flowchart for comparing to standards

Step 3. Nonparametric K-M Method for large datasets
Use bootstrapping to compute the UCL / LCL.

Bootstrap the one-sided limit using the enparCensored command (K-M estimate) when there is more than 1 RL. Use at least 5000 repetitions.

With only one reporting limit, the K-M estimates are biased high. Use the robust ROS procedure instead, assuming a lognormal distribution for the nondetects. This is the elnormAltCensored command, using the method = "rROS" option.

16

16

Step 3. Distributional Method

Computing the LCL

```
> enparCensored(Benzene, BenzCen, ci=TRUE, ci.type = "lower", ci.method = "bootstrap", n.bootstraps
    = 5000)
Results of Distribution Parameter Estimation Based on Type I Censored Data
--------------------------------------------
Assumed Distribution:           None                    Estimation Method: Kaplan-Meier
Censoring Side:                 left
Censoring Level(s):             0.1565 0.1693 0.1757 0.1853 0.1949 0.2045 0.2268 0.2428 0.2620
Estimated Parameter(s):         mean    = 0.56122798
                                sd      = 0.29136678
                                se.mean = 0.01717623
Sample Size:                    269                     Conf Interval:  Pct.LCL = 0.5327565
Percent Censored:               3.717472%                               Pct.UCL = Inf
Confidence Interval for:        mean                                    BCa.LCL = 0.5331886
Confidence Interval Method:     Bootstrap                               BCa.UCL =       Inf
Number of Bootstraps:           5000                                    t.LCL   = 0.5336898
Confidence Interval Type:       95% lower                               t.UCL   =       Inf
```

Exceeds std of 0.50.
Out of compliance

17

17

Summary: Comparing data with nondetects to standards

1. Decide whether to assume compliance (use LCL), or assume non-compliance (use UCL).
2. For n ≦20, assume the best-fit distribution (MLE or rROS) and compute the UCL / LCL using that distribution.
3. For n >20, bootstrap the Kaplan-Meier (multiple RLs) or rROS (one RL) estimates of the UCL / LCL.

For n>100, the t-interval or bootstrap with Kaplan-Meier or rROS estimates of mean and std dev, or the best-fitting distribution, should all give similar results. Use any of these.

For n < 70 -100, the t-interval you learned in college is appropriate only when data follow a normal distribution, which is very unlikely for data with nondetects.

18

18

Exercise

The exercise in the handouts has a 2nd and 3rd part that is very interesting.

2nd How to compute these statistics when no detection limit is given? (nondetects only designated as “ND” or “trace” or “0”.

3rd How to compute the probability of getting a value above the detection limit out in the field when 100% of data to date are nondetects? Hint: there’s a lot of information there. Don’t miss it!

The writeup in the Solutions is a sufficient explanation of what to do.

19

19