Seven Perilous Errors in Environmental Statistics

Dennis R. Helsel

= Practical Stats

www.practicalstats.com

Seven Perilous Errors in Environmental Statistics

The 7 Most Repeated Errors I've seen in my 37 year career:

1. A significant p-value tells you all you need to know

---

2. Testing for a normal distribution to decide whether to employ a parametric or nonparametric test
3. Using-tests and ANOVA with small data sets
4. Testing logarithms to look for differences in means

"the Flowchart"

---

5. Using only r-squared to find the best regression equation
6. Using outlier tests to find and delete 'bad' data
7. Substituting one-half the detection limit for nondetects

Error#1. A significant p-value tells you all you need to know

- People don't understand what a p-value is telling them
- Statistical significance is NOT the same as practical significance (usefulness)

3

A p-value is…..

- The probability of seeing a signal (difference in means, correlation, trend, etc.) when there is no signal in the real world
- The probability of a 'false positive' signal

A low p-value (<0.05) does not prove that there is a signal. It only says that it is likely.

Its primary use: to help people do what is hardest for them – to make a decision. Which way do the data indicate – signal or no signal?

4

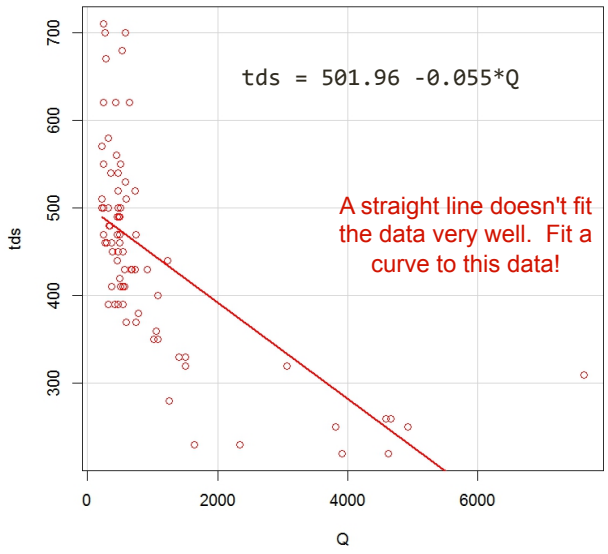
Statistical significance is not the same as a good regression model

```
tds = 501.96 -0.055*Q
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 501.958369  12.488199  40.195  < 2e-16 ***
Q            -0.054831   0.007473  -7.337 1.79e-10 ***
---
Residual standard error: 88.91 on 78 degrees of freedom
Multiple R-squared:  0.4083,Adjusted R-squared:  0.4008
F-statistic: 53.83 on 1 and 78 DF,  p-value: 1.788e-10
```

With this small of a p-value, isn't this a great regression model?

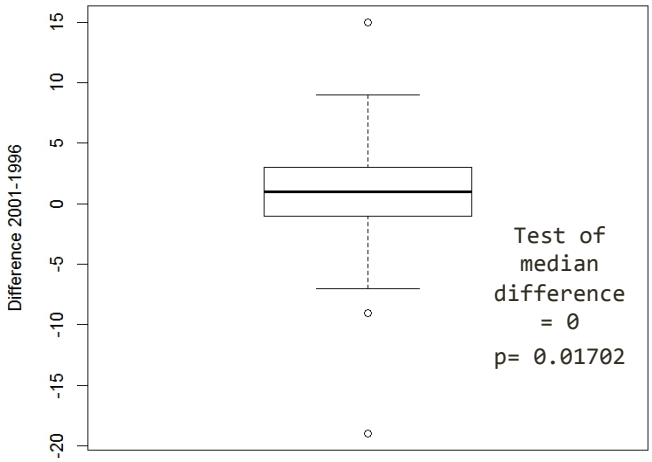
5

This is a terrible regression model !



6

Statistical significance is not the same as usefulness



Median difference (=1) is not zero, but small enough that we don't care about it. Instead of just the p-value, also report the 95% CI (0.1 to 2.0 ug/L) for the magnitude of difference

7

Summary for Error #1. A significant p-value tells you all you need to know

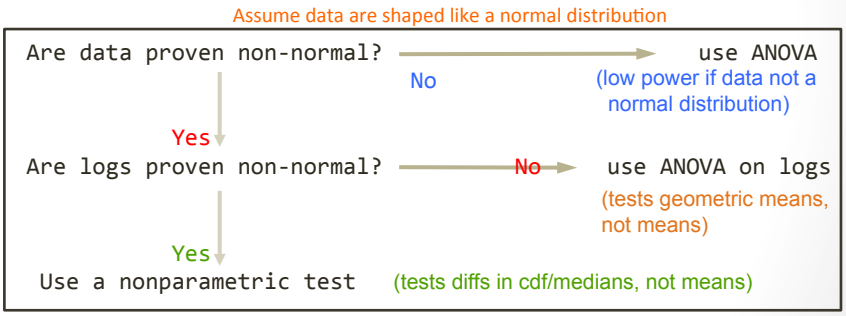
- Understand what a p-value is: the probability of a false signal / difference between groups/ trend
- A small p-value in regression doesn't mean that you have a wonderful equation. Plot the data!
- Statistical significance is NOT the same as usefulness. Always look at the magnitude of the difference after a statistical test indicates it is non-zero. That difference may be small enough to be of no practical interest.

8

The "Test for a Normal Distribution" Flowchart:

The Source of the Next 3 Errors

- Old-fashioned guidance docs start by assuming data follow a normal distribution, or that it doesn't matter
- The parametric test for means may have low power (low ability to see differences) – it DOES matter
- There are better tests for difference in means – permutation tests



9

Error#2. Testing for a normal distribution to decide whether to employ a parametric or nonparametric test

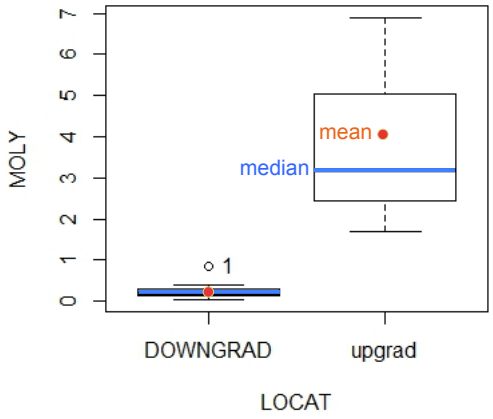
Instead, decide which type of test to run based on your objectives – do you want to test a measure of total amount (mean), or typical values and patterns (median)?

-- The tests on mean and median answer two totally different questions

-- It doesn't make sense to decide which question to answer based on the shape of the data

10

Test on means answers whether the totals in the two groups differ. Test on medians answers whether one group typically is higher than the other.



All data in the upgrad group are higher than all data in DOWNGRAD group!

Mean= 4 vs ¼ certainly looks different!
But t-test p=0.14, so means not found different.

Median = 3.2 vs 0.2.
Rank-sum test p=0.01, so medians significantly differ – one group is frequently higher than the other.

11

An Alternate, and Better, Decision Tree

- What is your objective?

If you are interested in

Mass/Totals chronic effects	Frequency/Typical patterns acute effects
Test means using permutation tests	Test percentiles using nonparametric (rank-based) tests

12

A Second Concern: Power: the ability of tests to find a signal

- Older, parametric tests have low power whenever data have outliers, or are skewed, or groups have different variability.
- Field data in environmental sciences usually have all three characteristics. So ANOVA, t-tests, and t confidence intervals don't work well for the type of data we usually encounter.

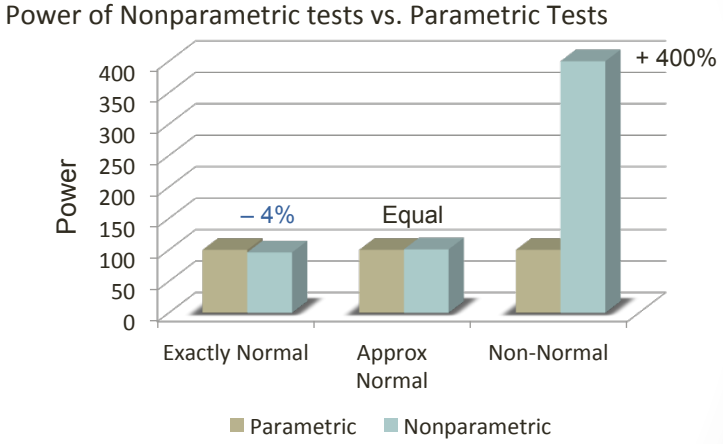
Alternatives:

Permutation tests. Do not assume a normal distribution. Not bothered by outliers. Can test for differences in means.

Nonparametric tests. Do not assume a normal distribution. Not bothered by outliers or changing variance. Test for differences in percentiles (typical patterns).

13

Assuming data follow a normal distribution unless proven otherwise is a bad start



Remember: field data almost never follow a normal distribution

14

Choose the type of test based on your objective, not the shape of the data!

- Are concentrations higher in one group than the other group?
  A Question of Frequency. Use nonparametric test
- Have concentrations increased over time?
  A Question of Frequency. Use nonparametric test
- The typically occurring concentration is __
  A Question of Frequency. Use percentile (median)
- What is the total amount washed into the estuary this year?
  A Question of Mass. Use the mean. For hypothesis tests, means can be tested without assuming normality using permutation tests

15

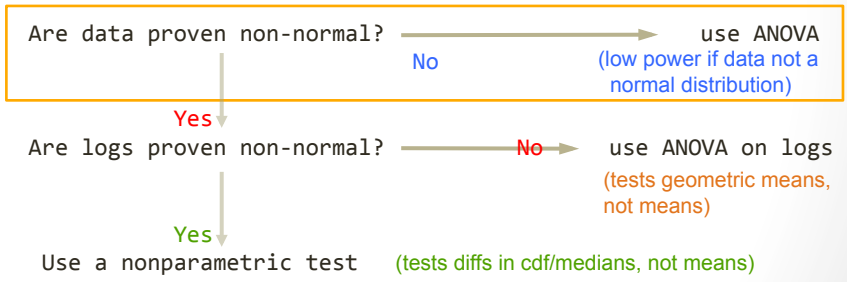
Summary for Error #2. Testing for a normal distribution to decide whether to employ a parametric or nonparametric test

- Field data rarely if ever follow a normal distribution
- For skewed data with outliers, nonparametric methods have a large power advantage (will find more differences when they are there) than do parametric tests
- Tests on means vs tests on medians answer different questions. Decide which question you want answered, and use that type of test

16

Error#3. Using t-tests and ANOVA with small data sets

- ANOVA and t-tests have low power for small data sets
- 'Small' is < ~70 observations per group
- There are better tests for difference in means – permutation tests

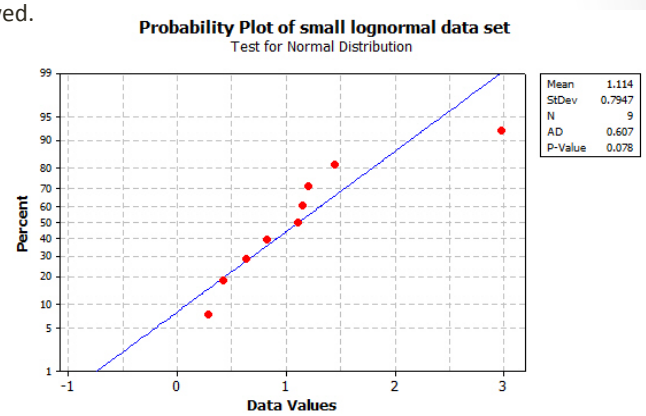


17

First: Not disproving normality DOES NOT MEAN that data follow a normal distribution

Tests for normality have low power to see non-normality with small data sets. So less-powerful parametric tests will be used on non-normal data if the flowchart is followed.

Generated lognormal data. n=9, p > 0.05 Cannot reject an assumed normal distribution, even though data do not follow a normal distribution



18

What if I want to test for differences in the mean, but my data are skewed?

- Testing / interpreting the mean when data are skewed has been one of the most difficult issues in statistics for decades
- In past, people assumed it didn't matter and ran the parametric test anyway. This was based on the large-sample properties (n>70 per group) called the Central Limit Theorem, which when invoked for n<70 is "hope"
- In the 1990s/2000s the problem of skewness and outliers was largely solved with the development of permutation tests
- Permutation tests can test hypotheses about means (and other statistics) without assuming a normal distribution

19

Permutation Tests

- Make no distributional assumptions about the population sampled. (Does not require assumption of normality)
- Do not rely on the Central Limit Theorem
- Use only the observed data and all possible rearrangements or permutations of the data
- Are still affected by unequal variance, but in the same way that the null hypothesis is itself

20

Is the Mean of Above the same as Below?

If so, the group assignment is arbitrary. Data in both groups are just noise around the same mean.

CONC	SITE
6	Above
5	Above
10	Above
16	Below
8	Below
22	Below
18	Below

21

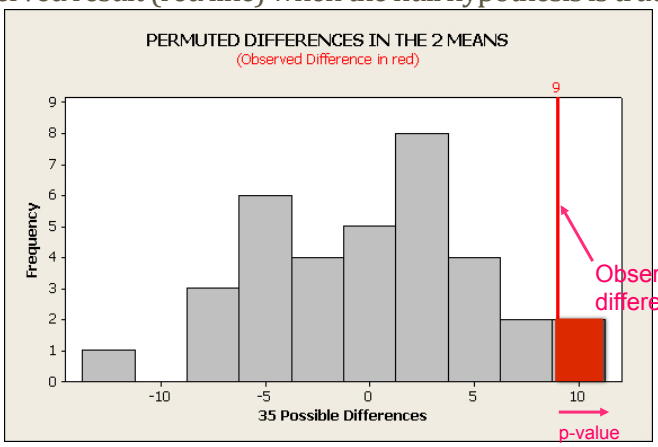
Shuffle the group names, either all possible arrangements or thousands of times, and compute the test result for each shuffle. This represents the "null hypothesis" (no difference) situation.

Concentration	Location
6	Below
5	Above
10	Below
16	Above
8	Below
22	Above
18	Below

← One example of a shuffle

22

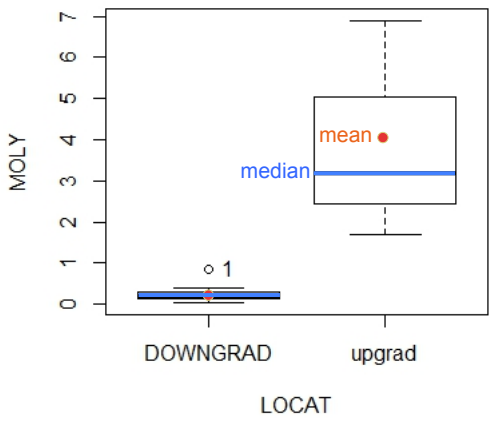
The bars are shuffled differences in the means, representing the null hypothesis. The p-value is the probability of getting the observed result (red line) when the null hypothesis is true



If the histogram bars for a t-test do not follow a t-distribution, the p-value for the t-test is incorrect. The permutation p-value (red area) always holds true regardless of the shape of the bars.

23

The means weren't found to significantly differ using the t-test, but are with the permutation test.



Mean= 4 vs ¼
The t-test p=0.14, so means not seen as different.

The permutation p=0.0018 for the same data, so means do differ when a test with more power is used.

The difference in test results measures how much is lost by assuming grouped data follow a normal distribution with equal variance (t-test), when they do not.

24

Error #3. Using t-tests and ANOVA with small data sets

Summary:

ANOVA and t-tests are only accurate when data follow a normal distribution. Or with 70+ observations per group.

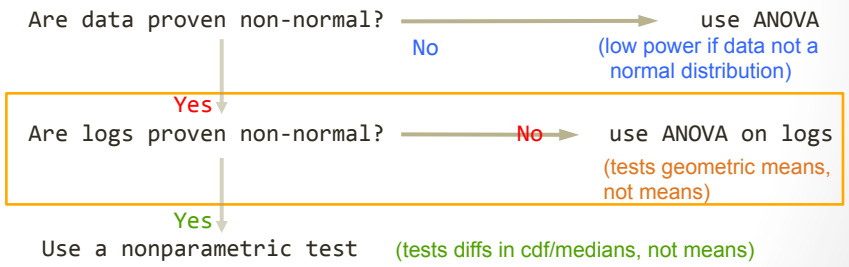
When a test for difference in means is the objective, a nonparametric test doesn't help. It tests for differences in percentiles (frequencies).

Test for differences in means using a permutation test. Older normal-theory tests were only approximations, which are no longer necessary.

25

Error #4. Testing logarithms to look for differences in means

With the flowchart, when data are non-normal, logs are often computed and the mean of the logs tested using a t-test or ANOVA.



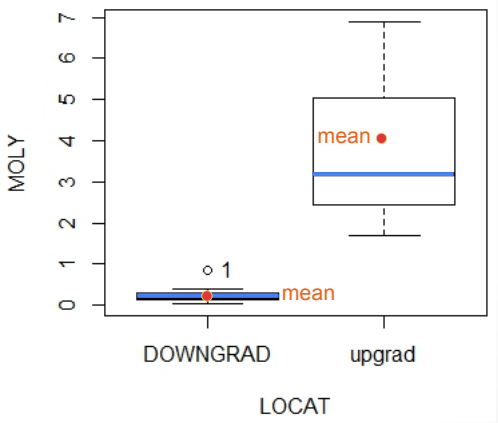
26

For skewed data, the mean and median are not the same

t-test: p=0.14
Means cannot be seen to significantly differ.

Is non-normality lowering the power and confusing the t-test?

If so, should we take logs and try again! ???

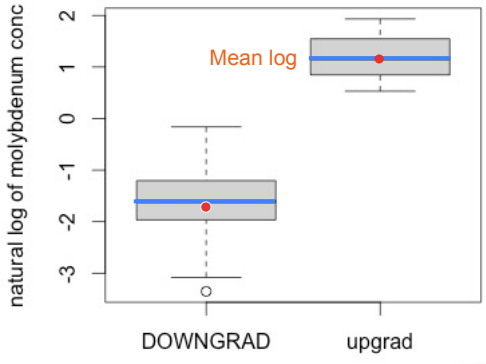


27

Testing in log units changes what is being tested

- The means of the logs are significantly different by the t-test. In original units, this tests for difference in geometric means!

t-test on logs:
p=0.005.
The mean logs (geometric means in original units) do significantly differ.



28

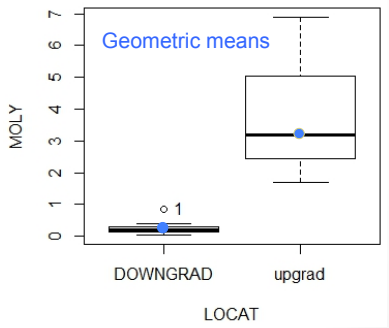
When a t-test is performed in log units.....

The null hypothesis of:

Mean log(Gp2) – Mean log(Gp1) = 0 in log units

$$\frac{\text{Geometric mean (Gp2)}}{\text{Geometric mean (Gp1)}} = 1$$

in original units



29

Error #4. Testing logarithms to look for differences in means

Summary:

- Testing differences between group means in log units (with ANOVA or t-tests) does NOT test for differences in means in original units
- The test in log units determines if geometric means differ
- The geometric mean is an estimate of the median, not the mean
- If you want to test differences between means of non-normal data, use a permutation test

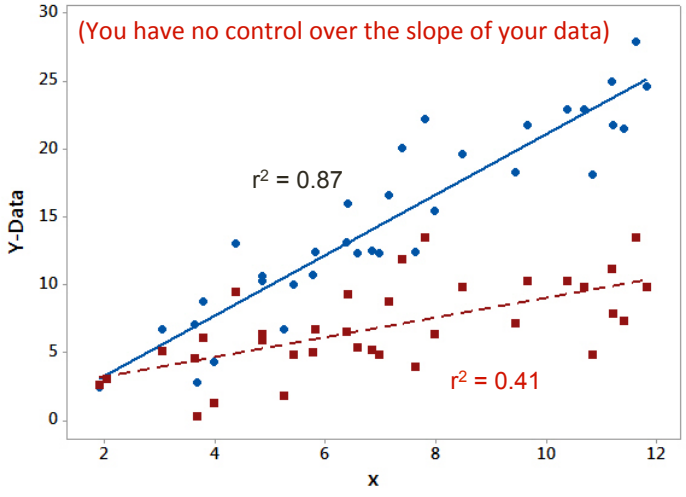
30

Error #5. Using only r-squared to find the best regression equation

- Many people maximize r-squared to get their "best" regression equation
- r^2 is in units of the y-variable. Changing the units (log etc) of y puts the statistic in a different set of units. Cannot directly compare it to r^2 in the original units.
- r^2 depends on the slope. Higher slope = higher r^2, all else being equal. We have no control over slopes
- Better numerical criteria are available for determining the best regression equation

31

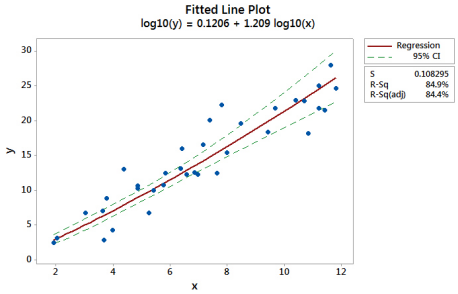
Two lines with identical variation, differing slopes. Is the upper equation better?



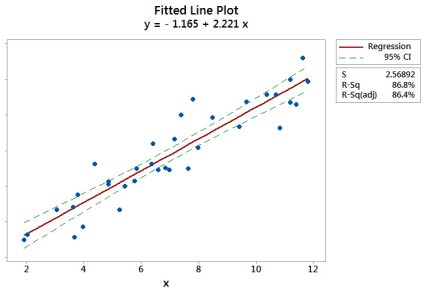
32

Explaining 85% of variance in log units may be better than 87% in original units

r^2 = 85% of log Y



r^2 = 87% of Y



Lower r^2 for the left (log) plot, but if variation increases for larger x (and this often occurs in the natural world), it is a more accurate model

33

AIC and other "cost-benefit" statistics better describe the quality of a regression equation

$$AIC = 2p - 2\ln(L)$$

 # parameters. Improves fit, but decreases df. Cost.

Unexplained noise, as expressed by the log likelihood L. Smaller is better. Benefit.

34

Smallest AIC, PRESS, Cp is best. Highest r^2 not necessarily so

to predict lnSO2

Vars	R-Sq	AIC	PRESS	Mallows Cp	S	M P N R T F W E D E C P I C A M T O N I Y P R P D P S w
1	29.9	77.8	15.1	31.9	0.59566	X
1	23.5	81.4	16.4	38.2	0.62226	X
2	44.9	69.9	12.2	19.2	0.53495	X X
2	41.0	72.8	13.2	23.0	0.55364	X X
3	53.6	64.8	11.1	12.6	0.49721	X X X
3	51.8	66.4	11.7	14.3	0.50679	X X X
4	63.0	57.6	9.2	5.3	0.45010	X X X X
4	59.9	60.9	9.9	8.4	0.46890	X X X X
5	65.3	56.8	9.4	5.0	0.44160	X X X X X
5	63.0	59.6	10.0	7.3	0.45646	X X X X X
6	65.5	58.8	10.2	7.0	0.44800	X X X X X X

35

Error #5 Summary. Using r^2 to find the 'best' regression equation isn't best

- r^2 is in units of the y-variable. Changing the units (log of y, etc) puts the statistic in a different set of units. Cannot directly compare it to the r^2 in original units.
- r^2 depends on the slope. Higher slope = higher r^2, all else being equal. We have no control over slopes
- Better, more modern numerical criteria are available for determining the best regression equation

36

Error #6. Using outlier tests to find 'bad' data

- Dixon and Rossner's tests are the most common
- Dixon's test determines whether the single max or min value is an outlier
- Rossner's test determines whether there is a group of outliers. You must specify the number of outliers in the group prior to testing
- These tests determine the likelihood of that observation occuring if data followed a normal distribution
- The big issue: that's not what many people are using the test for!

37

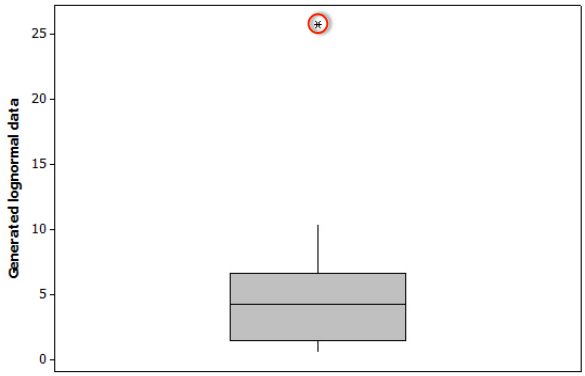
Three Major Causes of Outliers

1. An error in measurement. If you can determine there was an error, drop the measurement.
2. "Contamination" from another population. If the outlier represents conditions you do not wish to describe, drop the outlier or include it in a second, separately described population.
3. Skewed distributions. Adsorption, diffusion and other natural processes lead to skewed data being common. Most data from the natural world follow skewed, not normal, distributions

38

Dixon's test on lognormal data

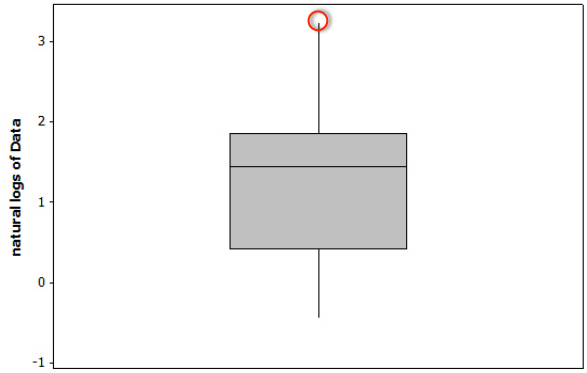
- Generated from a lognormal distribution
- Significant outlier for the highest observation (data are not from a normal distribution)



39

Dixon's test on same data after taking logs

- Now the upper value is not an outlier
- Is the same observation no longer "wrong"??



40

Outlier……..s Deal with Them!

- Deleting outliers is a science decision. It is reasonable for occasional wild values from a stable (say electronic) process. It is not reasonable for field data collected by a representative sampling method
- Field data will contain values that are 'outliers' in comparison to a normal distribution, because they are usually skewed and not from a normal distribution.
- A statistical test cannot tell you whether the observation is 'bad' or not. It only tells you whether it was likely to have come from a normal distribution. You already know that field data are not likely from a normal distribution.

41

Outlier……..s Deal with Them!

- A Better Option: Use methods that don't require a specific distributional shape
  - The primary reason people have run outlier tests is to 'normalize' data prior to running a parametric test
  - Normality is not required for either nonparametric or permutation tests. Use those tests instead.
  - Don't remove data in order to use approximate tests developed in the 1930-40s. Use more modern methods.

42

Outliers are often the most valuable observations

Outliers can tell you:

- Different conditions were used
- An unusual event happened. Infrequent conditions (floods, etc.) are very important -- though they may be considered from a different population

Suppose you are collecting samples of rock and measuring gold content. In one or two samples the content is unusually high – you hit a vein. Are you going to throw these data away because they're not like the others?

43

Summary for Error #6. Outlier Tests

- Outlier tests cannot tell you whether data are 'wrong', only that they aren't likely to have come from a normal distribution
- Environmental field data (water, air, soils, rock, biota) are usually skewed distributions, not originating from the normal distribution. There are physical reasons for this. Outliers are to be expected.
- Albert Einstein was an outlier

44

Error #7. Substituting one-half the detection limit for nondetects

What's wrong with substitution?

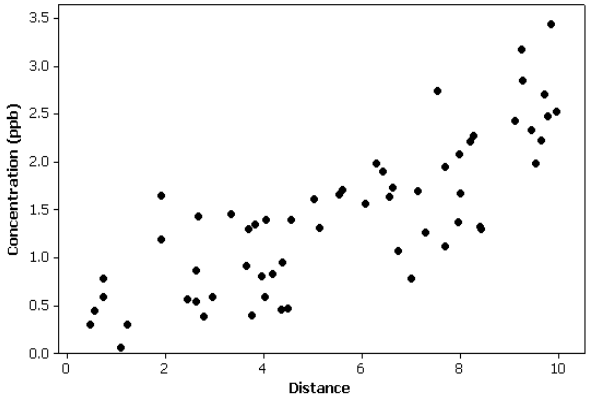
- Substitution is not neutral
- Produces invasive data alien to the concentrations actually in samples
- Substituting a constant always results in a poor estimate of std dev, and flat line
- Results in poor estimates and incorrect statistical tests
- There are much better alternative methods

45

Example 1 of what's wrong with substitution

Correlation and Regression

Before censoring. True correlation r=0.81

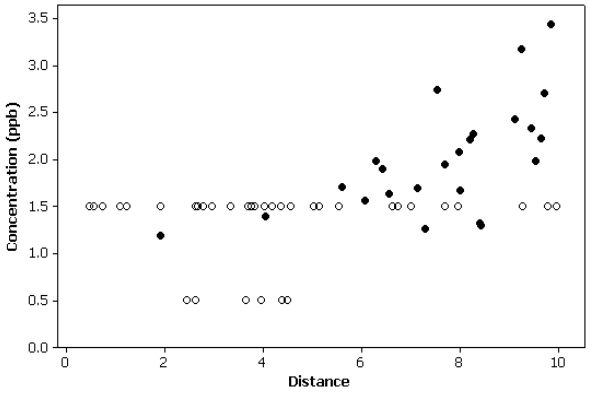


46

Example 1 of what's wrong with substitution

Correlation and Regression

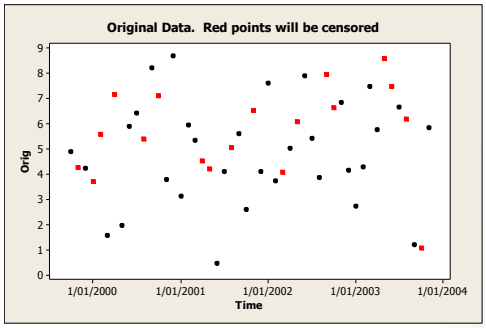
Some of the previous data are now declared to be nondetects. After substitution. invasive data form flat lines, lowering correlation to r=0.55



47

Example 2: finding a trend that isn't there

- True situation: No change over time
- Red dots become nondetects



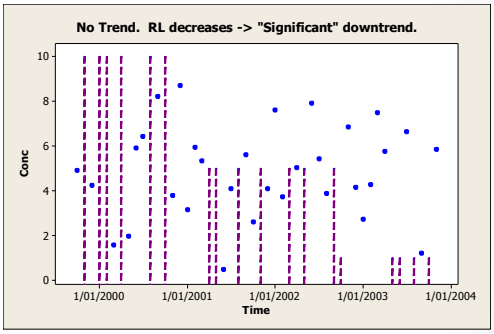
48

48

Example 2: finding a trend that isn't there

Invasive pattern

- Detection limits decrease over time
- By substituting, you put in a downtrend that wasn't there in the original data
- A correlation with time (trend) becomes significant



How long have we known this?

Gilliom and Helsel (1986)

- Compared substitution to other methods for estimating means, medians, std dev, percentiles
- Found that the other methods were generally better than substitution
- One-half DL gave reasonable estimates for the mean with one DL, but not other statistics, and not with multiple limits
- For example, the bias of subbing 1/2DL for estimating the median was about 4.5 times that for a probability plot method

Recent USEPA guidance documents do not recommend substitution

- Singh et al (2006), developers of the ProUCL software, determined that substituting ½ DL "does not provide adequate coverage [UCL95 is not high enough] …even for censoring levels as low as 10%"
- They summarize their results with "Do not use DL/2 (t) method to compute a UCL".
- Recommended methods were based on a Kaplan-Meier estimator

51

Many other papers point out these problems:

- Thompson and Nelson (2003) found that for censored response (y) variables in regression, substituting one-half the DL for nondetects produced
  1. biased parameter estimates (slopes too low) and
  2. artificially small standard error estimates (slopes artificially significant, so variables included in the equation that should not be).

Why do people continue? It is easy and cheap --- until you understand the consequences.

52

There are Three Better Approaches

1. Binary methods

   Simple. Data are either below or above a specified, single limit. Report % above, test percentages, logistic regression.

2. Nonparametric methods

   Simple. Rank all data below highest RL as tied. Report percentiles, NP tests, tau correlation coeff.

3. Survival Analysis methods

   More complicated. Can use data below multiple DLs. See *Statistics for Censored Environmental Data using Minitab and R* (Helsel, 2012)

53

Error #7. Substituting ½ the detection limit for nondetects works fine

Summary: No it doesn't!

Substitution produces an invasive, false signal (or false no-signal). Fortunately, there are several better methods available.

54

Seven Perilous Errors in Environmental Statistics

- You can find a pdf of these slides on the Practical Stats site, at:

  http://practicalstats.com/downloads/

55

Resources – more info

- Newsletters
  http://www.PracticalStats.com/news/
- Webinars and Classes
  http://www.PracticalStats.com/training/
- Textbooks http://practicalstats.com/books/

  Statistical Methods in Water Resources (Helsel & Hirsch, 2002)
  a new edition coming in 2016

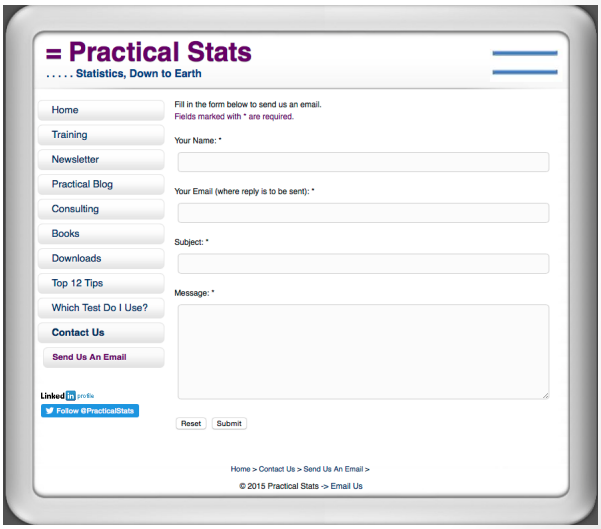
  Statistical Methods for Censored Environmental Data using Minitab and R Helsel (2012)

56

For more information:

Email using the Contact Us page at PracticalStats.com

Thank you for attending today!



57