



Module 10: Censored Data

10.1 Methods for Censored Data



Censored Data Analysis

- ◆ Censored data are the result of an analytical chemistry practice of deciding whether or not a measured value can be considered to be different from zero
- ◆ If not, the data is reported as non-detect or as below a calculated limit of detection (LOD)
- ◆ Censored data values create statistical difficulties since there are no numbers to use in statistical calculations

Module 10.1





Censored Data Analysis

- ◆ Some common censored data practices:
 - Ignore them (biases mean high)
 - Replace them with some constant
 - zero (biases mean low)
 - limit of detection (biases mean high)
 - half of the limit of detection

Module 10.1




Censored Data Analysis

- Replace them with
 - Randomly generated data from a $\text{uniform}(0, \text{LOD})$
- Assume a distributional type and use theoretical results to give maximum likelihood estimates of the distribution's parameters.
 - Works very well if your assumption is right and poorly if it's wrong.
 - Requires specialized software


Module 10.1






Censored Data Analysis


- Regression on order statistics
 - Fit a regression line on a normal probability plot of the uncensored data (using censored values as place holders) and estimate the model parameters from the line coefficients
- Fill-in methods
 - Estimate the parameters of the distribution and then set the censored data equal to their expected values
 - Usually iterative

Module 10.1 



Censored Data Analysis

- Robust parametric method
 - Construct a normal probability plot
 - (Transform to normality if necessary)
 - Fit a line by linear regression
 - Use the line to calculate expected values for the censored points
 - Replace the censored data with their expected values
 - Use all of the data, including both uncensored and the new values, to calculate the statistics

Module 10.1 

Censored Data Analysis

◆ Example

11.7	<5
9.3	16.4
10.4	11.9
9.6	9.1
<5	12.0
7.3	<5
14.5	15.8
10.9	<5
13.4	8.3
8.0	23.7


Module 10.1



Original Data	Ignore	Replace w/ zero	Replace w/ LOD	Replace w/ LOD/2	Random Uniform
11.7	11.7	11.7	11.7	11.7	11.7
9.3	9.3	9.3	9.3	9.3	9.3
10.4	10.4	10.4	10.4	10.4	10.4
9.6	9.6	9.6	9.6	9.6	9.6
<5		0.0	5.0	2.5	1.9100
7.3	7.3	7.3	7.3	7.3	7.3
14.5	14.5	14.5	14.5	14.5	14.5
10.9	10.9	10.9	10.9	10.9	10.9
13.4	13.4	13.4	13.4	13.4	13.4
8.0	8.0	8.0	8.0	8.0	8.0
<5		0.0	5.0	2.5	0.5034
16.4	16.4	16.4	16.4	16.4	16.4
11.9	11.9	11.9	11.9	11.9	11.9
9.1	9.1	9.1	9.1	9.1	9.1
12.0	12.0	12.0	12.0	12.0	12.0
<5		0.0	5.0	2.5	2.9824
15.8	15.8	15.8	15.8	15.8	15.8
<5		0.0	5.0	2.5	4.4955
8.3	8.3	8.3	8.3	8.3	8.3
23.7	23.7	23.7	23.7	23.7	23.7
True Mean = 10	12.01	9.61	10.61	10.11	10.10
True S.D. = 4	4.13	6.14	4.66	5.36	5.41
LOD = 5					

Module 10.1






Regression on Order Statistics

- Rank the n data values from smallest to largest with the censored values as smallest
- Calculate $P_i = (i - 3/8)/(n + 1/4)$
- Find the Z_i scores that correspond to those probabilities using the =NORMSINV() function in Excel
- Construct a normal probability plot with the Z scores as the independent variable and the uncensored data values as the dependent variable

Module 10.1




Regression on Order Statistics

- Carry out a regression analysis
- Make sure the line is a good fit to the data
- The mean of the distribution is estimated by the Y- intercept and the standard deviation by the slope


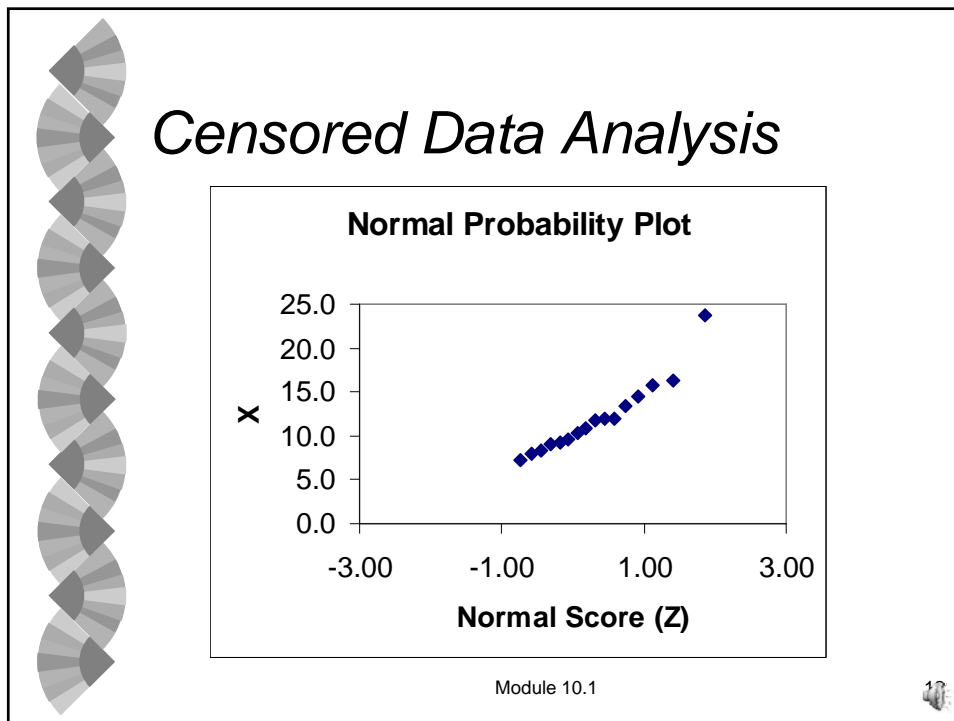
Module 10.1





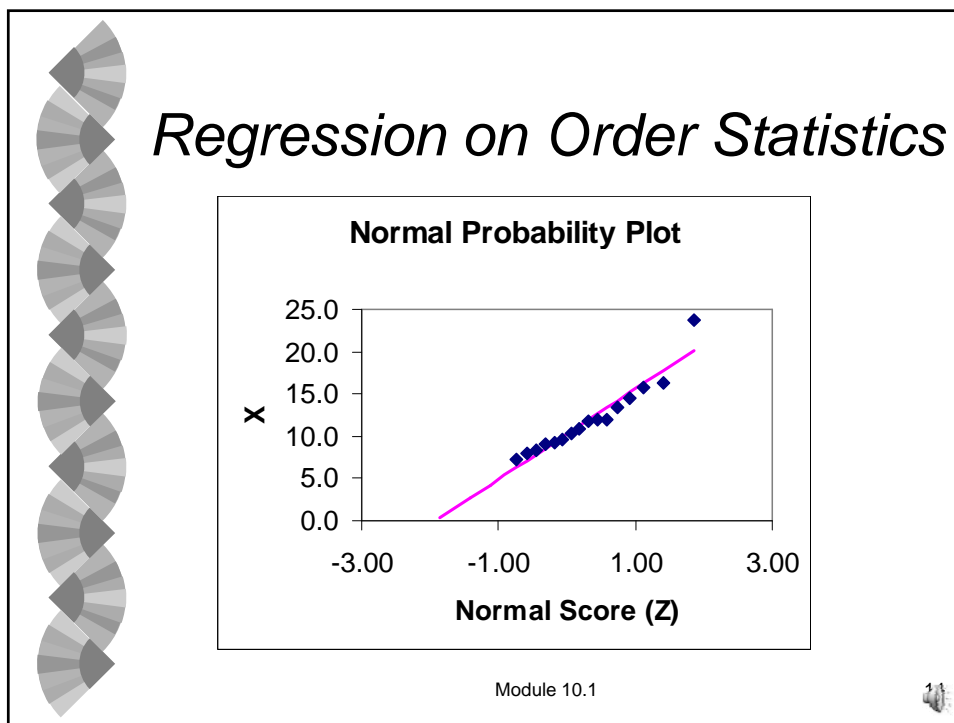
i	P_i	Z_i	X_i
1	0.03	-1.87	<5
2	0.08	-1.40	<5
3	0.13	-1.13	<5
4	0.18	-0.92	<5
5	0.23	-0.74	7.3
6	0.28	-0.59	8.0
7	0.33	-0.45	8.3
8	0.38	-0.31	9.1
9	0.43	-0.19	9.3
10	0.48	-0.06	9.6
11	0.52	0.06	10.4
12	0.57	0.19	10.9
13	0.62	0.31	11.7
14	0.67	0.45	11.9
15	0.72	0.59	12.0
16	0.77	0.74	13.4
17	0.82	0.92	14.5
18	0.87	1.13	15.8
19	0.92	1.40	16.4
20	0.97	1.87	23.7

Module 10.1

Regression on Order Statistics					
SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.9589039				
R Square	0.9194967			Estimate of Mean = 10.24	
Adjusted R	0.9137465			Estimate of Stan. Dev. = 5.33	
Standard E	1.2123278				
Observatic	16				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	235.02	235.02	159.91	4.761E-09
Residual	14	20.57634	1.469739		
Total	15	255.5963			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
Intercept	10.24	0.333851	30.67893	3E-14	
X Variable	5.33	0.421111	12.64539	5E-09	

Module 10.1



Robust Parametric

- ◆ Robust parametric method
 - Create a normal probability plot and fit a regression line as above
 - Use the regression coefficients to estimate values for the censored data
 - Use all of the data including the estimated values to calculate the statistics


Module 10.1




**Calculations for Regression on Order Statistics
and Robust Parametric Techniques**


I	P_i	Z_i	X_i	Fitted
1	0.03	-1.87	<5	0.29
2	0.08	-1.40	<5	2.77
3	0.13	-1.13	<5	4.23
4	0.18	-0.92	<5	5.35
5	0.23	-0.74	7.3	6.28
6	0.28	-0.59	8.0	7.10
7	0.33	-0.45	8.3	7.86
8	0.38	-0.31	9.1	8.57
9	0.43	-0.19	9.3	9.25
10	0.48	-0.06	9.6	9.91
11	0.52	0.06	10.4	10.57
12	0.57	0.19	10.9	11.24
13	0.62	0.31	11.7	11.92
14	0.67	0.45	11.9	12.63
15	0.72	0.59	12.0	13.38
16	0.77	0.74	13.4	14.20
17	0.82	0.92	14.5	15.14
18	0.87	1.13	15.8	16.25
19	0.92	1.40	16.4	17.72
20	0.97	1.87	23.7	20.19






	Robust Parametric	
11.7	11.7	
9.3	9.3	
10.4	10.4	
9.6	9.6	
<5	0.2936	
7.3	7.3	
14.5	14.5	
10.9	10.9	
13.4	13.4	
8.0	8.0	
<5	4.2347	
16.4	16.4	
11.9	11.9	
9.1	9.1	
12.0	12.0	
<5	2.7688	
15.8	15.8	
<5	5.3477	
8.3	8.3	
23.7	23.7	
True Mean = 10	10.24	
True S.D. = 4	5.23	
LOD = 5		


Module 10.1 



Other Solutions


- ◆ Use the median and other percentiles instead of the mean and standard deviation
- ◆ The median can be calculated if less than 50% of the data values are censored
- ◆ Other useful statistics are the minimum, maximum, and other percentiles


Module 10.1 



Other Solutions


- ◆ Five number summary:
 - Minimum
 - 25th percentile
 - Median (50th percentile)
 - 75th percentile
 - Maximum


Module 10.1 



Other Solutions


- ◆ Five number summary (example):
 - Minimum = <5
 - 25th percentile = 7.84
 - Median (50th percentile) = 10.00
 - 75th percentile = 12.38
 - Maximum = 23.7
- ◆ (Linear interpolation was used to estimate the percentiles)


Module 10.1 



Conclusions

- ◆ Censored data pose some special statistical problems
- ◆ If a very simple method must be used, replacement with half of the detection limit is probably the best of a bad lot
- ◆ A better method is the robust parametric method

Module 10.1 



Conclusions

- ◆ Other solutions include
 - Don't censor the values in the first place
 - Use the median and percentiles instead of mean and standard deviation

Module 10.1 