2. Filling Data Gaps, Data validation & Descriptive Statistics

Dr. Prasad Modak

Environmental Management Centre, Mumbai

Background

- Data collected from field may suffer from these problems
  - Data may contain gaps (= no readings during this period)
  - Data may exhibit “suspect” values

- Addressing these deficiencies is essential before processing/ analyzing the data

How to fill in data gaps

- Identify data gaps
- Use any of the following techniques:
  - Use the average of the distribution to fill in data gaps
  - Taking mean of adjacent values
  - Linear interpolation
  - Using a “Regression Model”

- Remember that there is no substitute to real value
- Filling missing values creates a “bias” and “distortion” in the data when it comes to interpretation

Techniques for Filling Missing Values

- Suppose x_a and x_c are two values with value of x_b missing, then:

$$x_b = \frac{x_a + x_c}{2}$$

- Linear interpolation. Suppose x_a and x_b are the two adjacent values to n missing values. Then the k^{th} missing value (from x_a) will have the value:

$$x_k = x_a + k\frac{x_b - x_a}{n}$$

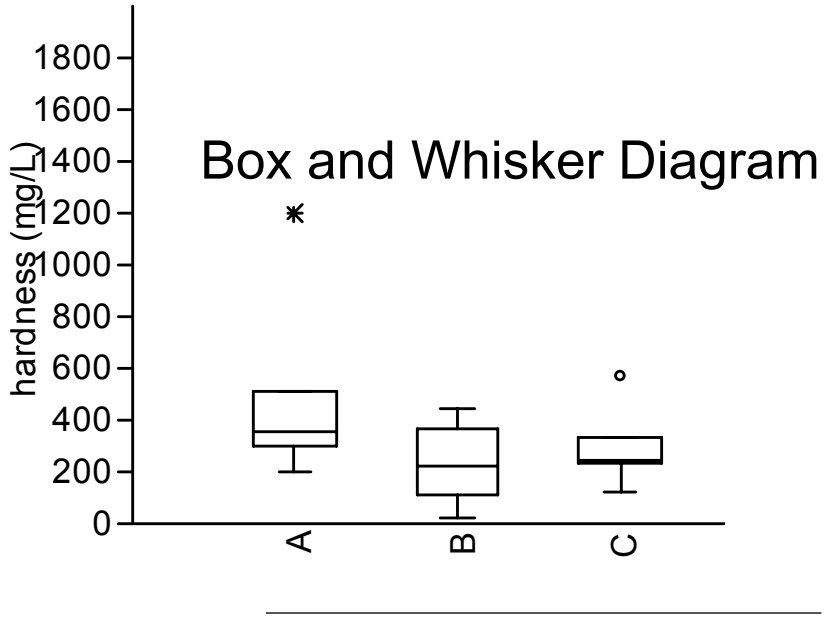
- If there m data points missing out of n and m << n, then assign to m, average value of the series.

Outliers

- What are outliers?
- How does the outliers affect analysis of data?
- How to find outliers
  - Dixon's 4 sigma (σ) test
  - Outliers are those values that lie beyond the mean ± 4 standard deviations
  - Box and whisker diagram
  - Use water quality standards (IS 10500:1991)

Date & Time	Hardness
3/26/2004 13:00	122.3
3/26/2004 14:00	178.6
3/26/2004 15:00	347.4
3/26/2004 18:00	368.3
3/26/2004 19:00	67942
3/26/2004 20:00	22175.7
3/26/2004 21:00	5875.2
3/26/2004 22:00	1840.4
3/26/2004 23:00	840.6
3/27/2004 0:00	643
3/27/2004 1:00	464.8
3/27/2004 2:00	275.2
3/27/2004 3:00	194.6
3/27/2004 4:00	168.2
3/27/2004 5:00	162.1
3/27/2004 6:00	162
3/27/2004 7:00	172.4
3/27/2004 8:00	298.8
3/27/2004 9:00	334
3/27/2004 10:00	398.3
3/27/2004 11:00	394.8
3/27/2004 12:00	480.9
μ	4719.98
σ	14893.7
μ+4σ	64294.8
μ-4σ	-54855
Q1	182.6
Q3	602.475
Max	67942
Min	122.3

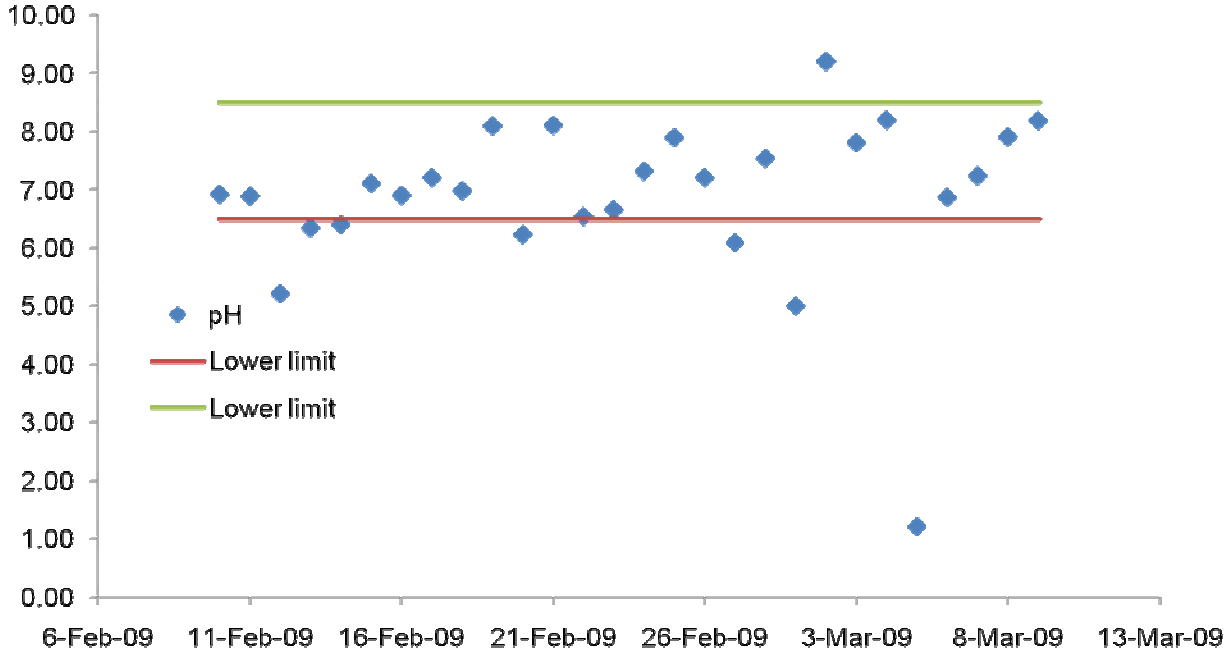
Dixons test



Time	A	B	C
t	313	202	245
t+1	356	220	123
t+2	298	312	332
t+3	1200	111	321
t+4	200	20	234
t+5	398	450	234
t+6	512	367	567

Control Chart

		as per IS 10500: 1991	
day	pH	Lower limit	Lower limit
10-Feb-09	6.92	6.5	8.5
11-Feb-09	6.89	6.5	8.5
12-Feb-09	5.21	6.5	8.5
13-Feb-09	6.34	6.5	8.5
14-Feb-09	6.40	6.5	8.5
15-Feb-09	7.11	6.5	8.5
16-Feb-09	6.90	6.5	8.5
17-Feb-09	7.21	6.5	8.5
18-Feb-09	6.98	6.5	8.5
19-Feb-09	8.10	6.5	8.5
20-Feb-09	6.23	6.5	8.5
21-Feb-09	8.11	6.5	8.5
22-Feb-09	6.54	6.5	8.5
23-Feb-09	6.66	6.5	8.5
24-Feb-09	7.32	6.5	8.5
25-Feb-09	7.90	6.5	8.5
26-Feb-09	7.21	6.5	8.5
27-Feb-09	6.09	6.5	8.5
28-Feb-09	7.54	6.5	8.5
1-Mar-09	5.00	6.5	8.5
2-Mar-09	9.21	6.5	8.5
3-Mar-09	7.81	6.5	8.5
4-Mar-09	8.20	6.5	8.5
5-Mar-09	1.20	6.5	8.5
6-Mar-09	6.87	6.5	8.5
7-Mar-09	7.24	6.5	8.5
8-Mar-09	7.91	6.5	8.5
9-Mar-09	8.19	6.5	8.5



Descriptive statistics

- No. of observations (n)
- No. of missing values
- Minimum
- Maximum
- Range
- 1st Quartile
- Median
- 3rd Quartile
- Sum (If relevant)

- Mean (μ)
- Standard error (σ^2)

- Standard deviation (σ)
- Variance (v)
- Skewness
- Kurtosis (k)

Arithmetical mean

- Mean is the arithmetic mean or the average of the data. It is calculated using the following formula:
- Mean = (Sum of data / Total number of observations)
- Mean is important, why?
  - Mean is the most used measure of a distribution
  - Other measures of dispersion (SE. SD, CV) are calculated based on mean
- Mean might not be always a very good measure of data, why?
  - Two distributions with same mean could have widely different ranges
  - Mean might not reflect the right attributes of the distribution (in case of highly tilted distributions)

Median, quartiles and percentiles

- a **quartile** is any of the three values which divide the sorted data set into four equal parts, so that each part represents one fourth of the "sampled population"
  - Q1 = first quartile (25% data points are under this value),
  - Q2 = second quartile (or median) (50% data points in sample is below this value)
  - Q3 = third quartile (75% data points in sample are below this value)
- Median is a real value located in the sample
- Sample mean and median could be different
- Similarly, percentiles divide the "population" into 100 equal unit
- So, median = 2^{nd} Quartile = 50^{th} Percentile

Median

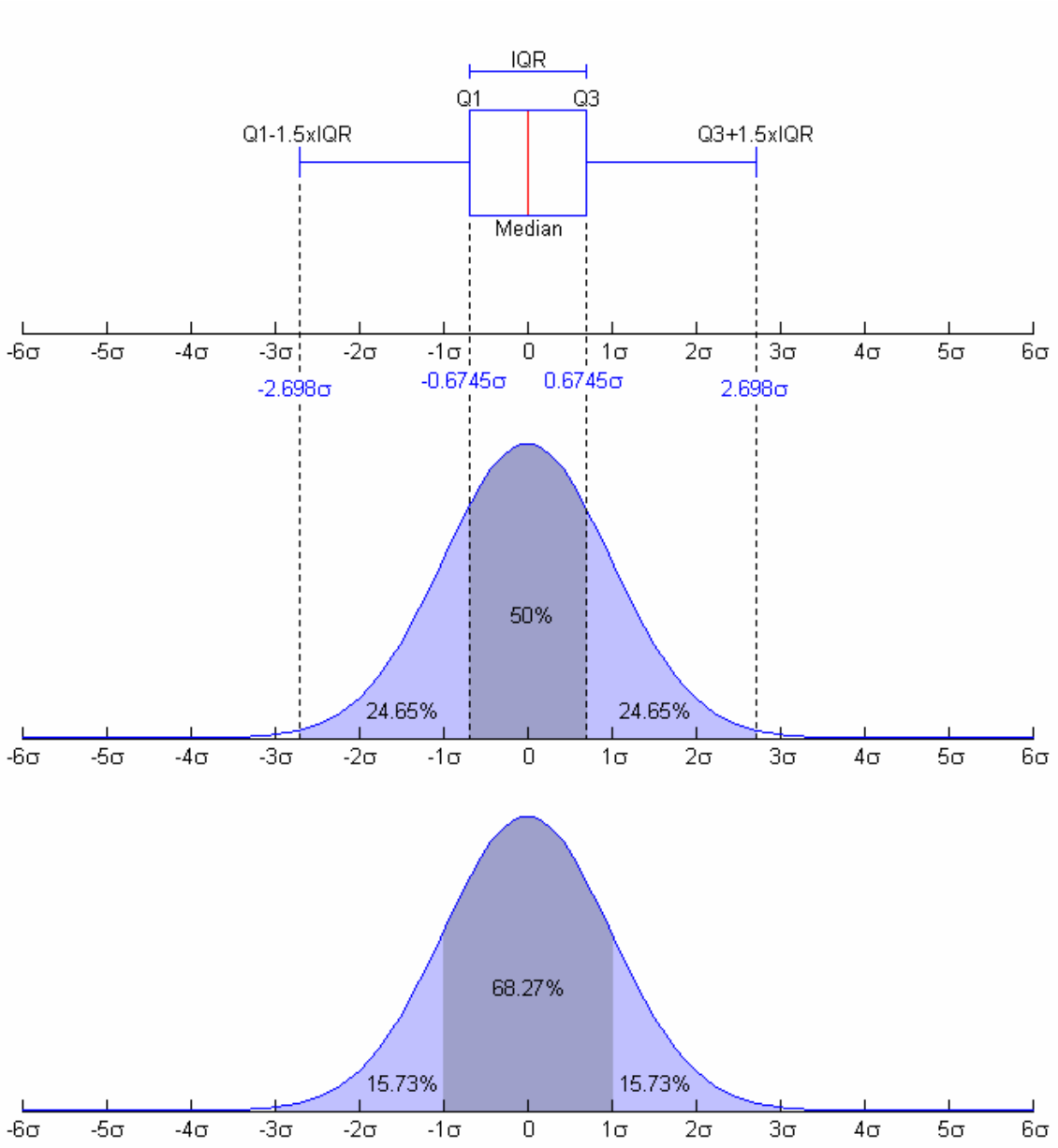
- Median
  - For even number of value (arranged in ascending order)

$$x_{med} = \frac{n}{2}$$

Where n is the number of values in the sample population

  - For odd number of values (arranged in ascending order)

$$x_{med} = \frac{n+1}{2}$$


IQR
Q1
Q3
Q1-1.5xIQR
Q3+1.5xIQR
Median
-6σ
-5σ
-4σ
-3σ
-2σ
-1σ
0
1σ
2σ
3σ
4σ
5σ
6σ
-2.698σ
-0.6745σ
0.6745σ
2.698σ
50%
24.65%
24.65%
68.27%
15.73%
15.73%

Standard Error and Standard Deviation

- A Standard Error (SE) is the estimate of variation of a statistics
- The SE of the mean tell about the spread of sample observations (x) about the mean (μ).
- The Standard Deviation (SD) is a widely used measure of the variability or dispersion along the mean
- It shows how much variation there is from the average or *measuring how far the data values lie from the mean*
- Chebyshev's rule: for any distribution, and for any positive k, the proportion of the data that lies within k nos. SD of the mean is at least:

$$p = 1 - \frac{1}{k^2}$$

http://www.ltcconline.net/greenl/courses/201/descstat/mean.htm

How to calculate SE & SD

- Sample standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

- Sample standard error of the mean

$$SE = \frac{\sigma}{\sqrt{n}}$$

Example

day	pH	x-μ	$(x-\mu)^2$
10-Feb-09	6.92	0.02	0.00
11-Feb-09	6.89	-0.01	0.00
12-Feb-09	5.21	-1.69	2.87
13-Feb-09	6.34	-0.56	0.32
14-Feb-09	6.40	-0.50	0.25
15-Feb-09	7.11	0.21	0.04
16-Feb-09	6.90	0.00	0.00
17-Feb-09	7.21	0.31	0.09
18-Feb-09	6.98	0.08	0.01
19-Feb-09	8.10	1.20	1.43
20-Feb-09	6.23	-0.67	0.45
21-Feb-09	8.11	1.21	1.46
22-Feb-09	6.54	-0.36	0.13
23-Feb-09	6.66	-0.24	0.06
24-Feb-09	7.32	0.42	0.17
25-Feb-09	7.90	1.00	0.99
26-Feb-09	7.21	0.31	0.09
27-Feb-09	6.09	-0.81	0.66
28-Feb-09	7.54	0.64	0.41
1-Mar-09	5.00	-1.90	3.62
2-Mar-09	9.21	2.31	5.32
3-Mar-09	7.81	0.91	0.82
4-Mar-09	8.20	1.30	1.68
5-Mar-09	1.20	-5.70	32.53
6-Mar-09	6.87	-0.03	0.00
7-Mar-09	7.24	0.34	0.11
8-Mar-09	7.91	1.01	1.01
9-Mar-09	8.19	1.29	1.66
			56.20
n = 28		Var	2.08
		SD	1.442732
		SE	0.272651

count (n)	28
Max	9.21
Min	1.2
Range	1.2-9.21
Mean	6.90
Q1	6.51
Median	7.05
Q3	7.83
Variance	2.08
Std. error	0.27
Std. dev	1.44

Discussion Points

- Mean, SE and Compliance with Standards
- Annual average DO concentration at a station was found to be 4.5 mg/l with SE of 1.0. Is the station “compliant”?
- SE and Number of Samples (Sampling Frequency)
- Sampling frequency was increased to weekly. What will happen now to annual average and SE?
- Coefficient of Variation (CV) = SD / Mean
- What will CV signify? What inferences could we draw?