Statistical Analyses of Stormwater Characterization and Control Data

Mostly excerpted from:

Bob Pitt
*University of Alabama*

---

Recommended Exploratory Data Analysis Reference Books

*Exploratory Data Analysis*. John W. Tukey. Addison-Wesley Publishing Co. 1977. This is a basic book with many simple ways to examine data to find patterns and relationships.


*Visualizing Data*. William S. Cleveland. Hobart Press, P.O. Box 1473, Summitt, NJ 07902, 1993 and *The Elements of Graphing Data*, 1994 are both continuations of the concept of beautiful and information books on elements of style for elegant graphical presentations of data.

---

Recommended Experimental Design Books (with some basic statistical methods)


---

Recommended General Statistics Books

*Statistics for Environmental Engineers*. Paul Mac Berthouex and Linfield C. Brown, Lewis, 2nd ed. 2001. This excellent book reviews short-comings and benefits of many common statistical procedures, enabling much more thoughtful evaluations of environmental data.


*Primer on Biostatistics*. Stanton A. Glantz. McGraw-Hill. 1992. This is one of the easiest to read and understand introductory texts on basic statistics available.
Recommended Books for Specialized Statistical Methods


Experimental Design

- Numbers of samples to satisfy data quality objectives
- Arrangement of experiments to maximize sensitivity and to identify major factors and interactions

\[ n = \left[ \text{COV} \left( Z_{1-\alpha} + Z_{1-\beta} \right) / \text{error} \right]^2 \]

- \( n \) = number of samples needed
- \( \alpha \) = false positive rate (1-\( \alpha \) is the degree of confidence. A value of \( \alpha \) of 0.05 is usually considered statistically significant, corresponding to a 1-\( \alpha \) degree of confidence of 0.95, or 95%.)
- \( \beta \) = false negative rate (1-\( \beta \) is the power. If used, a value of \( \beta \) of 0.2 is common, but it is frequently ignored, corresponding to a \( \beta \) of 0.5.)
- \( Z_{1-\alpha} \) = Z score (associated with area under normal curve) corresponding to 1-\( \alpha \). If \( \alpha \) is 0.05 (95% degree of confidence), then the corresponding \( Z_{1-\alpha} \) score is 1.645 (from standard statistical tables).
- \( Z_{1-\beta} \) = Z score corresponding to 1-\( \beta \) value. If \( \beta \) is 0.2 (power of 80%), then the corresponding \( Z_{1-\beta} \) score is 0.85 (from standard statistical tables). However, if power is ignored and \( \beta \) is 0.5, then the corresponding \( Z_{1-\beta} \) score is 0.
- error = allowable error, as a fraction of the true value of the mean
- COV = coefficient of variation (sometimes notes as CV), the standard deviation divided by the mean (Data set assumed to be normally distributed.)

Accuracy Definitions:
(a) low precision, large bias,
(b) low precision, small bias,
(c) high precision, large bias, and
(d) high precision, small bias (the only “accurate” case)

Gilbert 1987
Error Types

- (alpha) (type 1 error) - a false positive, or assuming something is true when it is actually false. An example would be concluding that a tested water was adversely contaminated, when it actually was clean. The most common value of is 0.05 (accepting a 5% risk of having a type 1 error). Confidence is $1 - \alpha$, or the confidence of not having a false positive.

- (beta) (type 2 error) - a false negative, or assuming something is false when it is actually true. An example would be concluding that a tested water was clean when it actually was contaminated. If this was an effluent, it would therefore be an illegal discharge with the possible imposition of severe penalties from the regulatory agency. In most statistical tests, is usually ignored (if ignored, is 0.5). If it is considered, a typical value is 0.2, implying accepting a 20% risk of having a type 2 error. Power is $1 - \beta$, or the certainty of not having a false negative.

Experimental Design - Number of Samples Needed

The number of samples needed to characterize stormwater conditions for a specific site is dependent on the COV and allowable error. For most constituents and conditions, about 20 to 30 samples may be sufficient for most objectives. Most Phase 1 sites only have about 10 events, but each stratification category usually has much more.

Experimental Design Example using Preliminary Data

<table>
<thead>
<tr>
<th>Preliminary Data set</th>
<th>#1</th>
<th>#2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Samples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>preliminary data set #1</td>
<td>60</td>
<td>26</td>
</tr>
<tr>
<td>preliminary data set #2</td>
<td>55</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>84</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>38</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>98</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>39</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>55</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>45</td>
</tr>
</tbody>
</table>

Burton and Pitt 2002
### Set A | Set B
--- | ---
mean: & 61.7 & 38.6 
standard deviation: & 19.32 & 16.00 
COV: & 0.31 & 0.41 

\[ u_1 = 61.7 \]
\[ u_2 = 38.6 \]
\[ u_1 - u_2 = 23.1 \]
\[ \text{avg st dev} = 17.66 \]
\[ \text{avg COV} = 0.36 \]
\[ \% \text{difference of means} = 37.44 \]

---

### Factorial Analysis

- A basic and powerful tool to identify significant factors and significant interacting factors.
- Use as the first step in sensitivity analysis and model building.
- Far superior to “holding all variables constant except for changing one variable at a time” classical approach (which doesn’t consider interactions).
- Should be used in almost all experimental evaluations, especially valuable in controlled laboratory tests, and very useful to organize “environmental” test results.

---

Box, Hunter and Hunter 1987
Particle Size Distribution of Street Dirt

Measured Particle Sizes, Including Bed Load Component, at Monroe St. Detention Pond, Madison, WI

Pitt, et al., 1999
Washoff Plots for Heavy Rain Intensities, Dirty Streets, and Rough Pavement Textures

Ratio of Available SS to Total SS Street Dirt Loadings

\[
I = 0.08 \pm 0.04 \\
T = -0.08 \pm 0.05 \\
\hat{Y} = 0.097 + 0.04(I) - 0.04(T)
\]

Pitt 1987

Exploratory Data Analyses

- Basic QA/QC data plots
- Probability plots and histograms
- Scatterplots
- Grouped box and whisker plots
- Simple line plots

These data plots on regular probability graphs indicate few Normal distributions (pH is most obvious and expected).
These log-normal probability plots indicate much better straight-line fits, indicating likely log-normal probability distributions of the data.
**Probability Plots for First-Flush Analyses**

- Normal Probability Plot for TSS_CO_CM...TSS_CO_FF
- ML Estimates

**Comparison of Sewage with Dry Weather Source Samples**

- p=0.05, % Sewage= 0.43
- E. coli = 12,000 MPN/100 mL

- p=0.05, % Sewage= 0.95
- Enterococci = 5,000 MPN/100 mL

**MEDIA CAPACITIES FOR COPPER**

- Plots of concentrations vs. rain depth typically show random patterns.
Plots of expected relationships are being used to identify data redundancies that can reduce future analytical costs.

3-D plot showing lack of obvious relationship between rain depth, geographical area, and drainage area for residential suspended solids data.

Paired observations of data

Parametric tests (data require normality and equal variance)
- Paired Student’s $t$-test (more power than non-parametric tests)

Non-parametric tests
- Sign test (no data distribution requirements, some missing data accommodated)
- Friedman’s test (can accommodate a moderate number of “non-detectable” values, but no missing values are allowed
- Wilcoxon signed rank test (more power than sign test, but requires symmetrical data distributions)
**Solids Removal in Swales: Flow Length**

- Box Plot for Location
- Normal Probability Plot for Location

**Solids Removal in Swales: Flow Depth**

- Box Plot for Flow depth
- Normal Probability Plot for Flow depth

**Two independent groups of data**

- Parametric tests (data require normality and equal variance)
  - Independent Student’s $t$-test (more power than non-parametric tests)

- Non-parametric tests
  - Mann-Whitney rank sum test (probability distributions of the two data sets must be the same and have the same variances, but do not have to be symmetrical; a moderate number of “non-detectable” values can be accommodated)

**Many groups (use multiple comparison tests, such as the Bonferroni $t$-test, to identify which groups are different from the others if the group test results are significant).**

- Parametric tests (data require normality and equal variance)
  - One-way ANOVA for single factor, but for >2 “locations” (if 2 “locations, use Student’s $t$-test)
  - Two-way ANOVA for two factors simultaneously at multiple “locations”
  - Three-way ANOVA for three factors simultaneously at multiple “locations”
  - One factor repeated measures ANOVA (same as paired $t$ test, except that there can be multiple treatments on the same group)
  - Two factor repeated measures ANOVA (can be multiple treatments on two groups)

**Many Groups (cont.)**

- Non-parametric tests:
  - Kurskal-Wallis ANOVA on ranks (use when samples are from non-normal populations or the samples do not have equal variances).
  - Friedman repeated measures ANOVA on ranks (use when paired observations are available in many groups).
Many Groups (cont.)

Nominal observations of frequencies (used when counts are recorded in contingency tables)

- Chi-square ($\chi^2$) test (use if more than two groups or categories, or if the number of observations per cell in a 2X2 table are > 5).

- Fisher Exact test (use when the expected number of observations is <5 in any cell of a 2X2 table).

- McNamar’s test (use for a “paired” contingency table, such as when the same individual or site is examined both before and after treatment)

These grouped box-whisker plots sort all of the data by land use. Kruskal-Wallis analyses indicate that all constituents have at least one significantly different category from the others. Heavy metal differences are most obvious.

Example 2-way ANOVA

- Want to investigate the differences between different strata.
- Are the variations between groups more important than the variations within the groups?
- What about interactions between different variables?
- ANOVA requires normally distributed data. In most stormwater cases, log-transformed values need to be used.
The rain group factor and the season factor are both highly significant. The prior 2-way ANOVA found that the interaction term was not significant; the ANOVA was therefore re-run without that term.

The first and third rain categories are significant.
14

Only Fall and Summer are significant.

Example 1-way ANOVA

- Is at least one member of a group significantly different from the other members?
- Complement analysis with group box-whisker plot
- This doesn’t identify which one(s) is(are) different.
- If a significant member, should be able to recognize from box-whisker plot and with Bonferroni T-test (multiple pair-wise comparisons).

Are any of these sites different from the others?
### ANOVA Single Factor (using Excel)

<table>
<thead>
<tr>
<th>Groups</th>
<th>Count</th>
<th>Sum</th>
<th>Average</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column 1</td>
<td>5</td>
<td>264</td>
<td>52.8</td>
<td>407.7</td>
</tr>
<tr>
<td>Column 2</td>
<td>3</td>
<td>176</td>
<td>58.6667</td>
<td>340.3333</td>
</tr>
<tr>
<td>Column 3</td>
<td>6</td>
<td>1124</td>
<td>187.3333</td>
<td>19161.87</td>
</tr>
<tr>
<td>Column 4</td>
<td>5</td>
<td>196</td>
<td>39.2</td>
<td>427.7</td>
</tr>
<tr>
<td>Column 5</td>
<td>4</td>
<td>69</td>
<td>17.25</td>
<td>128.9167</td>
</tr>
</tbody>
</table>

### ANOVA

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
<th>F crit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>98255</td>
<td>4</td>
<td>24564</td>
<td>4.41</td>
<td>0.0116</td>
<td>2.9277</td>
</tr>
<tr>
<td>Within Groups</td>
<td>100218</td>
<td>18</td>
<td>5567</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>198473</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Pilot-Scale Test Results

**Anaerobic Stripping of Sorbed Pollutants**
Soluble Phosphorus
Star Lake, Winter, Hoover, Alabama

![Graphs showing pilot-scale test results](image-url)
Design Configuration Optimization using Pool Sand Filter Media

Model building/equation fitting (these are parametric tests and the data must satisfy various assumptions regarding behavior of the residuals)

- Linear equation fitting (statistically-based models)
  - Simple linear regression ($y=b_0+b_1x$, with a single independent variable, the slope term, and an intercept. It is possible to simplify even further if the intercept term is not significant).
  - Multiple linear regression ($y=b_0+b_1x_1+b_2x_2+b_3x_3+…+b_kx_k$, having $k$ independent variables. The equation is a multi-dimensional plane describing the data).
  - Stepwise regression (a method generally used with multiple linear regression to assist in identifying the significant terms to use in the model.)
  - Polynomial regression ($y=b_0+b_1x_1+b_2x_2+b_3x_3+…+b_kx_k$, having one independent variable describing a curve through the data).
Non-linear equation fitting (generally developed from theoretical considerations)

- Nonlinear regression (a nonlinear equation in the form: $y=bx$, where $x$ is the independent variable. Solved by iteration to minimize the residual sum of squares).

**Model Building Steps**

1) Re-examine the hypothesis of cause and effect (an original component of the experimental design previously conducted and was the basis for the selected sampling activities).

2) Prepare preliminary examinations of the data, as described previously (most significantly, prepare scatter plots and grouped box/whisker plots).

3) Conduct comparison tests to identify significant groupings of data. As an example, if seasonal factors are significant, then cause and effect may vary for different times of the year.

4) Conduct correlation matrix analyses to identify simple relationships between parameters. Again, if significant groupings were identified, the data should be separated into these groupings for separate analyses, in addition to an overall analysis.

**Modeling Building (cont.)**

5) Further examine complex inter-relationships between parameters by possibly using combinations of hierarchical cluster analyses, principal component analyses (PCA), and factor analyses.

6) Compare the apparent relationships observed with the hypothesized relationships and with information from the literature. Potential theoretical relationships should be emphasized.

7) Develop initial models containing the significant factors affecting the parameter outcomes. Simple apparent relationships between dependent and independent parameters should lead to reasonably simple models, while complex relationships will likely require further work and more complex models.

**Plots to Assist in Model Building**

- Simple Correlation Matrices
- Hierarchical Cluster Analyses
- Principal Component Analyses (PCA) and Factor Analyses
Simple Data Associations

- Pearson Correlation (residuals, the distances of the data points from the regression line, must be normally distributed. Calculates correlation coefficients between all possible data variables. Must be supplemented with scatterplots, or scatter plot matrix, to illustrate these correlations. Also identifies redundant independent variables for simplifying models).

- Spearman Rank Order Correlation (a non-parametric equivalent to the Pearson test).

Complex Data Associations (typically only available in advanced software packages)

- Hierarchical Cluster Analyses (graphical presentation of simple and complex inter-relationships. Data should be standardized to reduce scaling influence. Supplements simple correlation analyses).

- Principal Component Analyses (identifies groupings of parameters by factors so that variables within each factor are more highly correlated with variables in that factor than with variables in other factors. Useful to identify similar sites or parameters).
Regression Analyses

1) Formulate the objectives of the curve-fitting exercise (a subset of the experimental design previously conducted).

2) Prepare preliminary examinations of the data, as described previously (most significantly, prepare scatter plots and probability plots of the data, plus correlation evaluations to examine independence between multiple parameters that may be included in the models).

3) Identify alternative models from the literature that have been successfully applied for similar problems (part of the previously conducted experimental design activities in order to identify which parameters to measure, or to modify or control).

4) Evaluate the data to ensure that regression is applicable and make suitable data transformations.

Regression (cont.)

5) Apply regression procedures to the selected alternative models.

6) Evaluate the regression results by examining the coefficient of determination (R2) and the results of the analysis of variance of the model (standard error analyses and p values for individual equation parameters and overall model).

7) Conduct an analysis of the residuals (as described below).

8) Evaluate the results and select the most appropriate model(s).

9) If not satisfied, it may be necessary to examine alternative models, especially based on data patterns (through cluster analyses and principal component analyses) and re-examinations and modification of the theoretical basis of existing models. Statistical based models can be developed using step-wise regression routines.
Indoor vs. Historical Stillwater, Oklahoma, Retardance Curves

From such graphs swale hydraulic characteristics can be predicted on the basis of flow rate, cross sectional geometry, slope, and vegetation type.

Preferential Capture of Large Particles in Grass Swales

Low flow, blue grass, 5%
Regression Example with ANOVA

- Examining treatment data with regression and associated plots and ANOVA

### Total Suspended Solids mg/L

<table>
<thead>
<tr>
<th>STORM</th>
<th>INLET</th>
<th>OUTLET</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>137</td>
<td>55</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>23</td>
<td>&lt;2.5</td>
</tr>
<tr>
<td>8</td>
<td>75</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>77</td>
<td>&lt;2.5</td>
</tr>
<tr>
<td>10</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>103</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>41</td>
<td>&lt;2.5</td>
</tr>
</tbody>
</table>
### MCTT Performance - Total Suspended Solids mg/L

**Influent**
- N: 12
- Detected Observations: 12
- Mean: 48.6
- Median: 39.5
- StDev: 41.1
- SE Mean: 11.9
- Minimum: 7
- Maximum: 137
- Q1: 16.3
- Q3: 76.5

**Effluent**
- N: 12
- Detected Observations: 9
- Mean: 11.22
- Median: 5.5
- StDev: 16.5
- SE Mean: 5.5
- Minimum: 3
- Maximum: 55
- Q1: 2.7
- Q3: 7
Dependent variable is: LOGOUTLET
No Selector
R squared = 15.6%  R squared (adjusted) = 7.2%
s = 0.4332 with 12 - 2 = 10 degrees of freedom

Source       Sum of Squares  df  Mean Square  F-ratio
Regression   0.347854  1   0.347854   1.85
Residual     1.87625   10  0.187625

Variable       Coefficient    s.e. of Coeff    t-ratio    prob
Constant    0.8333252       0.4876       0.0083       0.9469
LOGINLET    0.421692       0.3097       1.36       0.2032
Residual Analyses of Regression Model

- the residuals are independent
- the residuals have zero mean
- the residuals have a constant variance ($S^2$)
- the residuals have a normal distribution (required for making F-tests)

Plots to Check Residuals

- Check for normality of the residuals (preferably by constructing a probability plot on normal probability paper and having the residuals form a straight line, or at least use an overall plot,

- plot the residuals against the predicted values,

- plot the residuals against the predictor variables, and

- plot the residuals against time in the order the measurements were made.
Data Trends

- Graphical methods (simple plots of concentrations versus time of data collection).
- Regression methods (perform a least-squares linear regression on the above data plot and examine ANOVA for the regression to determine if the slope term is significant. Can be misleading due to cyclic data, correlated data, and data that are not normally distributed).
- Mann-Kendall test (a nonparametric test that can handle missing data and trends at multiple stations. Short-term cycles and other data relationships affect this test and must be corrected).

Data Trends (cont.)

- Sen’s estimator of slope (a nonparametric test based on ranks closely related to the Mann-Kendall test. It is not sensitive to extreme values and can tolerate missing data).
- Seasonal Kendall test (preferred over regression methods if the data are skewed, serially correlated, or cyclic. Can be used for data sets having missing values, tied values, censored values, or single or multiple data observations in each time period. Data correlations and dependence also affect this test and must be considered in the analysis).
Concentration plots vs. time indicate possible trends. Lead has historically dropped significantly from the earliest stormwater studies to the present due to increased use of unleaded gasoline (simple regression trend line shown).