Module 1: Experimental Design for Monitoring

Introduction
This module begins by describing experimental design methods enabling the user to determine the sampling effort needed to accomplish project objectives. The statistical basis for this approach is required to justify the allocation of limited resources. In many cases, certain elements of a multi-faceted study program, as required for practically all stormwater monitoring activities, require much more time and money than other elements of the program. The approach and tools given in this module enable one to balance project resources and scope with expected outcomes. It can be devastating to project conclusions if needed numbers of samples were not obtained at the only time possible. The tools in this module enable one to better plan and conduct a sampling program to minimize this possibility. Of course, all projects conclude with some unresolved issues that were not considered at the initiation of the project. This can only be minimized with increased experience and subject knowledge, plus retaining some flexibility during project execution.

The tools presented here assume some prior knowledge of the situation (especially expected variation in a variable to be measured) in order to determine the sampling effort. This is initially obtained through professional judgment (based on one’s experience in similar situations and from the literature), and generally followed up with a multi-staged sampling effort where an initial experimental design sampling effort is conducted to obtain a better estimate of parameter variability. That better estimate can then be used to better estimate the needed sampling effort during later sampling periods. In all cases, the tools presented here enable one to obtain a level of confidence concerning the significance of the project conclusions. As an example, if it is necessary to compare two sampling locations, such as influent and effluent of a stormwater control device, or comparing a test and control area, or different land uses (common objectives in a stormwater monitoring program), the sampling effort will determine the sensitivity of the study. Depending on the variability of the parameter of interest, a few samples collected may be useful to identify only very large differences in conditions between the sampling locations. Of course, the objective of the study may be to only confirm large differences (such as between influent and effluent conditions for a stormwater
measure known to be very effective). Unfortunately, in most cases involving stormwater discharges, the differences are likely to be much more subtle, requiring numerous samples and careful allocations of the project resources. The tools presented in this module enable one to predict the statistical sensitivity of different sampling schemes, allowing informed decisions and budget requests to be made.

The information included in this module is summarized from the monitoring book by Burton and Pitt (2002) with further additions. While this module discusses experimental design elements of a sampling program, it is obvious that some understanding of the anticipated statistical analyses that will be conducted with the data (including model calibration, for example) must be considered. This module includes some brief numerical examples, while additional examples, especially using factorial analyses, are also presented in the statistical analysis module.

**Experimental Design: Sampling Number and Frequency**

The first part of any study is to formulate the questions being addressed. The expected statistical analysis tools that are expected to be used for evaluating the data should also be an early part of the experimental design. Alternative study plans can then be examined, and finally, the sampling effort can be estimated.

**Sampling Plans**

All sampling plans attempt to obtain certain information (usually average values, totals, ranges, etc.) of a large population by sampling and analyzing a much smaller sample. The first step in this process is to select the sampling plan and then to determine the appropriate number of samples needed. Many sampling plans have been well described in the environmental literature. Gilbert (1987) has defined the following four main categories, plus subcategories, of sampling plans:

- **Haphazard sampling.** Samples are taken in a haphazard (not random) manner, usually at the convenience of the sampler when time permits. This is especially common when the weather is pleasant or the sampling locations most convenient. This is only possible with a very homogeneous condition over time and space, otherwise biases are introduced in the measured population parameters. This strategy is therefore not recommended because of the difficulty of verifying the homogeneous assumption. This sampling strategy may occur when untrained personnel are used for sampling.

- **Judgment sampling.** This strategy is used when only a specific subset of the total population is to be evaluated, with no desire to obtain “universal” characteristics. The target population must be clearly defined (such as during wet weather conditions only) and sampling is conducted appropriately. This could be the first stage of later, more comprehensive, sampling of other target population groups (multistage sampling).

- **Probability sampling.** Several subcategories of probability sampling include:

  - **Simple random sampling.** Samples are taken randomly from the complete population. This usually results in total population information, but it is usually inefficient as a greater sampling effort may be required than if the population was sub-divided into distinct groups. Simple random sampling doesn’t allow information to be obtained for trends or patterns in the population. This method is used when there is no reason to believe that the sample variation is dependent on any known or measurable factor.

  - **Stratified random sampling.** This may the most appropriate sampling strategy for most stormwater studies, especially if combined with an initial limited field effort as part of a multistage sampling effort. The goal is to define strata that results in little variation within any one strata, and great variation between different strata. Samples are randomly obtained from several population groups that are assumed to be internally more homogeneous than the population as a whole, such as separating an annual sampling effort by season and/or rain depth. This results in the individual groups having smaller variations in the characteristics of interest than in the population as a whole. Therefore, sample efforts within each group will vary, depending on the variability of characteristics for each group, and the total sum of the sampling effort may be less than if the complete population was sampled as a whole. In addition, much additional useful information is
likely if the groups are shown to actually be different. This is likely the most suitable sampling strategy that can be used in most stormwater monitoring programs.

- multistage sampling. One type of multistage sampling commonly used is associated with the required subsampling of samples obtained in the field and brought to the laboratory for subsequent splitting for several different analyses. Another type of multistage sampling is when an initial sampling effort is used to examine major categories of the population that may be divided into separate clusters during later sampling activities. This is especially useful when reasonable estimates of variability within a potential cluster are needed for the determination of the sampling effort for composite sampling. These variability measurements may need to be periodically re-verified during the monitoring program.

- cluster sampling. Gilbert (1987) illustrates this sampling plan by targeting specific population units that cluster together, such as an area of deposition near an influent location in a wet pond vs. other locations in the pond. Every unit in each randomly selected cluster can then be monitored.

- systematic sampling. This approach is most useful for basic trend analyses, where evenly spaced samples are collected for an extended time. Evenly spaced sampling is also most efficient when trying to find localized hot spots that randomly occur over an area. However, in wet weather sampling, the rain events are not evenly spaced, but rain events may be selected from within evenly spaced time frames (such as every month), but the events selected need to be equivalent in all other ways, a difficult assumption. This may be most suitable in a receiving water study. Gilbert (1987) presents guidelines for spacing of sampling locations for specific project objectives relating to the size of the hot spot to be found, which would be a suitable approach for lake or pond sediment sampling. Spatial gradient sampling is a systematic sampling strategy that may be worthy of consideration when historical information implies a spatial variation of conditions in a river or other receiving water. One example would be to examine the effects of a point source discharge on receiving sediment quality. A grid would be described in the receiving water in the discharge vicinity whose spacing would be determined by preliminary investigations.

- Search sampling. This sampling plan is used to find specific conditions where prior knowledge is available, such as the location of a historical (but now absence) waste discharger affecting receiving waters. Therefore, the sampling pattern is not systematic or random over an area, but stresses areas thought to have a greater probability of success.

Box, et al. (1978) contains much information concerning sampling strategies, specifically addressing problems associated with randomizing the experiments and blocking the sampling experiments. Blocking (such as in paired analyses to determine the effectiveness of a control device) eliminates unwanted sources of variability. Another way of blocking is to conduct repeated analyses (such for different seasons) at the same locations. Most of the above probability sampling strategies should include randomization and blocking within the final sampling plans (as demonstrated in the following example and in the use of factorial experiments).

**Example Use of Stratified Random Sampling Plan**

Street dirt samples were collected in San Jose, CA, during an early EPA project to identify sources of urban runoff pollutants (Pitt 1979). The samples were collected from narrow strips, from curb to curb, using an industrial vacuum. Many of these strips were to be collected in each area and combined to determine the dust and dirt loadings and their associated characteristics (particle size and pollutant concentrations). Each area (strata) was to be frequently sampled to determine the changes in loadings with time and to measure the effects of street cleaning and rains in reducing the loadings. The analytical procedure used to determine the number of subsamples needed for each composite sample involved weighing individual subsamples in each study area to calculate the coefficient of variation \( \text{COV} = \text{standard deviation/mean} \) of the street surface loading. The number of subsamples necessary \( (N) \), depending on the allowable error \( (L) \), were then determined. An allowable error value of about 25 percent, or less, was selected in order to keep the precision and sampling effort at reasonable levels. The formula used (after Cochran 1963) was:

\[
N = 4\sigma^2/L^2
\]
With 95 percent confidence, this equation calculates the number of sub-samples necessary to determine the true mean value for the street dirt loading within a range of \( \pm L \). As to be shown in the following discussions, more samples are required for a specific allowable error as the COV increases. Similarly, as the allowable error decreases for a specific COV, more samples are also required. Therefore, with an allowable error of 25 percent, the required number of subsamples for a study area with a COV of 0.8 would be 36.

Initially, individual samples were taken at 49 locations in the three study areas to determine the loading variabilities. The loadings averaged about 2700 lb/curb-mile in the Downtown and Keyes Street areas, but were found to vary greatly within these two areas. The Tropicana area loadings were not as high and averaged 310 lb/curb-mile. The Cochran (1963) equation was then used to determine the required number of subsamples in each test area. The data were then examined to determine if the study areas should be divided into meaningful test area groups.

The purpose of these divisions was to identify a small number of meaningful test area-groupings (strata) that would require a reasonable number of subsamples and to increase the usefulness of the test data by identifying important groupings. Five different strata were identified for this research: two of the areas were divided by street texture conditions into two separate strata each representing relatively smooth pavement and rough pavement associated with oil and screens overlies on the street, while the other area was left undivided, as the street texture did not vary greatly. The total number of individual sub-samples for all five areas combined was 111, and the number of subsamples per strata ranged from 10 to 35. In contrast, 150 subsamples would have been needed if the individual areas were not sub-divided. Sub-dividing the main sampling areas into separate strata not only resulted in a savings of about 25% in the sampling effort, but also resulted in much more useful information concerning the factors affecting the values measured. The loading variations in each stratum were re-examined seasonally and the sampling effort was re-adjusted accordingly.

**Factorial Experimental Designs**

Factorial experiments are described in Box, et al. (1978) and in Berthouex and Brown (2002). Both of these books include many alternative experimental designs and examples of this method. Berthouex and Brown (2002) state that “experiments are done to:

1. screen a set of factors (independent variables) and learn which produce an effect,
2. estimate the magnitude of effects produced by experimental factors,
3. develop an empirical model, and
4. develop a mechanistic model.”

They concluded that factorial experiments are efficient tools in meeting the first two objectives and are also excellent for meeting the third objective in many cases. Information obtained during the experiments can also be very helpful in planning the strategy for developing mechanistic models. The main feature of factorial experimental designs is that they enable a large number of possible factors that may influence the experimental outcome to be simultaneously evaluated.

Box, et al. (1978) presents a comprehensive description of many variations of factorial experimental designs. A simple \( 2^3 \) design (three factors: temperature, catalyst, and concentrations at two levels each) is shown in Figure 1 (Box, et al. 1978). All possible combinations of these three factors are tested, representing each corner of the cube. The experimental results are placed at the appropriate corners. Significant main effects can usually be easily seen by comparing the values on opposite faces of the cube. If the values on one face are consistently larger than on the opposite face, then the experimental factor separating the faces likely has a significant effect on the outcome of the experiments. Figure 2 (Box, et al. 1978) shows how these main effects are represented, along with all possible two-factor interactions and the one three-factor interaction. The analysis of the results to identify the significant factors is straight-forward.
Figure 1. Basic Cubic Design of $2^3$ Factorial Test (Box, et al. 1978).
One of the major advantages of factorial experimental designs is that the main effect of each factor, plus the effects of all possible interactions of all of the factors can be examined with relatively few experiments. The initial experiments are usually conducted with each factor tested at two levels (a high and a low level). In monitoring projects, where the conditions are not absolutely controllable, this same strategy can be used to organize the sampling results into suitable strata.

All possible combinations of these factors are then tested (or represented in the monitoring program). Table 1 shows an experimental design for testing 4 factors. This experiment therefore requires $2^4 (=16)$ separate experiments to examine the main effects and all possible interactions of these four factors. The signs signify the experimental conditions for each main factor during each of the 16 experiments. The shaded main factors are the experimental conditions, while the other columns specify the data reduction procedures for the other interactions. A plus sign shows when the factor is to be held at the high level, while a minus sign indicates the low level for the main factors. This table also shows all possible two-way, three-way, and four-way interactions, in addition to the main factors. Simple analyses of the experimental results allow the significance of each of these factors and interactions to be determined. As an example, the following list shows the four factors and the associated levels for organizing the monitoring results used to identify factors affecting runoff quality:

A: Season (plus: winter; minus: summer)
B: Land Use (plus: industrial; minus: residential)
C: Age of Development (plus: old; minus: new)
D: Rain Depth (plus: >1 inch; minus: <1 inch)

<table>
<thead>
<tr>
<th>Experiment #</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>AB</th>
<th>AC</th>
<th>AD</th>
<th>BC</th>
<th>BD</th>
<th>CD</th>
<th>ABC</th>
<th>ABD</th>
<th>BCD</th>
<th>ABCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>7</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>9</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>11</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>16</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

These factors would require the selection of four sampling locations:

1) old industrial area
2) new industrial area
3) old residential area
4) new residential area

The above experiments are designed to collect stormwater runoff data from four test locations. Obviously, both winter and summer seasons must be monitored, and rainfall events of varying depths would be sampled. Rains both less than one inch and greater than one inch would need to be sampled at all monitoring stations in both seasons in order to obtain the needed information.

Even though factorial experiments are best suited in controlled laboratory settings, they have been very useful in organizing environmental data for analysis. Table 2 shows another example where environmental data was organized using a simple factorial design. The design called for a 2^3 experiment to investigate the effects of soil moisture, soil texture, and soil compaction on observed soil infiltration rates (Pitt, et al. 1999). This table shows the calculations from 152 double-ring infiltration tests for the Horton (1939) equation final infiltration rate coefficient (f_c).

Replicate observations enhance the data analysis efforts and grouped standard error values can be calculated (Box, et al. 1978) to identify the significant factors affecting runoff quality. In Table 2, at least 12 replicates were conducted for each test condition to improve the statistical basis for the conclusions. These unusually large numbers of replicates were needed because of the inherently large variability within each test category. If the variability was less, then the number of required replicates could have been much less (as described later in this module). In addition, the site test conditions were not know with certainty when the field tests were run, as some field estimates required confirmation with later laboratory tests that resulted in the reclassification of some of the data.

If observations are not available for some of the needed conditions (such as the monitoring equipment failing during the only large event that occurred at the old industrial site during the summer), then a fractional factorial design can still be used to organize the data and calculate the effects for all of the main factors, and for most of the interactions (as noted in the above experiment). Once the initial experiments are completed, follow-up experiments can be
efficiently designed to examine the linearity of the effects of the significant factors by conducting response surface experimental designs. In addition, further experiments can be conducted and merged with these initial experiments to examine other factors that were not considered in the first experiments. Because of the usefulness and adaptability of factorial experimental designs, Berthouex and Brown (2002) recommend that they “should be the backbone of an experimenter’s design strategy.”

Number of Samples Needed to Characterize Conditions
An important aspect of any research is the assurance that the samples collected represent the conditions to be tested and that the number of samples to be collected are sufficient to provide statistically relevant conclusions. Unfortunately, sample numbers are most often not based on a statistically-based process and follow traditional “best professional judgments,” or are resource driven. The sample numbers should be equal between sampling locations if comparing station data (EPA 1983) and paired sampling should be conducted, if at all possible (the samples at the two comparison sites should be collected at the “same” time, for example), allowing for much more powerful paired statistical comparison tests (see module on statistical analyses). In addition, subsamples must also be collected and then combined to provide a single sample for analysis for many types of sampling, such as collecting discrete subsamples during a rain event. The subsamples are then combined before a single analysis (to reduce analysis expenses) or kept as separate samples (more costly, but provides a legitimate measure of variation/precision) to represent the runoff conditions.

An experimental design process can be used that estimates the number of needed samples based on the allowable error, the variance of the observations, and the degree of confidence and power needed for each parameter. A basic equation that can be used (after Cameron, undated) is as follows:

\[ n = \left[ \text{COV}(Z_{1,\alpha} + Z_{1,\beta})/(\text{error}) \right]^2 \]
Table 2. Example factorial experiment analysis for field project investigating infiltration into disturbed urban soils (Pitt, et al. 1999).
where:

\( n \) = 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.)

This equation is only approximate, as it requires that the data set be normally distributed. However, if the coefficient of variation (COV) values are low (less than about 0.4), then there is likely no significant difference in the predicted sampling effort. This equation is only appropriate as an approximation in many cases, as normal distributions are rare (log-normal distributions are appropriate for most water quality parameters) and the COV values are typically relatively large (closer to 1). The presentation of the results and the statistical procedures used to evaluate the data need to consider the exact degree of confidence of the measured values.

Figure 3 (Pitt and Parmer 1995) is a plot of this equation showing the approximate number of samples needed for an \( \alpha \) of 0.05 (degree of confidence of 95%), and a \( \beta \) of 0.2 (power of 80%). As an example, if an allowable error of about 25% is desired and the COV is estimated to be 0.4, then about 20 samples would have to be collected and analyzed. The samples could be composited and a single analysis conducted, but this would not allow the COV assumption to be confirmed, or the actual confidence range of the concentration to be determined. The use of stratified random sampling can usually be used to advantage by significantly reducing the COV of the subpopulation in the strata, requiring fewer samples for characterization, as noted previously.
Gilbert (1987) presents variations of this basic equation that considers determining the number of samples needed to determine the probability of occurrence within a specified range (such as to calculate the frequency of standard violations). He also presents equations that consider correlated data, such as when the observations are not truly independent, as when very high pollutant concentrations affect values in close spatial or temporal proximity. As expected, correlated data results in needing more samples than indicated from the basic equations.

**Types of Errors Associated with Sampling**

Unfortunately, there are many errors associated with a receiving water study. Errors associated with too few (or too many) samples for a parameter of interest is only one category. Sampling and analytical errors may also be significant and would add to these other errors. Hopefully, the collective sum of all errors is known (through QA/QC activities and adequate experimental design) and manageable. An important aspect of a monitoring program is recognizing the levels of errors and considering the resulting uncertainties in developing recommendations and conclusions.

Generally, errors can be divided into precision and bias problems. Both of these errors, either together or separately, have dramatic effects on the final conclusions of a study. Figure 4 is a classical figure from Gilbert (1987) that
shows the effects of these errors. Bias is a measure of how close the measured median value is to the true median value, while precision is a measure of how “fuzzy” the median estimate is (the repeatability of the analyses and is used to determine the confidence of the measurements).

Errors in decision making are usually divided into type 1 ($\alpha$: alpha) and type 2 ($\beta$: beta) errors:

$\alpha$ (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 $\alpha$ is 0.05 (accepting a 5% risk of having a type 1 error), although other values may be appropriate for specific project objectives and stages. Confidence is $1-\alpha$, or the confidence of not having a false positive.

$\beta$ (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, $\beta$ is usually not directly considered (if ignored, $\beta$ is 0.5), but is assumed to be considered during the experimental design phase with adequate samples collected to control the false negative rate. A typical value of $\beta$ 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. Again, other levels of power may be appropriate for the specific project objectives.

![Figure 4. 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).](image)

It is important that power and confidence be balanced for an effective monitoring program. Most experimental designs ignore power, while providing a high value (typically 95%) for the level of confidence. This is an unrealistic approach as both false negatives and false positives are important. In many environmental programs, power (false
negative problems) may actually be more critical than confidence. If a tested water had a type 2 error (false negative), inappropriate discharges would occur. Typical fines imposed by regulatory agencies are $10,000 per day for non-permitted discharges. Future liability for discharges of waste that were discharged due to an error in measurement or negligence can easily reach into millions of dollars for clean up and health effects. Clearly, one wants to minimize costs, yet have the assurance that the correct decision is being made. However, errors will always be present in any analysis, and some uncertainty in the conclusions must be accepted. Obviously, it can become prohibitively expensive to attempt to reduce monitoring errors to extremely low levels, especially when the monitoring program is affected by uncontrollable environmental factors.

**Determining Sample Concentration Variations**

An important requirement for using the above sampling effort equation is estimating the COV of the parameter of interest. In many cases, the approximate range of likely concentrations can be estimated for a parameter of interest. Figure 5 (Pitt, *et al.* 1993, determined from many Monte Carlo analyses) can be used to estimate the COV value for a parameter by knowing the 10th and 90th percentile ratios (the “range ratio”), assuming a log-normal distribution. Extreme values are usually not well known, but the approximate 10th and 90th percentile values can be estimated with better confidence. As an example, assume that the 10th and 90th percentile values of a water quality constituent of interest are estimated to be about 0.7 and 1.5 mg/L, respectively. The resulting range ratio is therefore $1.5/0.7 = 2.1$ and the estimated COV value is 0.25.

![Figure 5. Determination of Coefficient of Variation from Range of Observations (Pitt, *et al.* 1993).](image)

Also shown on Figure 5 is an indication of the median value, compared to the 10th percentile value and the range ratio. As the range ratio decreases, the median becomes close to the midpoint between the 10th and 90th percentile values. Therefore, at low COV values, the differences between normal distributions and log-normal distributions diminish, as stated previously. As the COV values increase, the mean values are located much closer to the 10th percentile value. In log-normal distributions, no negative concentration values are allowed, but very large positive “outliers” can occur. In the previous example, the median location is about 0.4 for the range ratio of 2.1. The following calculation shows how the median value can be estimated using this “median location” value:
median location = 0.4 = \(\frac{(X_{50}-X_{10})}{(X_{90}-X_{10})}\)

Therefore \(X_{50}-X_{10} = 0.4(X_{90}-X_{10})\).

\((X_{90}-X_{10}) = 1.5 \text{ mg/L} - 0.7 \text{ mg/L} = 0.8 \text{ mg/L}\).

Therefore \(X_{50}-X_{10} = 0.4(0.8) = 0.32 \text{ mg/L}\),

and \(X_{10} = 0.7 \text{ mg/L}, X_{50} = 0.32 \text{ mg/L + } 0.7 \text{ mg/L} = 1.0 \text{ mg/L}\).

For comparison, the average of the 10th and 90th percentile values is 1.1 mg/L. Therefore, the concentration distribution is likely close to being normally distributed and the equation shown previously can be used to estimate the required number of samples needed because these two values are within about 10% of each other. The following paragraphs (from Pitt, et al. 1993) show how log transformations of real-space data descriptors (COV and median) can be used in modifications of these equations.

**Example of Log10 Transformations for Experimental Design Calculations**

For relatively large COV values, it may be necessary to transform the data from known log-normal distributions (checked using log-normal probability paper or suitable statistical test, for example) before calculating the actual error associated with the collected data. Much urban receiving water quality data from the 10th to 90th percentile typically can be suitably described as a normal probability distribution, after log10 transformations of the data. However, values less than the 10th percentile value are usually less than predicted from the log-normal probability plot, while values greater than the 90th percentile value are usually greater than predicted from the log-normal probability plot. Non-transformed water quality data do not typically fit normal probability distributions very well, except for pH values (which are log transformed values of the hydrogen ion concentration, by definition).

Figure 6 (Pitt, et al. 1993) presents a relationship between the COV value in real space (non-transformed) and the standard deviation of log10 transformed data. Knowing the log10 transformed standard deviation values enables certain statistical experimental design features to be determined. The most significant feature is determining the number of observations needed to enable the data to be described with a specific error level. It can also be used to calculate the error associated with any observation, based on the assumed population distribution characteristics and the number of observations. As an example, consider a pollutant having a COV of 0.23 and a median value of 0.14. The resulting log10 transformed standard deviation would be about 0.12. One equation that has been historically used to calculate the number of analyses needed, based on the allowable error is (Cochran 1963):

\[
\text{Number of samples} = 4(\text{standard deviation})^2 / (\text{allowable error})^2
\]
With a 95 percent level of confidence, this relationship determines the number of samples needed to obtain a value within the range of the sample mean, plus and minus the error. This equation can be re-arranged to obtain the error, based on the number of samples obtained and the standard deviation. As an example, for ten samples and the above standard deviation (0.12), the resulting approximate 95 percent confidence range (ignoring false negatives) of the median observation (0.14 mg/L) is:

\[
\text{Error} = 2(0.12)/(10)^{0.5} = 0.076 \text{ in log}_{10} \text{ space}
\]

The confidence interval is therefore \(\log_{10}(0.14) \pm 0.076\), which is -0.778 to -0.930 in \(\log_{10}\) space. This results in a conventional 95 percent confidence range of \(10^{-0.930} = 0.12\) to \(10^{-0.778} = 0.17\). The absolute value for the error in the estimate of the median value is therefore between 14% \((100\times(0.14-0.12)/0.14)\) and 21% \((100\times(0.17-0.14)/0.14)\) for ten samples. If the original untransformed data were used, the error associated with 10 samples is about 15%, within the range of the estimate after log transformations. These results are close because of the low COV value (0.23). If the COV value is large (>0.4), the need for log transformations before the experimental design calculation increases.

**Example Showing Improvement of Mean Concentrations with Increasing Sampling Effort**

Many stormwater discharge samples were obtained from two study areas during the Bellevue, Washington, Urban Runoff Program (Pitt 1985). The runoff from each drainage area was affected by different public works stormwater control practices and the outfall data were compared to identify if any runoff quality improvements were associated with this effort. These data offer an opportunity to examine how increasing numbers of outfall data decreased the uncertainty of the overall average concentrations of the stormwater pollutants. Table 3 shows how the accumulative average of the observed concentrations eventually become reasonable steady, but only after a significant sampling effort. As an example, the average of the first three observations would result in an EMC (Event-Mean Concentration) that would be in error by about 40%. It would require more than 15 samples before the average value is consistently less than 10% from the seasonal average value (which only had a total population of 25 storm events), even with the relatively small COV value of 0.65.
Table 3. Event-Mean Concentrations for Series of Storm Samples in Bellevue, Washington (Pitt 1985)

<table>
<thead>
<tr>
<th>Storm #</th>
<th>Lead Concentration (mg/L)</th>
<th>Moving Average Concentration (EMC)</th>
<th>Error from Seasonal Average (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.53</td>
<td>0.53</td>
<td>119</td>
</tr>
<tr>
<td>2</td>
<td>0.10</td>
<td>0.32</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>0.38</td>
<td>0.34</td>
<td>39</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
<td>0.29</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>0.12</td>
<td>0.26</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>0.12</td>
<td>0.23</td>
<td>-3</td>
</tr>
<tr>
<td>7</td>
<td>0.56</td>
<td>0.28</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>0.19</td>
<td>0.27</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>0.38</td>
<td>0.28</td>
<td>16</td>
</tr>
<tr>
<td>10</td>
<td>0.23</td>
<td>0.28</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>0.20</td>
<td>0.27</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>0.39</td>
<td>0.28</td>
<td>16</td>
</tr>
<tr>
<td>13</td>
<td>0.53</td>
<td>0.30</td>
<td>24</td>
</tr>
<tr>
<td>14</td>
<td>0.05</td>
<td>0.28</td>
<td>16</td>
</tr>
<tr>
<td>15</td>
<td>0.26</td>
<td>0.28</td>
<td>16</td>
</tr>
<tr>
<td>16</td>
<td>0.05</td>
<td>0.27</td>
<td>10</td>
</tr>
<tr>
<td>17</td>
<td>0.05</td>
<td>0.25</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>0.39</td>
<td>0.26</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>0.28</td>
<td>0.26</td>
<td>8</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>0.25</td>
<td>5</td>
</tr>
<tr>
<td>21</td>
<td>0.29</td>
<td>0.25</td>
<td>6</td>
</tr>
<tr>
<td>22</td>
<td>0.18</td>
<td>0.25</td>
<td>4</td>
</tr>
<tr>
<td>23</td>
<td>0.31</td>
<td>0.25</td>
<td>5</td>
</tr>
<tr>
<td>24</td>
<td>0.10</td>
<td>0.25</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>0.10</td>
<td>0.24</td>
<td>0</td>
</tr>
</tbody>
</table>

Albert and Horwitz (1988) point out that taking averages leads to a tighter distribution. As shown above, the extreme values have little effect on the overall average, even with a relatively few observations (for a Gaussian distribution). The reduction in the standard deviation is proportional to $1/n^{0.5}$, for $n$ observations. Even if the population is not Gaussian, the averages tend to be Gaussian-like. In addition, the larger the sample size, the more Gaussian-like is the population of averages.

Determining the Number of Sampling Locations (or Land Uses) Needed to be Represented in a Monitoring Program

The above example for characterizing a stormwater characteristic parameter briefly examined a method to determine the appropriate number of samples that should be collected and analyzed at a specific location. However, another aspect of sample design is determining how many components (specifically sampling locations) need to be characterized. The following example uses a marginal benefit analysis to help identify a basic characterization monitoring program. The sampling effort procedure discussed previously applies to the number of samples needed for each sampling location, while this discussion identifies the number of sampling locations that should be monitored. This example specifically examines which land use categories should be included in a city-wide monitoring program when the total city’s stormwater discharges need to be quantified with a reasonable error.

Land Use Monitoring for Wet Weather Discharge Characteristics. The following paragraphs outline the steps that can be used to select the specific land uses that need to be included in a monitoring program to characterize stormwater runoff from an urban area to a specific receiving water. The following example is loosely based on analyses of data for the Waller Creek drainage in Austin, TX. The modules describing development characteristics and the National Stormwater Quality Database, also contain much information that needs to be considered along with the topic in this subsection.

Step 1 - This step identifies the land use categories that exist in the area of study. The information collected during the preliminary site selection activities will enable effective monitoring sites to be selected. In addition, this
information will provide very useful information needed to extrapolate the monitoring results across the whole drainage area (by understanding the locations and areas of similar areas represented by the land use-specific monitoring stations) and to help identify the retro-fit control programs that may be suitable for these types of areas and to understand the benefits of the most cost-effective controls for new development.

The initial list of land use areas to be considered for monitoring should be based on available land use maps, but they will have to be modified by overlaying additional information that should have an obvious effect on stormwater quality and quantity. The most obvious overlays would be the age of development (an “easy” surrogate for directly connected imperviousness, maturity of vegetation, width of streets, conditions of streets, etc., that all affect runoff conditions and control measure applications) and the presence of grass swale drainage (which has a major effect on mass discharges and runoff frequency). Some of these areas may not be important (such as a very small area represented in the study area, especially with known very low concentrations or runoff mass) and may be eliminated at this step. After this initial list (with subcategories) is developed, locations that are representative of each potential category need to be identified for preliminary surveys. About ten representative neighborhoods in each category that reflect the full range of development conditions for each category should be identified. The 10 locations in each land use would be relatively small areas, such as a square block for residential areas, a single school or church, a few blocks of strip commercial, etc. The ten sites would be selected over a wide geographical area of the study area to include topographical effects, distance from ocean, etc.

Step 2 - This step includes preliminary surveys of the land uses identified above. For each of the 10 neighborhoods identified in each category, simple field sheets are filled out with information that may affect runoff quality or quantity, including: type of roof connections, type of drainage, age of development, housing density, socio-economic conditions, quantity and maintenance of landscaping, condition of pavement, soils, inspections of storm drainage to ensure no inappropriate discharges, and existing stormwater control practices. These are simple field surveys that can be completed by a team of two people at the rate of about ten locations a day, depending on navigation problems, traffic, and how spread-out the sites are. Several photographs are also made of each site and are archived with the field sheets for future reference.

Step 3 - In this step, measurements of important surface area components are made for each of the neighborhoods surveyed above. These measurements are made using aerial photographs of each of the ten areas in each land use category. Measurements will include areas of: rooftops, streets, driveways, sidewalks, parking areas, storage areas, front grass strips, sidewalks and streets, playgrounds, backyards, front yards, large turf areas, undeveloped areas, decks and sheds, pools, railroad rows, alleyways, and other paved and non-paved areas. This step requires the use of good aerial photography in order to resolve the elements of interest for measurement. Print scales of about 100 ft per 1 inch are probably adequate, if the photographs are sharp. Photographic prints for each of the homogeneous neighborhoods examined on the ground in step 2 are needed. The actual measurements require about an hour per site. These measurements can be supplemented with automated GIS systems, but the automated systems are seldom sufficiently accurate or detailed enough.

Step 4 - In this step, the site survey and measurement information is used to confirm the groupings of the individual examples for each land use category. This step would finalize the categories to be examined, based on the actual measured values. As an example, some of the sites selected for field measurement may actually belong in another category (based on actual housing density, for example) and would be then reassigned before the final data evaluation. More importantly, the development characteristics (especially drainage paths) and areas of important elements (especially directly connected pavement) may indicate greater variability within an initial category than between other categories in the same land use (such as for differently aged residential areas, or high density residential and duplex home areas). A simple ANOVA test would indicate if differences exist and additional statistical tests can be used to identify the specific areas that are similar. If there is no other reason to suspect differences that would affect drainage quality or quantity (such as landscaping maintenance for golf courses vs. undeveloped areas), then these areas could be combined to reduce the total number of individual land use categories/sub-categories used in subsequent evaluations.

Step 5 - This step includes the ranking of the selected land use categories according to their predominance and pollutant generation. A marginal benefit analysis can be used to identify which land use categories should be
monitored. Each land use category has a known area in the drainage area and an estimated pollutant mass discharge. This step involves estimating the total annual mass discharges associated with each land use category for the complete study area. These sums are then ranked, from the largest to the lowest, and an accumulated percentage contribution is then produced. These accumulated percentage values are plotted against the number of land use categories. The curve will be relatively steep initially and then level off as it approaches 100%. A marginal benefit analysis can then be used to select the most effective number of land uses that should be monitored.

The following is an example of this marginal benefit analysis to help select the most appropriate number of land uses to monitor. The numbers and categories are based on the Waller Creek, Austin, Texas, watershed. Table 4 shows 16 initial land use categories, their land cover (as a percentage), and the estimated unit area loadings for each category for a critical pollutant. These loading numbers will have to be obtained using best judgment and prior knowledge (such as from the National Stormwater Quality Database, Maestre and Pitt 2005). This table then shows the relative masses of the pollutant for each land use category (simply the % area times the unit area loading). The land uses are shown ranked by their relative mass discharges and a summed total is shown. This sum is then used to calculate the percentage of the pollutant associated with each land use category. These are then accumulated. The “straight-line model” is the straight line from 0 mass at 0 stations to 100% of the mass at 16 stations. The final column is the difference between these two lines (the marginal benefit).

Figure 7 is a marginal benefit plot of these values. The most effective monitoring strategy is to monitor seven land uses in this example. After this number, the marginal benefit starts to decrease. Seven (out of 16) land uses will also account for about 75% of the total annual emissions from these land uses in this area. A basic examination of the plot shows a strong leveling of the curve at 12 land uses, where the marginal benefit dramatically decreases and where there is little doubt of additional benefit for additional effort. The basic interpretation of this data should include:

- the marginal benefit (as shown to include 7 out of the 16 land uses for monitoring in this example)
- land uses that have expected high unit area mass discharges that may not be included in the above list because of relatively low abundance, such as shopping malls in this example, should also be considered for monitoring
- land uses that are expected to increase in the future to become a significant component (such as the new medium density residential area in this example)
- land uses that have special conditions, such as a grass swale site in this example, that may need to be demonstrated/evaluated.
Step 6 - Final selection of monitoring locations. These top ranked land uses will then be selected for monitoring. In most cases, a maximum of about ten sites would be initiated each year. The remaining top-ranked land uses will then need to be monitored starting in future years because of the time needed to establish monitoring stations. In selecting sites for monitoring, sites draining homogeneous areas need to be found. In addition, monitoring locations will need to be selected that have sampling access, no safety problems, etc. To save laboratory resources, three categories of the land uses can be identified. The top group would have the most comprehensive monitoring efforts (including most of the critical source area monitoring activities), while the lowest group may only have flow monitoring (with possibly some manual sampling). The middle group would have a shorter list of constituents routinely monitored, with periodic checks for all constituents being investigated.

Step 7 - The monitoring facilities will need to be installed. The monitoring equipment should be comprised of automatic water samplers and flow sensors (velocity and depth of flow in areas expected to have surcharging flow problems), plus a tipping bucket rain gage. The samples should all be obtained as flow-weighted composites, requiring only one sample to be analyzed per event at each monitoring station.

The sampler should initiate sampling after three tips (about 0.03 inches of rain) of the tipping bucket rain gage at the sampling site. Another sample initiation method is to use an offset of the flow stage recorder to cause the sampler to begin sampling after a predetermined rise in flow conditions. False starts are then possible, caused by inappropriate discharges in the watershed above the sampling station. Frequent querying of sampler, flow, and rain conditions (using a data logger with phone connections) will detect this condition to enable retrieval of these dry-weather samples for analyses and to clean and reset the sampler. Both tripping methods can be used simultaneously to ensure that only wet weather samples are obtained. Of course, periodic (on random days about a month apart) dry-weather sampling (on a time composite basis over 24 hours) is also likely needed.

Figure 7. Marginal Benefit Associated with Increasing Sampling Effort.
Table 4. Example Marginal Benefit Analysis

<table>
<thead>
<tr>
<th>Land Use (ranked by % mass per category)</th>
<th>% of area</th>
<th>critical unit area loading</th>
<th>relative mass</th>
<th>% mass per category</th>
<th>accum. (% mass)</th>
<th>straight-line model</th>
<th>marginal benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Older medium density resid.</td>
<td>24</td>
<td>200</td>
<td>4800</td>
<td>22.8</td>
<td>22.8</td>
<td>6.25</td>
<td>16.5</td>
</tr>
<tr>
<td>2 High density resid.</td>
<td>7</td>
<td>300</td>
<td>2100</td>
<td>10.0</td>
<td>32.7</td>
<td>12.5</td>
<td>20.2</td>
</tr>
<tr>
<td>3 Office</td>
<td>7</td>
<td>300</td>
<td>2100</td>
<td>10.0</td>
<td>42.7</td>
<td>18.8</td>
<td>24.0</td>
</tr>
<tr>
<td>4 Strip commercial</td>
<td>8</td>
<td>250</td>
<td>2000</td>
<td>9.5</td>
<td>52.2</td>
<td>25.0</td>
<td>27.2</td>
</tr>
<tr>
<td>5 Multiple family</td>
<td>8</td>
<td>200</td>
<td>1600</td>
<td>7.6</td>
<td>59.8</td>
<td>31.3</td>
<td>28.5</td>
</tr>
<tr>
<td>6 Manufacturing industrial</td>
<td>3</td>
<td>500</td>
<td>1500</td>
<td>7.1</td>
<td>66.9</td>
<td>37.5</td>
<td>29.4</td>
</tr>
<tr>
<td>7 Warehousing</td>
<td>5</td>
<td>300</td>
<td>1500</td>
<td>7.1</td>
<td>74.0</td>
<td>43.8</td>
<td>30.3</td>
</tr>
<tr>
<td>8 New medium density resid.</td>
<td>5</td>
<td>250</td>
<td>1250</td>
<td>5.9</td>
<td>80.0</td>
<td>50.0</td>
<td>30.0</td>
</tr>
<tr>
<td>9 Light industrial</td>
<td>5</td>
<td>200</td>
<td>1000</td>
<td>4.7</td>
<td>84.7</td>
<td>56.3</td>
<td>28.4</td>
</tr>
<tr>
<td>10 Major roadways</td>
<td>5</td>
<td>200</td>
<td>1000</td>
<td>4.7</td>
<td>89.4</td>
<td>62.5</td>
<td>26.9</td>
</tr>
<tr>
<td>11 Civic/educational</td>
<td>10</td>
<td>100</td>
<td>1000</td>
<td>4.7</td>
<td>94.2</td>
<td>68.8</td>
<td>25.4</td>
</tr>
<tr>
<td>12 Shopping malls</td>
<td>3</td>
<td>250</td>
<td>750</td>
<td>3.6</td>
<td>97.7</td>
<td>75.0</td>
<td>22.7</td>
</tr>
<tr>
<td>13 Utilities</td>
<td>1</td>
<td>150</td>
<td>150</td>
<td>0.7</td>
<td>98.5</td>
<td>81.3</td>
<td>17.2</td>
</tr>
<tr>
<td>14 Low density resid. with swales</td>
<td>5</td>
<td>25</td>
<td>125</td>
<td>0.6</td>
<td>99.1</td>
<td>87.5</td>
<td>11.6</td>
</tr>
<tr>
<td>15 Vacant</td>
<td>2</td>
<td>50</td>
<td>100</td>
<td>0.5</td>
<td>99.5</td>
<td>93.8</td>
<td>5.8</td>
</tr>
<tr>
<td>16 Park</td>
<td>2</td>
<td>50</td>
<td>100</td>
<td>0.5</td>
<td>100.0</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>total</td>
<td>100</td>
<td>21075</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The base of the automatic sampler will need to be modified to use a larger sample bottle (as much as a 100 L Teflon lined drum, with a 10 L glass bottle suspended for small events) in order to automatically sample a wide range of rain conditions without problems. A refrigerated base may also be needed, depending on ambient air conditions and sample holding requirements. The large drum will need to be located in a small freezer, with a hole in the lid where the sample line from the automatic sampler passes through.

Each sampler should also be connected to a cell phone so the sampler status (including temperature of sample) and rainfall and flow conditions can be observed remotely. This significantly reduces personnel time and enables sampler problems to be identified quickly. Each sampler site will also need to be visited periodically (about weekly) to ensure that everything is ready to sample.

Step 8 - The monitoring initiation should continue down the list of ranked land use categories and repeat steps 6 and 7 for each category. At some point the marginal benefit from monitoring an additional land use category will not be sufficient to justify the additional cost.

As a very rough estimate, it could take the following time to complete each step for a large city: Steps l-3, one month each; Steps 4 and 5, 1 month combined; Step 6, three months; Step 7, three months; Step 8, continuous, for a total of about 10 months. As an example, this process was totally completed by Los Angeles County, for the unincorporated areas, in just a few months.

**Determining the Number of Samples Needed to Identify Unusual Conditions**

An important aspect of stormwater monitoring studies is investigating unusual conditions. The methods presented by Gilbert (1987) (“Locating Hot Spots”) can be used to select sampling locations that have acceptable probabilities of locating unusual conditions that are spatially different from other locations. This method would be most applicable
for studies of sediment in wet ponds or soils in the bottom of infiltration devices, for example. Gilbert concluded that the use of a regular spacing of samples over an area was more effective when the contamination pattern was irregular, and an irregular pattern was best if the contamination existed in a repeating pattern. In almost all cases, unusual contamination has an irregular pattern and a regular grid is recommended. Gilbert presents square, rectangular, and triangular grid patterns to help locate sampling locations over an area. The sampling locations are located at the nodes of the resulting grids. Figure 8 (Gilbert 1987) is for the rectangular grid pattern, where the grid has a 2 to 1 aspect ratio. The figure relates the ratio of the size of a circular hot spot to the rectangular grid dimensions (sampling spacing) to the probability of detection. β is the probability of not finding the spot, while S is shape factor for the hot spot (S = 1 for a circular spot, while S = 0.5 for an elliptical spot). For example, if a semi-elliptical spot was to be targeted (S=0.7), and the acceptable probability of not finding the spot was set at 25% (β = 0.25), the required L/G ratio would be about 0.95, with the rectangular width about equal to the minor radius of the desired target.

![Figure 8. Sample Spacing Needed to Identify Unusual Conditions (Gilbert 1987).](image)

**Number of Samples Needed for Comparisons between Different Sites or Times**

The comparison of paired data sets is commonly used when evaluating the differences between two situations (locations, times, practices, etc.). A related equation to the one given previously can be used to estimate the needed samples for a paired comparison (Cameron, undated):

\[ n = 2 \left[ \frac{(Z_{1-\alpha} + Z_{1-\beta})(\mu_1 - \mu_2)}{\sigma^2} \right]^2 \]

where \( \alpha \) = 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-\( \alpha \)

\( Z_{1-\beta} \) = Z score corresponding to 1-\( \beta \) value

\( \mu_1 \) = mean of data set one
\[ \mu_2 = \text{mean of data set two} \]
\[ \sigma = \text{standard deviation (same for both data sets, same units as } \mu. \text{ Both data sets are also assumed to be normally distributed.)} \]

This equation is also only approximate, as it requires that the two data sets be normally distributed and have the same standard deviations. As noted previously, most stormwater parameters of interest are likely closer to being log-normally distributed. Again, if the coefficient of variation (COV) values are low (less than about 0.4), then there is probably no real difference in the predicted sampling effort.

Figure 9 (Pitt and Parmer 1996) is a plot of this equation (normalized using COV and differences of sample means) showing the approximate number of sample pairs needed for an \( \alpha \) of 0.05 (degree of confidence of 95%), and a \( \beta \) of 0.2 (power of 80%). As an example, twelve sample pairs will be sufficient to detect significant differences (with at least a 50% difference in the parameter value) for two locations, if the coefficient of variations are no more than about 0.5. Appendix A (Pitt and Parmer 1996) contains similar plots for many combinations of other levels of power, confidence and expected differences.

![Figure 9. Sample Effort Needed for Paired Testing (Power of 80% and Confidence of 95%) (Pitt and Parmer 1995).](image-url)
**Need for Probability Information and Confidence Intervals**

The above discussions presented information mostly pertaining to a simple characteristic of the population being sampled: the “central tendency”, usually presented as the average, or mean, of the observations. However, much greater information is typically needed, especially when conducting statistical analyses of the information. Information concerning the probability distribution of the data (especially variance) was used previously as it affected sampling effort. However, many more uses of the probability distributions exist. Albert and Horwitz (1988) state that the researcher must be aware of how misleading an average value alone can be, because the average tells nothing about the underlying spread of values. Berthouex and Brown (2002) also point out the importance of knowing the confidence interval (and the probability) of a statistical conclusion. It can be misleading to simply state that the results of an analysis is significant (implying that the null hypothesis, the difference between the means of two sets of data is zero, is rejected at the 0.05 level), for example, when the difference may not be very important. It is much more informative to present the 95% confidence interval of the difference between the means of the two sets of data in most cases.

One important example of how probability affects decisions concerns the selection of critical and infrequent conditions. In hydrology analyses, the selection of a “design” rainfall dramatically affects the design of a drainage system. The probability that a high flow rate (or any other factor of interest having a recurrence interval of “T” years) will occur during “n” years is:

\[ P = 1 - (1 - 1/T)^n \]

As an example, the probability of a year rain occurring at least once in a 5 year period is not 1, but is:

\[ P = 1 - (1 - 1/5)^5 = 1 - (0.8)^5 = 1 - 0.328 = 0.67 \] (or 67%).

In another example, a flow having a recurrence interval of 20 years is of interest. That flow is likely to have the following probability of occurrence during a 100-year period:

\[ P = 1 - (1 - 1/20)^{100} = 1 - (0.95)^{100} = 1 - 0.0059 = 0.994 \] (99.4%)

but only the following probability of occurrence during a year period:

\[ P = 1 - (1 - 1/20)^5 = 1 - (0.95)^5 = 1 - 0.774 = 0.227 \] (22.7%)

Figure 10 (McGee 1991) illustrates this equation. If a construction project was to last for 2 years, but the erosion control practices need to be certain of survival at least at the 95% level, then a 40-year design storm condition must be used! Similarly, a 1,000-year design flow (one only having a 0.1% chance of occurring in any one year) would be needed if one needed to be 90% certain that it would not be exceeded during a 100-year period.

An entertaining example presented by Albert and Horwitz (1988) illustrates an interesting case concerning the upper limits of a confidence interval. In their example, an investigator wishes to determine if purple cows really exist. While traveling through a rural area, 20 cows are spotted, but none are purple. What is the actual percentage of cows that are purple (at a 95% confidence level), based on this sampling? The following formula can be used to calculate the upper limit of the 95% confidence interval:

\[ (1-0)^n - (1-x)^n = 0.95 \]

or

\[ 1 - (1-x)^n = 0.95 \]
where n is the number of absolute negative observations and x is the upper limit of the 95% confidence interval. Therefore, for a sampling of 20 cows (n = 20), the actual percentage of cows that are purple is between 0.0% and 13.9% (x = 0.139). If the sample was extended to 40 cows (n = 40), the actual percentage of cows that are purple would be between 0.0% and 7.2% (x = 0.072). The upper limit of both of these cases is well above zero and, for most people, these results generally conflict with common sense. Obviously, the main problem with the above purple cow example is the violation of the need for random sampling throughout the whole population. Also, the confidence interval includes the zero value (the likely correct answer). In later discussions of regression, it is shown that the confidence intervals of the equation coefficients need to be examined. If doing a trend analysis, for example, if the confidence interval of the “slope” term includes the zero value, the trend is not considered significant.

**Summary: Experimental Design for Monitoring**

This module presented methods to determine the needed sampling effort, including the number of samples and the number of sampling locations. These procedures can be utilized for many different conditions and situations, but some prior knowledge of the conditions to be monitored is needed. A phased sampling approach is therefore recommended, allowing some information to be initially collected and used to make preliminary estimates of the sampling effort. Later sampling phases are then utilized to obtain the total amount of data expected to be needed.
References

Appendix A: Sampling Requirements for Paired Tests

Number of Sample Pairs Needed
(Power = 0.5  Difference = 10%)

Number of Sample Pairs Needed
(Power = 0.8  Difference = 10%)
Number of Sample Pairs Needed
(Power = 0.9  Difference = 10%)

Number of Sample Pairs Needed
(Power = 0.5  Difference = 25%)
Number of Sample Pairs Needed
(Power = 0.8  Difference = 25%)

Coefficient of Variation

Number of Sample Pairs Needed
(Power = 0.9  Difference = 25%)

Coefficient of Variation
Number of Sample Pairs Needed
(Power = 0.5  Difference = 50%)

Number of Sample Pairs Needed
(Power = 0.8  Difference = 50%)
Number of Sample Pairs Needed
(Power = 0.8  Difference = 75%)

Number of Sample Pairs Needed
(Power = 0.9  Difference = 75%)

Coefficient of Variation

Degree of Confidence (1 - alpha)