Experiment Design and Data Analysis

When dealing with measurement and simulation, a careful experiment design and data analysis are essential for reducing costs and drawing meaningful conclusions. The two issues are coupled, since it is usually not possible to select all parameters of an experiment without doing a preliminary run and analyzing the data obtained.

Simulation Techniques

- Continuous-Time Simulation
- Discrete-Event Simulation

A Standard Uniform Random Variable Y

- Let us assume that Y is a random variable uniformly distributed between 0 and 1. That is

$$F_X(x) = P[X \leq x] = \begin{cases} 0, & \text{if } x < 0, \\ x, & \text{if } 0 \leq x \leq 1, \\ 1, & \text{if } x > 1. \end{cases}$$

Generating a Random Variable X with Distribution G(.)

- Define $X = G^{-1}(Y)$.

- Then,

$$\begin{aligned} F_X(x) &= P[X \leq x] \\ &= P[G^{-1}(Y) \leq x] \\ &= P[Y \leq G(x)] \\ &= G(x). \end{aligned}$$

Fundamentals of Data Analysis

The most fundamental aspect of the systems of interest is that they are driven by a nondeterministic workload. The randomness in the inputs makes the outputs also random. Thus, no single observation from the system would give a reliable indication of the performance of the system. One way to cope with this randomness is to use several observations in estimating how the system will behave "on average".

Some Questions

- How do we use several observations to estimate the average performance, i.e., what is a good estimator based on several observations?

- Is an estimate based on several observations necessarily more reliable than the one based on a single observation?

- How do we characterize the error in our estimate as a function of the number of observations? Or, put another way, given the tolerable error, how do we determine the number of observations?

Some Questions (continued)

- How do we perform experiments so that the error characterization is itself reliable?
- If the number of needed observations is found to be too large, what can we do to reduce it?

Some Assumptions

- Let X denote a performance measure of interest (e.g., the response time).

- We can regard X as a random variable with some unknown distribution. Let s and σ^2 denote its mean and variance respectively.

- Suppose that we obtain the observations $X_1, X_2, \cdots, X_n$ as a sequence of i.i.d. random variables where for each i, $E(X_i) = s$ and $Var(X_i) = \sigma^2$.

Sample Mean Estimator $\overline{X}$

- $$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- $\overline{X}$ is an unbiased estimator because

$$E[\overline{X}] = \frac{1}{n} E[\sum_{i=1}^{n} X_i] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = s$$

Variance of Sample Mean Estimator $\overline{X}$

$$\begin{aligned}
\sigma^2_{\overline{X}} &= E[(\overline{X} - s)^2] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} E[(X_i - s)(X_j - s)] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} E[X_i - s]^2 + \frac{1}{n^2} \\
&\quad \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} E[(X_i - s)(X_j - s)] \\
&= \frac{\sigma^2}{n} + \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} Cov(X_i, X_j) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Variance of Sample Mean Estimator $\overline{X}$ (continued)

- If σ is finite, then we have

$$\lim_{n\to\infty} \sigma^2_{\overline{X}} = 0.$$

- That is the sample mean will converge to the expected value as $n \to \infty$. This is one form of the *law of large numbers*.

Sample Variance Estimator δ_X^2

- $$\delta_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Sample Variance Estimator δ_X^2 (continued)

- $$\begin{aligned}\phi &= \sum_{i=1}^{n} E[X_i - s + s - \overline{X}]^2 \\ &= \sum_{i=1}^{n} E[(X_i - s) - \frac{1}{n}\sum_{j=1}^{n}(X_j - s)]^2\end{aligned}$$

- Expanding the square, taking the expectation operator inside, and noting that $E[(X_i - s)^2] = \sigma^2$ for any i, we can have as in the next page:

Sample Variance Estimator δ_X^2(continued)

$$\begin{aligned}
\phi &= \sum_{i=1}^{n}[\sigma^2 - \frac{\sigma^2}{n} - \frac{2}{n}\sum_{j\neq i} Cov(X_i, X_j) \\
&\quad + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k\neq j} Cov(X_j, X_k)] \\
&= (n-1)\sigma^2 - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} Cov(X_i, X_j)
\end{aligned}$$

Sample Variance Estimator δ_X^2(continued)

- It is easy to see that $E[\delta_X^2] = \frac{\phi}{n-1}$
- Thus, we see that if X_i's are mutually independent, δ_X is an unbiased estimator of σ, but not otherwise in general.
- Since $Var(\overline{X}) = \frac{\sigma^2}{n}$ in this case, we can also define an unbiased estimator of $Var(\overline{X})$, denoted $\delta_{\overline{X}}^2$, as simply $\frac{\delta_X^2}{n}$.

Characterization of the value of s

- The measures $\overline{X}$ and $\delta^2_{\overline{X}}$ give some idea about the value of s.

- For a more concrete characterization, we would like to obtain an interval of width e around $\overline{X}$, such that the real value s lies somewhere in the range of $\overline{X} \pm e$.

- Since $\overline{X}$ is a random variable, we can specify such a finite range only with a probability $P_0 < 1$.

- The parameter P_0 is called the *confidence level*, and must be chosen a priori.

Characterization of the value of s (continued)

- Thus, our problem is to determine e such that
$$Pr(|\overline{X} - s| \leq e) = P_0$$

- The parameter $2e$ is called the *confidence interval*, and is expected to increase as P_0 increases.

- To determine the value of e, we need to know the distribution of$\overline{X}$.

- To this end, we use the *central limit theorem*, and conclude that if n is large, the distribution of $\overline{X}$ can be approximated as $\mathcal{N}(s, \sigma/\sqrt{n})$, i.e., normal with mean s and variance σ^2/n.

Characterization of the value of s (continued)

- Let
$$Y = (\overline{X} - s)\sqrt{n}\,/\sigma$$

- Then, the distribution of Y must be $\mathcal{N}(0, 1)$.

- We can find e' such that
$$Pr(|Y| \leq e') = P_0 = 1 - \alpha$$

- Let $Pr(Y \leq Z_\beta) = 1 - \beta$. Z_β can be found from a standard table. Then, e' can be found as
$$e' = Z_{\alpha/2}$$

Characterization of the value of s (continued)

- Accordingly, we have

$$\begin{aligned}&Pr(|Y| \leq Z_{\alpha/2})\\&= Pr(|(\overline{X} - s)\sqrt{n}/\sigma| \leq Z_{\alpha/2})\\&= Pr(|(\overline{X} - s) \leq Z_{\alpha/2}\,\sigma/\sqrt{n})\end{aligned}$$

- Thus, we can have

$$e = Z_{\alpha/2}\,\sigma/\sqrt{n}$$

  where σ is unknown.

- We can substitute δ_X for σ, but that will not work because the distribution of the random variable $(\overline{X} - s)\sqrt{n}/\delta_X$ is unknown and may differ substantially from the normal distribution.

Characterization of the value of s (continued)

- T get around this difficulty, we assume that the distribution of each X_i itself is normal, i.e., $\mathcal{N}(s, \sigma)$. Then, $Y = (\overline{X} - s)\sqrt{n}/\delta_X$ has the standard *t-distribution* with $(n-1)$ degrees of freedom. We denote the latter as $\Phi_{t,n-1}(.)$.

- Let $Pr(Y \leq t_{n-1,\beta}) = 1 - \beta$. $t_{n-1,\beta}$ can be found from a standard table.

- Then, we can write

$$Pr(|Y| \leq t_{n-1,\alpha/2}) = 1 - \alpha$$

Characterization of the value of s (continued)

- Accordingly, we get

$$Pr(|(\overline{X} - s)\sqrt{n}\,/\delta_X| \leq t_{n-1,\alpha/2}) = 1 - \alpha$$

- We can put the above equation in the following alternate form

$$Pr[\overline{X} - \eta \leq s \leq \overline{X} + \eta] = 1 - \alpha$$

where

$$\eta = \frac{\delta_X\, t_{n-1,\alpha/2}}{\sqrt{n}}$$

Characterization of the value of s (continued)

- The last formula can be used in two ways:
  - to determine confidence interval for a given number of observations, or
  - to determine the number of observations needed to achieve a given confidence interval.
- For the latter, suppose that the desired error (i.e., fractional half-width of the confidence interval) is q. Then

$$\frac{\delta_X \, t_{n-1,\alpha/2}}{\sqrt{n}} \leq q\overline{X} \Rightarrow n \geq \frac{\delta_X^2 \, t_{n-1,\alpha/2}^2}{q^2 \overline{X}^2}$$

Characterization of the value of s (continued)

- For the latter, suppose that the desired error (i.e., fractional half-width of the confidence interval) is q. Then

$$\frac{\delta_X\, t_{n-1,\alpha/2}}{\sqrt{n}} \leq q\overline{X} \Rightarrow n \geq \frac{\delta_X^2\, t_{n-1,\alpha/2}^2}{q^2 \overline{X}^2}$$

- Since δ_X, $\overline{X}$, and $t_{n-1,\alpha/2}$ depend on n, we should first "guess" some value for n and determine δ_X, $\overline{X}$, and $t_{n-1,\alpha/2}$. Then, we can check if the above equation is satisfied. If it is not, more observations should be made.

Characterization of the value of s (continued)

- In the previous cases, we considered a two-sided confidence interval. In some applications, we only want to find out whether the performance measure of interest exceeds (or remains below) some given threshold.

- For example, to assert that the actual value s exceeds some threshold $\overline{X} - e$, let $Y = (\overline{X} - s)\sqrt{n}\,/\delta_X$. Then

$$Pr(s \geq \overline{X} - e) = P_0 = 1 - \alpha$$

Characterization of the value of s (continued)

- Accordingly, we get

  $$Pr(Y \leq e') = 1 - \alpha$$

  where $e' = t_{n-1,\alpha}$.

- Thus, we find

  $$e = \frac{\delta_X \, t_{n-1,\alpha}}{\sqrt{n}}$$

Example:

Five independent experiments were conducted for determining the average flow rate of the coolant discharged by the cooling system. One hundred observations were taken in each experiment, the means of which are reported below

$$3.07 \;\; 3.24 \;\; 3.14 \;\; 3.11 \;\; 3.07$$

Based on this data, could we say that the mean flow rate exceeds 3.00 at a confidence level of 99.5%? What happens if we degrade the confidence level to 97.5%?

Solution:

- The sample mean and sample standard deviation can be calculated from the data as: $\overline{X} = 3.126$, $\delta_X = 0.0702$.

- From the table, we get $t_{4,0.005} = 4.604$. Thus, we have:

$$\begin{aligned} & Pr(Y \leq 4.604) \\ & = Pr[(\overline{X} - s)\sqrt{n}\,/\delta_X \leq 4.604] \\ & = P[(3.126 - s)\sqrt{5}/0.0702 \leq 4.604] \\ & = Pr(s \geq 2.9815) \\ & = .995 \end{aligned}$$

- Therefore, with the confidence level of 0.995, we cannot be sure that the the flow rate exceeds 3.00.

Solution: (continued)

- The sample mean and sample standard deviation are the same as before.

- From the table, we get $t_{4,0.025} = 2.776$. Thus, we have:

$$
\begin{aligned}
& Pr(Y \leq 2.776) \\
& = Pr[(\overline{X} - s)\,\sqrt{n}\,/\delta_X \ \leq 2.776] \\
& = P[(3.126 - s)\sqrt{5}/0.0702 \leq 2.776] \\
& = Pr(s \geq 3.039) \\
& = .975
\end{aligned}
$$

- Therefore, with the confidence level of 0.975, we can be sure that the the flow rate exceeds 3.00.

Regression Analysis

- Let X and Y denote the input and output parameters of interest. Let X_i, $i = 1 \cdots n$ denote an increasing set of values of the input parameter, and Y_i, $i = 1 \cdots n$ the corresponding *observed values* of the output parameters.

- Then we want to determine a function $Y = f(X)$ that is consistent with these observations.

- Because of the effect of uncontrolled variables and measurement errors, we will not observe the true value $f(X_i)$ at point X_i. Instead, what we get is

$$Y_i = f(X_i) + \epsilon_i$$

  where ϵ_i is a random variable representing the unknown error such that $E(\epsilon_i) = 0$.

Regression Analysis (continued)

- Let $\alpha_1, \cdots, \alpha_k$ denote the k unknown parameters of the assumed function f. For clarity, we will write f as $f(X; \alpha_1, \cdots, \alpha_k)$. Presumably, α_j's have some *actual values*, that we do not know. All we can do is to estimate their values from the data. We shall denote the estimated values by using the circumflex (ˆ) symbol. Thus $\widehat{f}(X) = f(X; \widehat{\alpha_1}, \cdots, \widehat{\alpha_k})$.

Regression Analysis (continued)

- As for Y's, there are three types of values to consider:

  1. Actual values: denoted $\mathcal{Y}$, e.g., $\mathcal{Y}_i = f(X_i; \alpha_1, \cdots, \alpha_k)$.

  2. Observed values Y_i's, related to $\mathcal{Y}_i$ as $Y_i = \mathcal{Y}_i + \epsilon_i$.

  3. Estimated values: denoted $\widehat{Y}$, e.g., $\widehat{Y}_i = \widehat{f}(X_i) = f(X_i; \widehat{\alpha_1}, \cdots, \widehat{\alpha_k})$.

Regression Analysis (continued)

- The total variation of the observed values about the estimated values is given by Q_E,

$$Q_E = \sum_{i=1}^{n} [Y_i - \widehat{f}(X_i)]^2$$

  The estimation now involves finding values of α_i's such that Q_E is minimized. The resulting estimate is called the *regression* of Y over X.

Gauss-Markov Theorem

The least squares method yields an unbiased estimator of α_i's, and minimizes variance in estimated values if the following conditions are satisfied:

- $f(X)$ is linear in α_i's. That is, $f(X) = \alpha_1 g_1(X) + \cdots + \alpha_k g_k(X)$, where g_i's are arbitrary but fully known functions.
- There is no uncertainty (or error) in the values of X_i's.
- Error in observed values of Y (i.e. ϵ_i) has zero mean.
- All measurements are uncorrelated.

Gauss-Markov Theorem (continued)

- It can be shown that

  $$Var(\hat{Y}) = \sum_{i=1}^{k} g_i^2(X) Var(\widehat{\alpha}_i)$$

  Thus, minimum variance estimation of α_i's implies minimum variance for $\hat{Y}$.

- From the definition of Q_E, we have

  $$Q_E = \sum_{i=1}^{n} [Y_i - \widehat{\alpha}_1 g_1(X_i) - \cdots - \widehat{\alpha}_k g_k(X_i)]^2$$

Gauss-Markov Theorem (continued)

- To find the global minima, we set the partial derivatives with respect to $\widehat{\alpha_j}$'s to zero. Thus, we have

  $$\sum_{i=1}^{n} g_j(X_i)\left[Y_i - \widehat{\alpha_1} g_1(X_i) - \cdots - \widehat{\alpha_k} g_k(X_i)\right] = 0,$$

  for $j = 1 \cdots k$.

- The above equations can be put in the following matrix form

  $$\Pi\alpha = \theta$$

  where $\alpha = [\alpha_1, \cdots, \alpha_k]$ and $\theta = [\theta_1, \cdots, \theta_k]$ are column vectors, and $\Pi = [\pi_{jm}]$ is a $k \times k$ matrix. The elements of $\mathbf{\Pi}$ and θ are defined as follows:

  $$\pi_{jm} = \sum_{i=1}^{n} g_j(X_i) g_m(X_i) \text{ and } \theta_j = \sum_{i=1}^{n} g_j(X_i) Y_i$$

Example:

Suppose that the hypothesized “actual” function is quadratic and is given by

$$\mathcal{Y} = f(x) = a_1 + a_2 x + a_3 x^2$$

Find expressions for estimated values of a_i’s.

Solution:

Throughout this problem we assume that the summations are over i ranging from 1 to n. Let $z_j = \sum x_i^j$ for $j = 1 \cdots 4$. Then, teh equations satisfied by a_i's are as follows:

$$\begin{bmatrix} n & z_1 & z_2 \\ z_1 & z_2 & z_3 \\ z_2 & z_3 & z_4 \end{bmatrix} \begin{bmatrix} \widehat{a_1} \\ \widehat{a_2} \\ \widehat{a_3} \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum x_i^2 Y_i \end{bmatrix}$$

from which we can get expressions for a_i's. It is also easy to show from here that $E[\widehat{a_1}] = a_1$, i.e., the estimates are unbiased.

Linear Regression

- Linear regression arises in practice quite often. Let $\mathcal{Y} = \alpha + \beta x$, $\overline{x} = \sum_i x_i/n$, and $\overline{Y} = \sum_i Y_i/n$. It is easy to verify that the following estimates can be derived for α and β.

$$\widehat{\beta} = \sum_{i=1}^{n}(x_i - \overline{x})Y_i / \sum_{i=1}^{n}(x_i^2 - \overline{x}^2)$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{x}$$

- Since $\hat{Y} = \widehat{\alpha} + \widehat{\beta}x$, from the last equation above, we get

$$\hat{Y} = \overline{Y} + \widehat{\beta}(x - \overline{x})$$

Linear Regression (continued)

Let $\overline{\mathcal{Y}}$ denote the sample average of $\mathcal{Y}_i$'s. Since $\mathcal{Y}_i = \alpha + \beta x_i$, we have $\overline{\mathcal{Y}} = \alpha + \beta \overline{x}$. Thus, we get

$$\mathcal{Y}_i = \overline{\mathcal{Y}} + \beta(x_i - \overline{x})$$

Let $z_i = x_i - \overline{x}$ and $\gamma = \sum_{i=1}^{n} z_i^2$. Since $Y_i = \mathcal{Y}_i + \epsilon_i$, $E[Y_i] = \mathcal{Y}_i$. We can show that the estimates of α and β are unbiased as follows:

- $$E[\widehat{\beta}] = \frac{1}{\gamma}\sum_{i=1}^{n} z_i E[Y_i] = \frac{1}{\gamma}\sum_{i=1}^{n} z_i \mathcal{Y}_i$$
  $$= \frac{1}{\gamma}[\sum_{i=1}^{n} \overline{\mathcal{Y}} z_i + \sum_{i=1}^{n} \beta z_i^2] = \beta$$

- $$E(\widehat{\alpha}) = E[\overline{Y} - \widehat{\beta}\overline{x}] = \overline{Y} - \beta\overline{x} = \alpha$$

  which also means that $\widehat{Y}$ is an unbiased estimate of $\mathcal{Y}$, that is $E(\widehat{Y}) = \mathcal{Y}$.