Applied Statistical Design of Experiments: Applications in Natural Sciences

Ijaz A Rauf*

Department of Physics and Astronomy, 128 Petrie Science and Engineering Building, York University, 4700 Keele Street, Toronto, Ontario, M3J 1P3, Canada

Experiments are performed today in many scientific and engineering research fields to increase our understanding and knowledge of various scientific principles and processes. These experiments are often conducted in a series of trials or tests, which produce quantifiable outcomes. The quantitative as well as qualitative outcomes of these tests or trials are measured using scientific instruments as well as measurement equipment.

The most common approach to experimentation employed by many scientists today is One-Variable-At-a-Time (OVAT), where one variable is changed at a time keeping all other variables in the experiment fixed (or constant) and the change in the resulting outcome is observed. Quantifiable outcomes of the experiments are often measured using scientific instruments/equipment like mass spectroscopy, atomic force microscopy, or electron microscopy. The changes or variation observed in the outcome of an experiment may stem from two sources: a) A direct consequence of the intentional change in the input variable since the outcome variable is correlated to the changing input variable and/or b) error or variation in the measurement system.

The performance of an instrument or measurement equipment is influenced by several factors including natural variation of the instrument/equipment and the human factor. When dealing with science, materials and devices on a nanoscale and operating the measurement equipment near or at the performance limit, it becomes critically important that all sources of variability are properly quantified to draw correct logical conclusions.

When several variables influence a certain characteristic of a process or principle, the best strategy is then to design an experiment so that valid, reliable and sound conclusions can be drawn effectively, efficiently and economically. Statistical Design of Experiments (SDoE) is a methodology where multiple variables can be changed and tested in parallel and the impact of each variable can be quantified independent of other variables using quantitative statistical tools.

Inferential statistical analysis of experimental data plays a pivotal role in transforming the observed data into critical information for the advancement of underlying science and accelerated achievement of the research goals in the field of interest. The recent drastic funding cuts for resources to perform only 4 experiments, a scientist would choose to make one measurement at four levels spanning the region of interest. This does not allow the quantification of the error at each of the points and does not establish if the changes observed are due to changes in the experimental variables or caused by the measurement error (or variation).

The statistical design of experiments data when analysed using statistical tools produces two set of models; the structural model and the error model. The relationship between the output (or dependent) variable ‘y’ and input (or independent) variables ‘z’ especially when y is dependent on more than one variables can be expressed as:

\[ y = a x + b + e \]

(1)

Where f is an unknown function, which may be very complicated, and e represents other non-systematic sources of variability not accounted for in function f, such as measurement error. The goal is to approximate f by a relatively simple analytical function on the basis of experimental data. The variables \( z_1, z_2, \ldots, z_n \) are natural variable (for example, time and temperature in a synthesis experiment) that are measured in units used in the experiments (for example, seconds for time etc.).

In an experimental design the natural variables are converted into

\[ y(z_1, z_2, \ldots, z_n) + \epsilon \]

(2)

Received: September 29, 2015; Accepted: October 02, 2015; Published: October 05, 2015


Copyright: © 2015 Rauf IA. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
dimensionless coded variables $x_1, x_2, \ldots, x_n$ on a scale of -1 (the lowest value) to +1 (the highest value) as follows:

$$x_i = \frac{z_i - (z_{a1} + z_{a2})/2}{(z_{a1} - z_{a2})/2} \quad (3)$$

This converts all variables to a common scale and allows evaluation of the impact of each variable and its interaction on the outcome independent of its measurement units. The simplest model of the response surface can be approximated by multiple linear regression with k input variables as follows:

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \epsilon \quad (4)$$

Where the linear regression coefficients, $\beta$, are estimated from the experimental data through model fitting. When the change in the output variable due to a change in an input variable $x_j$ depends on the values of other variables, this can be expressed by as:

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{i=2}^{n} \sum_{j=1}^{k} \beta_{ij} x_i x_j + \epsilon \quad (5)$$

Here the term $\beta_{ij} x_i x_j$ represents the statistical interaction between variable $x_i$ and $x_j$. Both equations (4) and (5) are linear terms and thus represent the first order model. The relationship between variables may not always be linear and a more complex model may need to be developed. If a strong quadratic relationship is expected this can be expressed as a 2nd order model:

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{i=2}^{n} \sum_{j=1}^{k} \beta_{ij} x_i x_j + \sum_{j=1}^{k} \beta_{jj} x_j^2 + \epsilon \quad (6)$$

Parameters $\beta$ are often estimated from experimental data using computation software designed for this purpose and are donated by $\hat{\beta}$, the values predicted by (4) for $y_i$ based on the input variables of $x_1, x_2, \ldots, x_n$ is given as:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j + \epsilon \quad (7)$$

The difference between observed and predicted value is called residual and is useful in making inferences about the adequacy of the model. The residuals are calculated as:

$$\hat{e} = y_i - \hat{y}_i \quad (8)$$

In summary, statistical design of experiments and its analytical tools have an application niche in academic and scientific research and thus scientists should be trained in these concepts and take advantage of its application for faster advancement of science at the lowest cost. In contrast to OVAT approach SDoE addresses the issues of interactions of variable and can also quantitatively address the confounding problems. SDoE approach thus may reveal previously unknown scientific facts about a scientific process or principle.

References