Experimental and Quasi-Experimental Designs for Research

Design of tasks set to uncover from

The design of experiments (DOE, DOX, or experimental design) is the design of any task that aims to describe and explain the variation of information under conditions that are hypothesized to reflect the variation. The term is generally associated with experiments in which the design introduces conditions that directly affect the variation, but may also refer to the design of quasi-experiments, in which natural conditions that influence the variation are selected for observation.

In its simplest form, an experiment aims at predicting the outcome by introducing a change of the preconditions, which is represented by one or more independent variables, also referred to as "input variables" or "predictor variables." The change in one or more independent variables is generally hypothesized to result in a change in one or more dependent variables, also referred to as "output variables" or "response variables." The experimental design may also identify control variables that must be held constant to prevent external factors from affecting the results. Experimental design involves not only the selection of suitable independent, dependent, and control variables, but planning the delivery of the experiment under statistically optimal conditions given the constraints of available resources. There are multiple approaches for determining the set of design points (unique combinations of the settings of the independent variables) to be used in the experiment.

Main concerns in experimental design include the establishment of validity, reliability, and replicability. For example, these concerns can be partially addressed by carefully choosing the independent variable, reducing the risk of measurement error, and ensuring that the documentation of the method is sufficiently detailed. Related concerns include achieving appropriate levels of statistical power and sensitivity.

Correctly designed experiments advance knowledge in the natural and social sciences and engineering. Other applications include marketing and policy making. The study of the design of experiments is an important topic in metascience.

History [edit]

Statistical experiments, following Charles S. Peirce [edit]

A theory of statistical inference was developed by Charles S. Peirce in "Illustrations of the Logic of Science" (1877–1878)[1] and "A Theory of Probable Inference" (1883),[2] two publications that emphasized the importance of randomization-based inference in statistics.[3]

Randomized experiments [edit]

Charles S. Peirce randomly assigned volunteers to a blinded, repeated-measures design to evaluate their ability to discriminate weights.[4][5][6][7] Peirce's

experiment inspired other researchers in psychology and education, which developed a research tradition of randomized experiments in laboratories and specialized textbooks in the 1800s.[4][5][6][7]

Optimal designs for regression models [edit]

Charles S. Peirce also contributed the first English-language publication on an optimal design for regression models in 1876.[8] A pioneering optimal design for polynomial regression was suggested by Gergonne in 1815. In 1918, Kirstine Smith published optimal designs for polynomials of degree six (and less).[9][10]

Sequences of experiments [edit]

The use of a sequence of experiments, where the design of each may depend on the results of previous experiments, including the possible decision to stop experimenting, is within the scope of sequential analysis, a field that was pioneered[11] by Abraham Wald in the context of sequential tests of statistical hypotheses.[12] Herman Chernoff wrote an overview of optimal sequential designs,[13] while adaptive designs have been surveyed by S. Zacks.[14] One specific type of sequential design is the "two-armed bandit", generalized to the multi-armed bandit, on which early work was done by Herbert Robbins in 1952.[15]

Fisher's principles [edit]

A methodology for designing experiments was proposed by Ronald Fisher, in his innovative books: The Arrangement of Field Experiments (1926) and The Design of Experiments (1935). Much of his pioneering work dealt with agricultural applications of statistical methods. As a mundane example, he described how to test the lady tasting tea hypothesis, that a certain lady could distinguish by flavour alone whether the milk or the tea was first placed in the cup. These methods have been broadly adapted in biological, psychological, and agricultural research.[16]

Comparison In some fields of study it is not possible to have independent measurements to a traceable metrology standard. Comparisons between treatments are much more valuable and are usually preferable, and often compared against a scientific control or traditional treatment that acts as baseline.

Randomization Random assignment is the process of assigning individuals at random to groups or to different groups in an experiment, so that each individual of the population has the same chance of becoming a participant in the study. The random assignment of individuals to groups (or conditions within a group) distinguishes a rigorous, "true" experiment from an observational study or "quasi-experiment".[17] There is an extensive body of mathematical theory that explores the consequences of making the allocation of units to treatments by means of some random mechanism (such as tables of random numbers, or the use of randomization devices such as playing cards or dice). Assigning units to treatments at random tends to mitigate confounding, which makes effects due to factors other than the treatment to appear to result from the treatment.

The risks associated with random allocation (such as having a serious imbalance in a key characteristic between a treatment group and a control group) are calculable and hence can be managed down to an acceptable level by using enough experimental units. However, if the population is divided into several subpopulations that somehow differ, and the research requires each subpopulation to be equal in size, stratified sampling can be used. In that way, the units in each subpopulation are randomized, but not the whole sample. The results of an experiment can be generalized reliably from the experimental units to a larger statistical population of units only if the experimental units are a random sample from the larger population; the probable error of such an extrapolation depends on the sample size, among other things.

Statistical replication Measurements are usually subject to variation and measurement uncertainty; thus they are repeated and full experiments are replicated to help identify the sources of variation, to better estimate the true effects of treatments, to further strengthen the experiment's reliability and validity, and to add to the existing knowledge of the topic.[18] However, certain conditions must be met before the replication of the experiment is commenced: the original research question has been published in a peer-reviewed journal or widely cited, the researcher is independent of the original experiment, the researcher must first try to replicate the original findings using the original data, and the write-up should state that the study conducted is a replication study that tried to follow the original study as strictly as possible.[19]

Blocking Blocking is the non-random arrangement of experimental units into groups (blocks) consisting of units that are similar to one another. Blocking reduces known but irrelevant sources of variation between units and thus allows greater precision in the estimation of the source of variation under study.

Example of orthogonal factorial design

Orthogonality concerns the forms of comparison (contrasts) that can be legitimately and efficiently carried out. Contrasts can be represented by vectors and sets of orthogonal contrasts are uncorrelated and independently distributed if the data are normal. Because of this independence, each orthogonal treatment provides different information to the others. If there are T treatments and T – 1 orthogonal contrasts, all the information that can be captured from the experiment is obtainable from the set of contrasts.

Factorial experiments Use of factorial experiments instead of the one-factor-at-a-time method. These are efficient at evaluating the effects and possible interactions of several factors (independent variables). Analysis of experiment design is built on the foundation of the analysis of variance, a collection of models that partition the observed variance into components, according to what factors the experiment must estimate or test.

Example [edit]

This example of design experiments is attributed to Harold Hotelling, building on examples from Frank Yates.[20][21][13] The experiments designed in this example involve combinatorial designs.[22]

Weights of eight objects are measured using a pan balance and set of standard weights. Each weighing measures the weight difference between objects in the left pan and any objects in the right pan by adding calibrated weights to the lighter pan until the balance is in equilibrium. Each measurement has a random error. The average error is zero; the standard deviations of the probability distribution of the errors is the same number σ on different weighings; errors on different weighings are independent. Denote the true weights by

$\theta_1, \dots, \theta_8.$ {displaystyle theta _{1},dots ,theta _{8}.,}

We consider two different experiments:

Weigh each object in one pan, with the other pan empty. Let X i be the measured weight of the object, for i = 1, ..., 8. Do the eight weighings according to the following schedule—a weighing matrix—and let Y i be the measured difference for i = 1, ..., 8:

left pan right pan 1st weighing: 1 2 3 4 5 6 7 8 (empty) 2nd: 1 2 3 8 4 5 6 7 3rd: 1 4 5 8 2 3 6 7 4th: 1 6 7 8 2 3 4 5 5th: 2 4 6 8 1 3 5 7 6th: 2 5 7 8 1 3 4 6 7th: 3 4 7 8 1 2 5 6 8th: 3 5 6 8 1 2 4 7

$$\begin{array}{lcc}&\text{left pan}&\text{right pan}\\ \hline \text{1st weighing:}&1\ 2\ 3\ 4\ 5\ 6\ 7\ 8&\text{(empty)}\\ \text{2nd:}&1\ 2\ 3\ 8 &4\ 5\ 6\ 7\\ \text{3rd:}&1\ 4\ 5\ 8 &2\ 3\ 6\ 7\\ \text{4th:}&1\ 6\ 7\ 8 &2\ 3\ 4\ 5\\ \text{5th:}&2\ 4\ 6\ 8 &1\ 3\ 5\ 7\\ \text{6th:}&2\ 5\ 7\ 8 &1\ 3\ 4\ 6\\ \text{7th:}&3\ 4\ 7\ 8 &1\ 2\ 5\ 6\\ \text{8th:}&3\ 5\ 6\ 8 &1\ 2\ 4\ 7 \end{array}$$

Then the estimated value of the weight θ_1 is

$\theta^1 = Y1 + Y2 + Y3 + Y4 - Y5 - Y6 - Y7 - Y8$. $$\displaystyle \widehat{\theta}_{1}=\frac{Y_{1}+Y_{2}+Y_{3}+Y_{4}-Y_{5}-Y_{6}-Y_{7}-Y_{8}}{8}.$$

Similar estimates can be found for the weights of the other items. For example

$\theta^2 = Y1 + Y2 - Y3 - Y4 + Y5 + Y6 - Y7 - Y8$. $\theta^3 = Y1 + Y2 - Y3 - Y4 - Y5 - Y6 + Y7 + Y8$. $\theta^4 = Y1 - Y2 + Y3 - Y4 + Y5 - Y6 + Y7 - Y8$. $\theta^5 = Y1 - Y2 + Y3 - Y4 - Y5 + Y6 - Y7 + Y8$. $\theta^6 = Y1 - Y2 - Y3 + Y4 + Y5 - Y6 - Y7 + Y8$. $\theta^7 = Y1 - Y2 - Y3 + Y4 - Y5 + Y6 + Y7 - Y8$. $\theta^8 = Y1 + Y2 + Y3 + Y4 + Y5 + Y6 + Y7 + Y8$.

$$\displaystyle \begin{aligned}\widehat{\theta}_{2}&=\frac{Y_{1}+Y_{2}-Y_{3}-Y_{4}+Y_{5}+Y_{6}-Y_{7}-Y_{8}}{8}.\\[5pt]\widehat{\theta}_{3}&=\frac{Y_{1}+Y_{2}-Y_{3}-Y_{4}-Y_{5}-Y_{6}+Y_{7}+Y_{8}}{8}.\\[5pt]\widehat{\theta}_{4}&=\frac{Y_{1}-Y_{2}+Y_{3}-Y_{4}+Y_{5}-Y_{6}+Y_{7}-Y_{8}}{8}.\\[5pt]\widehat{\theta}_{5}&=\frac{Y_{1}-Y_{2}+Y_{3}-Y_{4}-Y_{5}+Y_{6}-Y_{7}+Y_{8}}{8}.\\[5pt]\widehat{\theta}_{6}&=\frac{Y_{1}-Y_{2}-Y_{3}+Y_{4}+Y_{5}-Y_{6}-Y_{7}+Y_{8}}{8}.\\[5pt]\widehat{\theta}_{7}&=\frac{Y_{1}-Y_{2}-Y_{3}+Y_{4}-Y_{5}+Y_{6}+Y_{7}-Y_{8}}{8}.\\[5pt]\widehat{\theta}_{8}&=\frac{Y_{1}+Y_{2}+Y_{3}+Y_{4}+Y_{5}+Y_{6}+Y_{7}+Y_{8}}{8}.\end{aligned}$$

The question of design of experiments is: which experiment is better?

The variance of the estimate X 1 of θ_1 is σ^2 if we use the first experiment. But if we use the second experiment, the variance of the estimate given above is $\sigma^2/8$. Thus the second experiment gives us 8 times as much precision for

the estimate of a single item, and estimates all items simultaneously, with the same precision. What the second experiment achieves with eight would require 64 weighings if the items are weighed separately. However, note that the estimates for the items obtained in the second experiment have errors that correlate with each other.

Many problems of the design of experiments involve combinatorial designs, as in this example and others.[22]

Avoiding false positives [edit]

False positive conclusions, often resulting from the pressure to publish or the author's own confirmation bias, are an inherent hazard in many fields. A good way to prevent biases potentially leading to false positives in the data collection phase is to use a double-blind design. When a double-blind design is used, participants are randomly assigned to experimental groups but the researcher is unaware of what participants belong to which group. Therefore, the researcher can not affect the participants' response to the intervention. Experimental designs with undisclosed degrees of freedom are a problem.[23] This can lead to conscious or unconscious "p-hacking": trying multiple things until you get the desired result. It typically involves the manipulation â€" perhaps unconsciously â€" of the process of statistical analysis and the degrees of freedom until they return a figure below the p

Reference

Observation and Experiment: An Introduction to Causal Inference

Methods in Behavioral Research