Design of Experiments

DOE

Radu T. Trîmbiţaş

April 20, 2016

1 Introduction

The Elements Affecting the Information in a Sample

- Generally, the design of experiments (DOE) is a very broad subject concerned with methods of sampling to reduce the variation in an experiment and thereby to acquire a specified quantity of information at minimum cost.
- The width of a CI for mean

  $$\overline{Y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

  σ and n affect this
- If we wish to compare two populations, based on a total of n observations, how many observations should be taken from each population?
- If we have decided to fit a simple linear regression model and wish to maximize the information in the resulting data, how should we choose the values of the independent variable?
- Example: how to find the sample size if we know the maximum error E, α and the proportion?

  $$E = |z_{\alpha/2}| \sqrt{\frac{pq}{n}} \Longrightarrow n = \frac{|z_{\alpha/2}|^2 pq}{E^2}$$

2 Designing Experiments to Increase Accuracy

Designing Experiments to Increase Accuracy

- For the same total number of observations, some methods of data collection (designs) provide more information concerning specific population parameters than others.

- Consider the problem of estimating the difference between a pair of population means, $\mu_1 - \mu_2$, based on independent random samples.
- How many observations should she select from populations 1 and 2—say, n_1 and n_2 ($n_1 + n_2 = n$), respectively—to maximize the information in the data pertinent to $\mu_1 - \mu_2$?

Example 1. If n observations are to be used to estimate $\mu_1 - \mu_2$,, based on independent random samples from the two populations of interest, find n_1 and n_2 so that $V\left(\overline{Y}_1 - \overline{Y}_2\right)$ is minimized (assume that $n_1 + n_2 = n$).

Let b denote the fraction of the n observations assigned to the sample from population 1; that is, $n_1 = bn$ şi $n_2 = (1-b)n$. Then,

$$V\left(\overline{Y}_1 - \overline{Y}_2\right) = \frac{\sigma_1^2}{bn} + \frac{\sigma_2^2}{(1-b)n}.$$

To find b that minimizes the variance, we set the derivative with respect to b to zero. this yields

$$\frac{1}{b^2 n}\sigma_1^2 - \frac{1}{n}\frac{\sigma_2^2}{(b-1)^2} = 0$$

Solving for b, we obtain

$$b = \frac{\sigma_1}{\sigma_1 + \sigma_2},\ 1 - b = \frac{\sigma_2}{\sigma_1 + \sigma_2}.$$

Thus, $V\left(\overline{Y}_1 - \overline{Y}_2\right)$ is minimized for

$$n_1 = \frac{\sigma_1}{\sigma_1 + \sigma_2} n,\ \ n_2 = \frac{\sigma_2}{\sigma_1 + \sigma_2} n,$$

i.e., when sample sizes are proportional to the standard deviations. Note that, if $\sigma_1 = \sigma_2$, then $n_1 = n_2 = n/2$.

Simple Linear Regression

- Suppose that we are primarily interested in the slope β_1 of the line in the linear model
$$Y = \beta_0 + \beta_1 X + \varepsilon.$$
- If we have the option of selecting the n-values of x for which y will be observed, which values of x will maximize the quantity of information on β_1?
- The best design for estimating the slope β_1 can be determined by considering the standard deviation of $\widehat{\beta}_1$:
$$\sigma_{\widehat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

- The larger S_{xx}, the sum of squares of deviations of $x_1, x_2, \ldots, x_n$ about their mean, the smaller the standard deviation of $\widehat{\beta}_1$ will be. That is, we obtain a better estimator for the slope if the values of x are spread farther apart. In some cases, the experimenter has some experimental region—say, $x_1 < x < x_2$ — over which he or she wishes to observe Y, and this range is frequently selected prior to experimentation.

- The smallest value for $\sigma_{\widehat{\beta}_1}$ occurs when the n data points are equally divided, with half located at x_1 and half at x_2. (Proof, homework.)An experimenter who wished to fit a line by using $n = 10$ data points in the interval $2 \leq x \leq 6$ would select five data points at $x = 2$ and five at $x = 6$.

- Before concluding the discussion of this example, you should notice that observing all values of Y at only two values of x will not provide information on curvature of the response curve in case the assumption of linearity in the relation of $E(Y)$ and x is incorrect. It is frequently safer to select a few points (as few as one or two) somewhere near the middle of the experimental region to detect curvature if it should be present (see Figure 3).

- A further comment is in order. One of the assumptions that we have made regarding the simple linear regression model is that the variance of the error term ε does not depend on the value of the independent variable x. If the x values are more spread out, the validity of this assumption may become more questionable.

3 The Matched-Pairs Experiment

The Matched-Pairs Experiment

- A commonly occurring situation – repeated observations are made on the same sampling unit: weighing the same individual before and after a weight-loss program; in a medical experiment, pair of individuals of similar gender, weights and ages: one individual from each pair is randomly selected to receive one of two competing medications to control hypertension whereas the other individual from the same pair receives the other medication.

- Comparing two populations on the basis of paired data can be a very effective experimental design that can control for extraneous sources of variability and result in decreasing the standard error of the estimator $\overline{Y}_1 - \overline{Y}_2$ for the difference in the population means $\mu_1 - \mu_2$.

- (Y_{1i}, Y_{2i}), $i = 1, 2, \ldots, n$, denote a random sample of paired observations. Assume that

$$\begin{aligned} &E(Y_{1i}) = \mu_1, \qquad V(Y_{1i}) = \sigma_1^2, \qquad \mathrm{Cov}(Y_{1i}, Y_{2i}) = \rho\sigma_1\sigma_2, \\ &E(Y_{2i}) = \mu_2, \qquad V(Y_{2i}) = \sigma_2^2, \end{aligned}$$

  ρ is the common correlation coefficient of the variables within each pair

- Define $D_i = Y_{1i} - Y_{2i}$, for $i = 1, 2, \ldots, n$, D_is are IID and

$$\begin{aligned} \mu_D &= E(D_i) = E(Y_{1i}) - E(Y_{2i}) = \mu_1 - \mu_2 \\ \sigma_D^2 &= V(D_i) = V(Y_{1i}) + V(Y_{2i}) - 2\mathrm{Cov}(Y_{1i}, Y_{2i}) \\ &= \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2. \end{aligned}$$

- A natural estimator for $\mu_1 - \mu_2$ is the average of the differences $\overline{D} = \overline{Y}_1 - \overline{Y}_2$, and

$$\begin{aligned} E(\overline{D}) &= \mu_D = \mu_1 - \mu_2 \\ \sigma_{\overline{D}}^2 = V(\overline{D}) &= \frac{\sigma_D^2}{n} = \frac{1}{n}\left[\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right]. \end{aligned}$$

- If the data had been obtained from an independent samples experiment and $n_1 = n_2 = n$,

$$\begin{aligned} E(\overline{D}) &= \mu_1 - \mu_2 \\ \sigma_{\overline{Y}_1 - \overline{Y}_2}^2 &= \frac{1}{n}\left[\sigma_1^2 + \sigma_2^2\right] \end{aligned}$$

- If $\rho > 0$, then $\sigma_{\overline{D}}^2 < \sigma_{\overline{Y}_1 - \overline{Y}_2}^2$ (for example when the values of Y_{1i} and Y_{2i} will tend to increase or decrease together)

Sample	Method 1	Method 2	d_i
1	38.25	38.27	$-.02$
2	31.68	31.71	$-.03$
3	26.24	26.22	$+.02$
4	41.29	41.33	$-.04$
5	44.81	44.80	$+.01$
6	46.37	46.39	$-.02$
7	35.42	35.46	$-.04$
8	38.41	38.39	$+.02$
9	42.68	42.72	$-.04$
10	46.71	46.76	$-.05$
11	29.20	29.18	$+.02$
12	30.76	30.79	$-.03$
			$\overline{d} = -0.0167$

Table 1: Data for the matched-pairs experiment in Example 2

- Assumptions: the differences D_i must be normally distributed. This does not mean the two populations are normally distributed
- Counterexample: Suppose matched pairs and the i-th measurement ($i = 1, 2$), in the j-th pair , $j = 1, 2, \ldots, n$, is

  $$Y_{ij} = \mu_i + U_j + \varepsilon_{ij},$$

  where μ_i = mean value of population i, $i = 1, 2$, U_j = RV distributed $U[-1, +1]$, ε_{ij} = error for measurement i pair j. Suppose ε_{ij} are normally distr. indep. RVs with $E(\varepsilon_{ij}) = 0$ and $V(\varepsilon_{ij}) = \sigma^2$ and U_j and ε_{ij} are independant. Then Y_{ij} are not normal but differences $D_j = Y_{1j} - Y_{2j}$ are independent, normally distributed RVs.

Example

Example 2. We wish to compare two methods for determining the percentage of iron ore in ore samples. Because inherent differences in the ore samples would be likely to contribute unwanted variability in the measurements that we observe, a matched pairs experiment was created by splitting each of 12 ore samples into two parts. One-half of each sample was randomly selected and subjected to method 1; the other half was subjected to method 2. The results are presented in Table 12.1. Do the data provide sufficient evidence that method 2 yields a higher average percentage than method 1? Test using $\alpha = .05$.

Solution. See `ex12_2.pdf` □

4 Some Elementary Experimental Designs

4.1 Basic notions

Example

- Suppose that we wish to compare five teaching techniques, A, B, C, D, and E, and that we use 125 students in the study.
- The objective is to compare the mean scores on a standardized test for students taught by each of the five methods.
- Sources of variability:
  - gender
  - differences in the native abilities of the students in the group
  - different students may come from families that place different emphases on education, and this could have an impact on the scores on the standardized test.
- We decide that it might be wise to randomly assign 25 students to each of five groups. Each group will be taught using one of the techniques under study.
- Objectives of the random division of the students into the five groups
  1. we eliminate the possible biasing effect of individual characteristics of the students
  2. it provides a probabilistic basis for the selection of the sample that permits the statistician to calculate probabilities associated with the observations in the sample and to use these probabilities in making inferences.

Basic notions

Definition 3. *Experimental units* are the objects upon which measurements are taken.

The experimental units in this study are the individual students.

Definition 4. *Factors* are variables completely controlled by the experimenter. The intensity level (distinct subcategory) of a factor is called its *level*.

This experiment involves a single *factor*—namely, method of teaching. In this experiment, the factor has five *levels*: A, B, C, D, and E.

Definition 5. A *treatment* is a specific combination of factor levels.

- In a single-factor experiment like the preceding one, each level of the single factor represents a treatment. Thus, in our education example, there are five treatments, one corresponding to each of the teaching methods.

- As another example, consider an experiment conducted to investigate the effect of various amounts of nitrogen and phosphate on the yield of a variety of corn.

  – An experimental unit would be a specified surface—say, 1 ha—of corn.

  – A treatment would be a fixed number of kgs of nitrogen x_1 and of phosphate x_2 applied to a given ha of corn. For example, one treatment might be to use $x_1 = 100$ kgs of nitrogen per ha and $x_2 = 200$ kgs of phosphate. A second treatment might correspond to $x_1 = 150$ and $x_2 = 100$.

  – Notice that the experimenter could use different amounts (x_1, x_2) of nitrogen and phosphate and that each combination would represent a different treatment.

Definition 6. A *one-way layout* to compare k populations is an arrangement in which independent random samples are obtained from each of the populations of interest.

4.2 Completely randomized design

Completely randomized design

Definition 7. A *completely randomized design* to compare k treatments is one in which a group of n relatively homogeneous experimental units are randomly divided into k subgroups of sizes $n_1, n_2, \ldots, n_k$ (where $n_1 + n_2 + \cdots + n_k = n$). All experimental units in each subgroup receive the same treatment, with each treatment applied to exactly one subgroup.

- The preceding experiment for comparing teaching methods A, B, C, D, and E entailed randomly dividing the 125 students into five groups, each of size 25. Each group received exactly one of the treatments. This is an example of a completely randomized design.

- The observations obtained from a completely randomized design are typically viewed as being *independent random samples* taken from the populations corresponding to each of the treatments.

- Suppose that we wish to compare five brands of aspirin, A, B, C, D, and E, regarding the mean amount of active ingredient per tablet for each of the brands.

- We decide to select 100 tablets randomly from the production of each manufacturer and use the results to implement the comparison. In this case, we physically sampled five distinct populations.

- Regardless of whether we have implemented a completely randomized design or taken independent samples from each of several existing populations, a one-to-one correspondence is established between the populations and the treatments. Both of these scenarios, in which independent samples are taken from each of k populations, are examples of a one-way layout.

Definition 8. A *one-way layout* to compare k populations is an arrangement in which independent random samples are obtained from each of the populations of interest.

4.3 Randomized block design

Randomized block design

Definition 9. A *randomized block design* containing b blocks and k treatments consists of b blocks of k experimental units each. The treatments are randomly assigned to the units in each block, with each treatment appearing exactly once in every block.

- It is an extension of matched-pairs design

- Suppose that we wanted to compare three different medications for controlling hypertension.

- We could form several groups, each containing three members matched on sex, weight, and age.

- Within each group of three, we would randomly select one individual to receive treatment 1 and another to receive treatment 2, and then we would administer treatment 3 to the remaining member of each group.

- The objective of this design is identical to that of the matchedpairs design—namely, to eliminate unwanted sources of variability that might creep into the observations in our experiment.

Differences between CRD and RBD

- Experiment designed to compare subject reaction to a set of four stimuli (treatments) in a stimulus–response psychological experiment. We will denote the treatments as T_1, T_2, T_3, and T_4.

- Suppose that eight subjects are to be randomly assigned to each of the four treatments.

- Random assignment of subjects to treatments (or vice versa) randomly distributes errors due to person-to-person variability in response to the four treatments and yields four samples that, for all practical purposes, are random and independent.
- This is a *completely randomized experimental design*.
- The experimental error associated with a completely randomized design has a number of components. Some of these are due to the differences between subjects, to the failure of repeated measurements within a subject to be identical (due to the variations in physical and psychological conditions), to the failure of the experimenter to administer a given stimulus with exactly the same intensity in repeated measurements, and to errors of measurement.
- Reduction of any of these causes of error will increase the information in the experiment.
- The subject-to-subject variation in the foregoing experiment can be eliminated by using subjects as blocks.
- Each subject would receive each of the four treatments assigned in a random sequence.
- The resulting randomized block design would appear as in Figure 1. Now only eight subjects are needed to obtain eight response measurements per treatment.
- Notice that each treatment occurs exactly once in each block.
- The word *randomized* in the name of the design implies that the treatments are randomly assigned within a block.
- The purpose of the randomization (that is, position in the block) is to eliminate bias caused by fatigue or learning.
- Blocks may represent time, location, or experimental material.
- Other examples
  - A comparison of the sale of competitive products in supermarkets should be made within supermarkets, thus using the supermarkets as blocks and removing store-to-store variability.
  - Animal experiments in agriculture and medicine often use animal litters as blocks, applying all the treatments, one each, to animals within a litter. Because of heredity, animals within a litter are more homogeneous than those between litters. This type of blocking removes litter-to-litter variation.

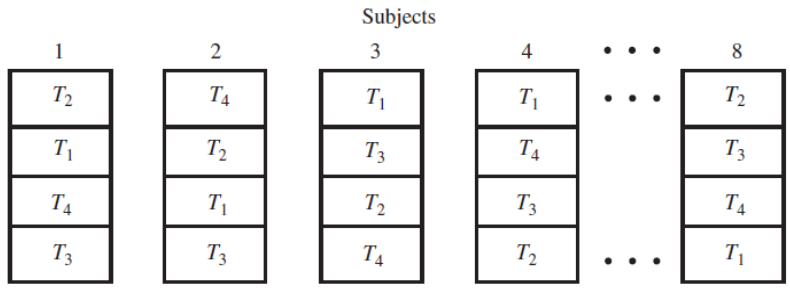

Figure 1: A randomized block design

Latin square design

- Blocking in two directions can be accomplished by using a *Latin square design*.
- Suppose that the subjects of the preceding example became fatigued as the stimuli were applied, so the last stimulus always produced a lower response than the first.
- Each stimulus is applied once to each subject and occurs exactly once in each position of the order of presentation. All four stimuli occur in each row and in each column of the 4×4 configuration. The resulting design is a 4×4 Latin square.
- A Latin square design for three treatments requires a 3×3 configuration; in general, p treatments require a $p \times p$ array of experimental units.

5 References

References

References

[1] Box, G. E. P., W. G. Hunter, and J. S. Hunter. 2005. *Statistics for Experimenters*, 2d ed. New York: Wiley Interscience.

[2] Cochran, W. G., and G. Cox. 1992. *Experimental Designs*, 2d ed. New York: Wiley.

[3] Graybill, F. 2000. *Theory and Application of the Linear Model*. Belmont Calif.: Duxbury.

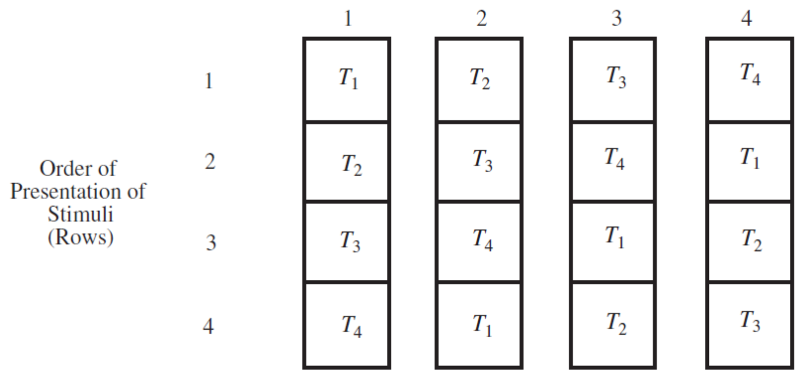

Figure 2: A Latin square design

[4] Hicks, C. R., and K. V. Turner. 1999. *Fundamental Concepts in the Design of Experiments*, 5th ed. New York: Oxford University Press.

[5] Hocking, R. R. 2003. *Methods and Applications of Linear Models: Regression and the Analysis of Variance*, 5th ed. New York: Wiley Interscience.

[6] Montgomery, D. C. 2006. Design and Analysis of Experiments, 6th ed. New York: Wiley.

[7] Scheaffer, R. L., W. Mendenhall, and L. Ott. 2006. *Elementary Survey Sampling*, 6th ed. Belmont Calif.: Duxbury.

[8] Scheffé, H. 2005. *The Analysis of Variance*. New York: Wiley Interscience.