經濟與管理論叢(Journal of Economics and Management), 2013, Vol. 9, No. 2, 271-290

Economic Field Experiments: Comments on Design Efficiency, Sample Size and Statistical Power

Johannes Ledolter*

Department of Management Science and Department of Statistics and Actuarial Science, University of Iowa, U.S.A.

This paper puts forward suggestions that could improve the efficiency of field experiments as they are currently carried out in experimental economics. Two recommendations are made: (1) Prior to the actual study, economic field experiments should include sample size calculations that confirm that meaningful effects can be detected. (2) Economic field experiments should take advantage of the power of multi-factor experimental plans.

Keywords: economic field experiments, experimental design, multi-factor experiments, sample size, statistical power

JEL classification: C9, C90, C93

1 Introduction

(1) In manufacturing, the value of well-designed and carefully-executed experiments has been well established. The outcomes of these experiments show which of several studied factors are important and how these factors relate to the response variables of interest. Box, Hunter and Hunter (2005) explain how to construct experiments efficiently and how to analyze the resulting data, and they illustrate their discussion with numerous case studies. (2) In medical research, investigators

*Correspondence to: Johannes Ledolter, Department of Management Science and Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242. Email: johannes-ledolter@uiowa.edu. I thank John List and participants at the University of Chicago brown bag seminar series on field experiments for stimulating discussions, and Arthur Swersey of Yale University for his comments on an earlier draft of this paper.

run experiments all the time, and evidence-based medicine relies on randomized experiments to confirm which of several treatments are the most effective. (3) Experiments are not only conducted in the natural and the health sciences, but also in the social sciences and in economics. An extensive literature on economic field experiments is summarized on John List's website http://www.fieldexperiments.com/. This website contains publications and discussion papers in experimental economics that make use of field experiments. The listed papers illustrate that well-designed experiments can help solve many open questions in economics. (4) More and more experiments address managerial business issues, and these experiments are not just run in laboratory settings, but in the real world (that is, "in the field"). The Six Sigma business management strategy, with its heavy focus on process improvement and experimentation, has contributed to an increase in the number of field experiments. Field experiments from business and marketing are discussed in Ledolter and Swersey (2007).

Questions about the most effective ways to design experiments and issues of sample size and statistical power are commonplace in scientific experimentation, in evidence-based medicine, and in economic field experiments. If experiments are executed poorly, little or even nothing will be learned from the resulting data. While it is true that most experiments increase knowledge (you usually learn "something" through experimentation), the experimenter wants to learn as efficiently as possible. Relatively few experimental runs (observations) are needed in efficient experimental designs to get precise estimates of the factor effects. Sir Ronald Fisher, the eminent statistician and scientist who developed this area, said that a well-designed experiment may improve the precision of the results tenfold, for the same cost in time and labor (R.A. Fisher (1935), page 217).

A thorough knowledge of experimental design principles can improve the efficiency of economic field experiments. Important principles of experimental design are replication, randomization, blocking, multi-factor instead of one factor at-a-time experimentation, and the sequential approach to experimentation. Each of these principles is discussed in Ledolter and Swersey (2007). The sequential approach to experimentation is important with the results of initial experiments used to determine the next experimental steps. Only a portion of the overall budget should be spent on the initial runs.

It may be difficult at times to convince engineers to run on-line experiments and to persuade firms to experiment on actual processes. The fear is that experiments on the production line and "in the field" (as compared to in the lab) may "mess up" the status quo and reduce throughput. Practitioners must be convinced that knowledge gained from experimentation is worth the risk. W. Edwards Deming (1982) maintained that an important reason why Japanese products (at that time, mostly cars and electronics) were better than US products was the Japanese emphasis on continuous quality improvement and experimentation. By experimenting, Japanese firms learned how to produce better products more efficiently.

Prior to running an experiment one needs to determine the sample size required to identify meaningful effects, that is, determine whether a certain sample size is sufficient to detect a specified change in the response. If the sample size is too small, first, observed effects may not be statistically significant and second, meaningful effects may not be uncovered. If the experimenter implements process changes based on observations that are large but not statistically significant, as Deming (1982) in his book *Out of the Crisis* has pointed out, such tampering with a stable system leads to an increase in variability.

It is very important to know prior to running the experiment whether the resulting data have a chance of detecting meaningful changes. For example, consider field experiments that study the effects of monetary incentives on the academic success of disadvantaged high school students. These studies are expensive: Students are paid for their efforts, and there are additional administrative costs. In such studies one must calculate the statistical power of detecting (practically) meaningful changes before implementing the experiment, in order to avoid the costs of inconclusive results.

An assessment of the literature on economic field experiments indicates that economic field experiments have several weaknesses.

(1) Rarely do field experiments address the required sample sizes, and the resulting studies are often underpowered statistically. As a consequence, often there is not enough information at the conclusion of the study that allows researchers to determine whether A is really better than B. I recommend that prior to running field experiments one carries out sample size calculations that ensure that practically

meaningful effects can be detected. If one cannot afford the required sample sizes, one should restructure or abandon the problem in favor of problems that can be solved. If there is little chance that meaningful changes can be detected, the money could be better spent elsewhere. I recommend that each funding proposal for an economic field experiment include a section on sample size and power, similar to the proposals for medical/drug studies that are being evaluated for funding by NIH.

(2) Frequently economic field experiments limit themselves to a simple comparison between two treatments, A and B. Usually just one factor is being studied. In this paper I illustrate that several factors can be studied with the same effort and costs, allowing for a more complete assessment of main and interaction effects. Economic field experiments should take advantage of the power of multi-factor experiments.

These two issues – sample size and statistical power, and the advantages of multi-factor experiments – are addressed in this paper. Section 2 expands on an earlier paper by List, Sadoff and Wagner (2011) and illustrates how to obtain appropriate sample sizes. Section 3 demonstrates that much can be gained by adopting multi-factor experiments. The discussion in this review is not entirely new as there is an extensive literature on the statistical design of experiments starting with the work of R. A. Fisher in the 1920s. The objective is to suggest improvements that could benefit researchers running field experiments.

2 Issues of Sample Size and Power: Determining the Appropriate Sample Size

In Sections 2.2, 2.3, 2.5 and 2.6 we add rigorous proofs for the sample size results given in the paper by List *et al*. (2011), and in Section 2.4 we add results that deal with sample sizes when comparing the means of two log-normal distributions. We also direct the investigator to flexible sample size/power software which makes sample size selection for detecting practically meaningful effects easy and straightforward. Having software readily available leaves no excuse for not addressing sample sizes prior to the start of an experiment.

2.1 Sample Size and Power when Testing a Single Mean

We consider the test of $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$, and test the research hypothesis of a reduction in the mean. Four quantities need to be specified: the standard deviation of an individual measurement Y, $\sigma = \sqrt{Var(Y)}$, the significance level α (usually $\alpha = 0.05$), and the power (usually 0.80) to detect a specified difference of interest $\delta = \mu_1 - \mu_0 < 0$. Note that $\beta = 1 - power = 0.2$ is the probability of a type II error.

We obtain the sample size n by solving two equations as follows.
From:

$$\alpha = P[reject\ H_0 \mid H_0\ true] = P[\bar{Y} \le c] = P[\frac{\bar{Y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \le \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}}]$$

$$= P[Z \le \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}}],$$

we express the known 100α percentile of the standard normal distribution as:

$$z_\alpha = \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

From:

$$Power = P[reject\ H_0 \mid H_1\ true] = P[reject\ H_0 \mid \mu_1 = \mu_0 + \delta]$$

$$= P[\bar{Y} \le c] = P[\frac{\bar{Y} - \mu_1}{\frac{\sigma}{\sqrt{n}}} \le \frac{c - \mu_1}{\frac{\sigma}{\sqrt{n}}}] = P[Z \le \frac{c - \mu_1}{\frac{\sigma}{\sqrt{n}}}],$$

we express the known $100(1-\beta)$ percentile of the standard normal distribution as:

$$z_{Power} = z_{1-\beta} = \frac{c - \mu_1}{\frac{\sigma}{\sqrt{n}}}.$$

We solve the two equations $z_\alpha = (c - \mu_0) \big/ (\sigma \big/ \sqrt{n})$ and $z_{Power} = z_{1-\beta}$

$= (c - \mu_1)/(\sigma/\sqrt{n})$ for the two unknown quantities n and c, obtaining the following result for the required sample size.

Result 1: $n = [z_\alpha - z_{1-\beta}]^2 / [(\mu_1 - \mu_0)/\sigma]^2 = (z_\alpha + z_\beta)^2 (\delta/\sigma)^2$.

Example 1: $\sigma = \sqrt{Var(Y)} = 1$, targeted detectable difference $\delta = \mu_1 - \mu_0 = -0.3$, $\alpha = 0.05$ and $z_\alpha = -1.645$, $\beta = 0.20$ (power = 0.80) and $z_{0.20} = -0.8416$.

Then:

$$n = \left[z_\alpha + z_\beta\right]^2 \left[\frac{\sigma}{\delta}\right]^2 = \frac{(-1.645 - 0.8416)^2}{(0.09)} = (100/9)(2.4816)^2 = 68.4 .$$

The required sample size is 69.

Comment 1: Here we have worked with the normal distribution, assuming implicitly that the sample size is fairly large. Normality for averages follows from the central limit effect. If the sample size n (for given σ and δ) is small, one can use the t-distribution and solve the equation $n = [t_{n-1,\alpha} + t_{n-1,\beta}]^2 (\sigma/\delta)^2$, where ($t_{n-1,\alpha}$, $t_{n-1,\beta}$) are percentiles of the t-distribution. This equation has to be solved iteratively. Excellent computer software is available to carry out these calculations, and we discuss useful packages in Section 2.6.

Comment 2:

- The sample size increases with the power. The more power one wants, the larger the sample size.
- The sample size increases with decreasing detectable difference. The smaller the difference one wants to detect, the larger the sample size must be.
- The sample size increases proportionally to the variance. The larger the uncertainty, the larger the sample size. The sample size quadruples with a doubling of the standard deviation.
- Two-sided tests require a larger sample size than one-sided tests.

Comment 3: This result can be applied to the paired (blocked, or within-subject) test with response $D = Y - X$. Here the same experimental unit is observed under

both treatments; for example, before (X) and after (Y) a certain treatment is applied. This type of design is common in drug studies where the effectiveness of a drug is assessed by comparing the response of a subject under treatment to his/her baseline response without treatment. This type of design is different from the fully randomized arrangement where the control and experimental groups are made up of different individuals; the data analysis that corresponds to the fully randomized design is given in the next section. As another illustration, imagine a study that compares the durabilities of two types of shoe soles. The fully randomized experiment assigns different subjects to the two experimental groups with a subject receiving either one or the other treatment, while a blocked (or within-subject) experiment assigns each of the two soles to the same individual (this is easy to do as there are two feet). The blocked experiment is more efficient as it allows us to isolate and remove the subject effect from the comparisons. In the paired experiment, $\sigma = \sqrt{Var(D)} = \sqrt{Var(Y) + Var(X) - 2\,\mathrm{cov}(Y, X)} < \sqrt{Var(Y) + Var(X)}$ if X and Y are positively correlated and if blocking has been effective. Blocking on factors that have a large influence on the results is very useful; this is what Fisher had in mind when he said that a complete overhaul of an experimental design may improve the precision of the results ten- or twelve-fold, for the same cost in time and labor.

2.2 Sample Size and Power when Comparing Means of Two Populations

We compare the means of two groups and test $H_0 : \mu_2 - \mu_1 = 0$ against $H_1 : \mu_2 - \mu_1 < 0$. We test the research hypothesis of a reduction. The relevant function of the data to test the above hypothesis is given by $(\bar{Y}_2 - \bar{Y}_1) \Big/ \sqrt{(\sigma_2^2 / n_2) + (\sigma_1^2 / n_1)}$.

For this test we need the following (now five) quantities: the two standard deviations σ_1 and σ_2 which don't have to be equal; the significance level α (usually $\alpha = 0.05$); and the power (usually 0.80) to detect a specified difference of interest $\delta = \mu_2 - \mu_1 < 0$. Note that $\beta = 1 - power = 0.2$ is the probability of a type II error.

We obtain the two sample sizes n_1 and n_2 by solving two equations as follows:

From:

$$\alpha = P[reject\ H_0 \mid H_0\ true] = P[\bar{Y}_2 - \bar{Y}_1 \le c] = P[\frac{\bar{Y}_2 - \bar{Y}_1 - (0)}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}} \le \frac{c - (0)}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}]$$

$$= P[Z \le \frac{c}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}],$$

we obtain the first equation:

$$z_\alpha = \frac{c}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}.$$

From:

$$Power = P[reject\ H_0 \mid H_1\ true] = P[reject\ H_0 \mid \mu_2 - \mu_1 = \delta < 0]$$

$$= P[\bar{Y}_2 - \bar{Y}_1 \le c] = P[\frac{\bar{Y}_2 - \bar{Y}_1 - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}} \le \frac{c - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}]$$

$$= P[Z \le \frac{c - \delta}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}],$$

we obtain the second equation:

$$z_{Power} = z_{1-\beta} = \frac{c - \delta}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}.$$

We solve the two equations:

$$z_\alpha = \frac{c}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}} \quad \text{and} \quad z_{Power} = z_{1-\beta} = \frac{c - \delta}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}},$$

for the three unknowns, c, n_1 and n_2. This leads to the equation:

$$z_\alpha\sqrt{\frac{\sigma_2^2}{n_2}+\frac{\sigma_1^2}{n_1}}=\delta+z_{1-\beta}\sqrt{\frac{\sigma_2^2}{n_2}+\frac{\sigma_1^2}{n_1}}\,.$$

An infinite number of combinations of n_1 and n_2 satisfies this equality as there are three unknowns and only two equations. Among these many combinations, we pick the combination of n_1 and n_2 that minimizes the total sample size $N=n_1+n_2$.

Result 2: The sample sizes of the two groups should be selected proportional to the standard deviations; that is $n_1/n_2=\sigma_1/\sigma_2$. Under this optimal allocation the total sample size $N=n_1+n_2$ is given by:

$$N=\left[\frac{z_\alpha+z_\beta}{\delta}\right]^2\left[\sigma_2+\sigma_1\right]^2.$$

Proof: Let $\sigma_2^2/n_2+\sigma_1^2/n_1=g$. Many combinations of n_1,n_2 will satisfy this equation. We want the combination such that $N=n_1+n_2$ is a minimum. Express $n_1=n_2\sigma_1^2/(gn_2-\sigma_2^2)$ and $N=n_2+[n_2\sigma_1^2/(gn_2-\sigma_2^2)]$. Take the derivative of $N=n_2+[n_2\sigma_1^2/(gn_2-\sigma_2^2)]$ with respect to n_2 and set it equal to 0. This gives the equation $1=\sigma_1^2\sigma_2^2/[gn_2-\sigma_2^2]^2$ or $1=\pm\sigma_1\sigma_2/(gn_2-\sigma_2^2)$. This leads to $gn_2=\sigma_2(\sigma_1+\sigma_2)$. Similarly (by setting equal to zero the derivative of $N=n_1+[n_1\sigma_2^2/(gn_1-\sigma_1^2)]$ with respect to n_1) we obtain $1=\pm\sigma_1\sigma_2/(gn_1-\sigma_1^2)$ and $gn_1=\sigma_1(\sigma_1+\sigma_2)$. This leads to $n_1/n_2=\sigma_1/\sigma_2$. Note that the solutions with $-\sigma_1\sigma_2$ in the numerators of the above two equations (that result from setting the first derivatives equal to zero) violate $n_1>0,n_2>0$. The result for the total sample size $N=n_1+n_2$ follows by substitution.

Comment 4: Here we have used the normal distribution assuming that the sample sizes are fairly large. If the combined sample size *N* is small, we can obtain a better value for *N* by using percentiles of the t-distribution, $t_{N-2,\alpha},t_{N-2,\beta}$. The equation $N=[t_{N-2,\alpha}+t_{N-2,\beta}]^2[(\sigma_2+\sigma_1)/\delta]^2$ must be solved iteratively.

Comment 5: Assuming (incorrectly) that $n_1=n_2=n$, and solving the equation $z_\alpha\sqrt{\sigma_2^2/n_2+\sigma_1^2/n_1}=\delta+z_{1-\beta}\sqrt{\sigma_2^2/n_2+\sigma_1^2/n_1}$ for $N=2n$, the total sample size is

$N = [(z_\alpha + z_\beta)/\delta]^2 2(\sigma_2^2 + \sigma_1^2)$.

This allocation increases the optimal total sample size by the factor of $2(\sigma_2^2 + \sigma_1^2)/(\sigma_2 + \sigma_1)^2$.

Result 3: Assuming that the standard deviations are the same ($\sigma_1 = \sigma_2 = \sigma$), the sample sizes for the two groups are the same, and the sample size for each group is $n = 2[(z_\alpha + z_\beta)/\delta]^2 \sigma^2$, for a combined sample size of $N = 2n = 4[(z_\alpha + z_\beta)/\delta]^2 \sigma^2$.

Comment 6:

- The sample sizes should be selected proportional to the standard deviations. Equal sample sizes should be selected if $\sigma_2 = \sigma_1$.
- The sample size increases with power. The more power you want, the larger the sample size.
- The sample size increases with decreasing detectable difference. The smaller the difference you want to detect, the larger the sample size.
- The sample size increases proportionally to the variances. The larger the uncertainty, the larger the sample size must be.
- Two-sided tests require a larger sample size than one-sided tests.
- Rule of thumb: With $\alpha = 0.025$ (or $\alpha = 0.05$ in a two-sided test) and $z_\alpha = -1.96$, and $\beta = 0.20$ and $z_{0.20} = -0.8416$, we have $2[z_\alpha + z_\beta]^2 \approx 16$. This leads to the following rule of thumb: Use (16)(16) = 256 subjects in each of two groups if you want to detect an effect that amounts to one quarter of a standard deviation (as then $\sigma/\delta = 4$).

Example 2: $\sigma_2 = 3$ and $\sigma_1 = 1$; $\alpha = 0.05$ and $z_\alpha = -1.645$; $\beta = 0.20$ (power = 0.80) and $z_{0.20} = -0.8416$, and detectable difference $\delta = \mu_2 - \mu_1 = -0.5$. Then:

$$N = \left[\frac{z_\alpha + z_\beta}{\delta}\right]^2 [\sigma_2 + \sigma_1]^2 = \left[\frac{-1.645 - 0.8416}{-0.5}\right]^2 (3+1)^2 = 394,$$

for a total sample size of about 400. We should put 300 subjects into the group with standard deviation $\sigma_2 = 3$, and 100 subjects into the group with standard deviation $\sigma_1 = 1$.

Putting the same number of subjects into both groups (different from the optimal allocation) leads to a total sample size of

$N = [(z_\alpha + z_\beta)/\delta]^2\, 2(\sigma_2^2 + \sigma_1^2) = [(-1.645 - 0.8416)/(-0.5)]^2\, 2(9+1) = 493$; and about 250 in each group. This allocation increases the sample size by 100(20/16) = 25 percent.

2.3 Sample Size and Power when Comparing Two Proportions

In many comparative studies we evaluate the success of a new strategy or a new method through the resulting change in a proportion. For example, we may have two different advertising strategies (strategies 1 and 2) and may be interested in whether or not strategy 2 increases the proportion of people who buy a certain product. Under the null hypothesis H_0: $\pi_1 = \pi_2 = \pi$, the distribution of the difference of the two sample proportions $p_2 - p_1$ is normal with mean 0 and variance $2\pi(1-\pi)/n$, where *n* is the size of the first (and second) sample, for a combined sample size 2*n*. For a test with significance level α , we reject the null hypothesis in favor of the one-sided alternative H_1: $\pi_2 - \pi_1 = \delta > 0$ whenever $p_2 - p_1 > z_{1-\alpha}\sqrt{2\pi(1-\pi)/n}$; $z_{1-\alpha}$ is the $100(1-\alpha)$ percentile of the standard normal distribution.

We are looking for a test with power $1-\beta$, which implies probability β of falsely accepting the null hypothesis if the alternative ($\pi_2 - \pi_1 = \delta > 0$) is actually true. This requirement implies the equality:

$$\frac{z_{1-\alpha}\sqrt{\frac{2\pi(1-\pi)}{n}} - \delta}{\sqrt{\frac{\pi(1-\pi)}{n} + \frac{(\pi+\delta)(1-\pi-\delta)}{n}}} = -z_{1-\beta}\,.$$

The above equation can be solved for the sample size *n*, leading to:

$$n = \frac{[z_{1-\alpha}\sqrt{2\pi(1-\pi)} + z_{1-\beta}\sqrt{\pi(1-\pi)+(\pi+\delta)(1-\pi-\delta)}\,]^2}{\delta^2}\,.$$

Setting $\delta = 0$ in the numerator, leads to the approximation:

$$n \cong \frac{2\pi(1-\pi)[z_{1-\alpha} + z_{1-\beta}]^2}{\delta^2} = \frac{2\pi(1-\pi)[z_\alpha + z_\beta]^2}{\delta^2}\,;$$

see Ledolter, J. and Swersey, A.: *Testing 1-2-3: Experimental Design with*

Applications in Marketing and Service Operations. Stanford University Press, 2007 (page 42).

Example 3: Consider the planning value $\pi = 0.03$ for the common success proportion, and assume that it is important to detect an increase of one half of a percent ($\delta = 0.005$). For $\alpha = \beta = 0.05$ and $z_{0.95} = 1.645$, we must sample:

$$n \cong \frac{2(0.03)(0.97)[1.645+1.645]^2}{(0.005)^2} = 25{,}200$$

subjects in each group, for a total of 50,400 people for the two groups combined.

Comment 7: The result is pertinent to the design of comparative experiments that attempt to estimate the difference between two unknown success proportions. It shows how to select the two sample sizes such that a certain specified difference (δ) in the success proportions is detected with reasonably large power. A planning value for the common success proportion (π) and a meaningful detectable difference (δ) of the two success proportions need to be specified. Information on the success rate is usually available from prior experiments, and worthwhile changes are determined with economic considerations in mind. In our illustration, taken from a marketing study with very small response rates, $\pi = 0.03$ and $\delta = 0.005$.

2.4 Sample Size and Power when Comparing Two Lognormal Distributions

In some situations the treatment affects the level as well as the variation, and many times the standard deviation of the response is proportional to the level. In such cases the logarithmic transformation of the response Y stabilizes the variability; see Box, Hunter and Hunter (2005), Abraham and Ledolter (2006, page 205). A normal distribution for the transformed response $X = \log Y$, with mean μ and standard deviation σ, implies that the response variable Y follows a lognormal distribution.

Usually we are given the coefficient of variation of the (untransformed) observations. Using results about the mean and variance of a lognormal distribution, the coefficient of variation is given by $c = \sqrt{Var(Y)}\,/\,E(Y) = \sqrt{\exp(\sigma^2)-1}$. We can

solve this equation for σ, the standard deviation of the log-transformed observations $X = \log Y$. It is $\sigma = \sqrt{\log(1+c^2)}$.

We are also given the (proportionate) effect in the level of the *Y*-observations that we want to detect. This means that:

$$E(Y_1) = E(Y_0)(1+f) \quad \text{or} \quad \exp(\mu_1 + 0.5\sigma^2) = (1+f)\exp(\mu_0 + 0.5\sigma^2).$$

Note that we have assumed that σ is the same in both groups and that the change is only in the means of the log-transformed observations $X = \log Y$. We assume that the coefficient of variation is the same under the null and the alternative hypothesis. This important assumption certainly needs to be checked on prior data. Under this assumption we can solve the above equation to obtain the difference in the means of log-transformed observations; it is:

$$\delta = \mu_1 - \mu_0 = \log(1+f).$$

Hence, for the power calculations, we transform the data to logs, $X = \log Y$, with $\sigma = \sqrt{\log(1+c^2)}$. We want to detect the difference $\delta = \mu_1 - \mu_0 = \log(1+f)$.

We apply Result 3 for the two-sample comparison. With one-sided significance α and power $1-\beta$, the required sample size for each group is:

$$n = 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2 \sigma^2 = 2\left[\frac{z_\alpha + z_\beta}{\log(1+f)}\right]^2 \log(1+c^2).$$

Comment 8: This equation is derived in Van Belle, G. and Martin, D.C.: "Sample size as a function of coefficient of variation and ratio of means," *The American Statistician*. Vol. 47 (1993), pages 165-167.

Comment 9: (Rule of thumb) With $\alpha = 0.025$ (or $\alpha = 0.05$ in a two-sided test) and $z_\alpha = -1.96$, and $\beta = 0.20$ and $z_{0.20} = -0.8416$, we have $2[z_\alpha + z_\beta]^2 \approx 16$. Furthermore, for small coefficient of variation c and small proportional change f, $\log(1+c^2) \approx c^2$ and $\log(1+f) \approx f$. Hence, for small c and f, $n \approx 16(c/f)^2$.

Example 4: Assume coefficient of variation $c = 0.15$. Then $\sigma = \sqrt{\log(1+(0.15)^2)} = 0.149$. Assume that we want to detect a 20 percent change in outcome ($f = 0.20$); then $\delta = \log(1+0.2) = 0.182$. Hence:

$$n = 2\left[\frac{z_\alpha + z_\beta}{\log(1+f)}\right]^2 \log(1+c^2) = 2\left[\frac{-1.645-0.8416}{0.182}\right]^2 (0.149)^2 = 8.29 .$$

That is, we need 9 subjects in each of the two groups.

2.5 Sample Size and Power for Cluster Designs: Your Sample Sizes May Need to Be Larger than You First Thought

Assume that we study two groups, with equal variances. Result 3 shows that the sample size for each of the two groups is $n = n_1 = n_2 = 2[(z_\alpha + z_\beta)/\delta]^2 \sigma^2$. The derivation assumes that the two treatments are assigned to the experimental units (subjects, objects, rats, etc) at random. It assumes a fully randomized arrangement in which the outcomes are independent across the experimental units.

Sometimes the randomization is applied to clusters that consist of groupings of the experimental units. Clusters may be communities, and experimental units may be people. The randomization is at the cluster level; that is, the treatment groups (experimental and control, such as absence and presence of a certain economic incentive) are assigned to clusters at random. Each of the *m* experimental units in a cluster is then assigned to the same treatment. While the data of interest comes from experimental units in the two experimental groups, the randomization is carried out on the clusters.

Usually subjects from the same cluster tend to be alike. Since observations from the same cluster are now correlated, with intra-cluster correlation coefficient $\rho > 0$ (that is, the correlation among units from the same cluster), the *m* observations in a cluster don't carry the same weight as *m* independent observations. Hence, in the presence of large intra-cluster correlation it is important to randomize over many, and preferably small clusters so as to maximize the efficiency of the experiment. Taking more and more replicates within a rather small number of clusters may get you larger sample sizes, but not the desired power.

Here is the theoretical justification of this result. The variability of an experimental unit is the sum of two variances, $\sigma^2 = \sigma_C^2 + \sigma_\varepsilon^2$: a cluster variance σ_C^2 and a unit-specific variance σ_ε^2. The intra-cluster correlation coefficient is $\rho = \sigma_C^2/(\sigma_C^2 + \sigma_\varepsilon^2)$. Assume that each cluster contains *m* experimental units. Cluster

averages have variance $\sigma_C^2 + (\sigma_\varepsilon^2/m)$. The required number of clusters in each treatment group (for specified significance level, and for specified power at given detectable difference δ) is:

$$k = 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2\left[\sigma_C^2 + \frac{\sigma_\varepsilon^2}{m}\right].$$

Hence the required number of observations n (number of clusters, k, times number of observations in the cluster, m) in each treatment group is:

$$\begin{aligned} n &= 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2\left[\sigma_\varepsilon^2 + m\sigma_C^2\right] = 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2\sigma^2\left[\frac{\sigma_\varepsilon^2 + m\sigma_C^2}{\sigma_C^2 + \sigma_\varepsilon^2}\right] \\ &= 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2\sigma^2\left[1 + (m-1)\frac{\sigma_C^2}{\sigma_C^2 + \sigma_\varepsilon^2}\right] \\ &= 2\left[\frac{z_\alpha + z_\beta}{\delta}\right]^2\sigma^2\left[1 + (m-1)\rho\right]. \end{aligned}$$

The intra-cluster correlation inflates the sample size that we obtain under complete random sampling, $2[(z_\alpha + z_\beta)/\delta]^2\sigma^2$, by the factor $[1 + (m-1)\rho]$. For $\rho = 0$, we are back at our earlier result. For $\rho = 1$, we multiply the sample size that we obtain under complete random sampling by the number of experimental units in the cluster (m). Each experimental unit in a cluster is a carbon-copy of the other units in that cluster. The m experimental units in the cluster basically count as one unit (and not as m); it is the number of clusters that matters, and not the number of observations within the cluster. Hence, in the presence of large intra-cluster correlation, it is important to randomize over many, small clusters so as to maximize the efficiency of the experiment. Taking more and more replicates within the cluster doesn't increase the power of the experiment. Taking more clusters does.

2.6 Sample Size and Power Calculations Using Statistical Software

Here we have described the sample size calculations from first principles. However excellent statistical software is available to carry out the sample size calculations. Most statistics packages have capabilities for calculating the appropriate sample

sizes. There are also programs dedicated to sample size exclusively such as Russ Lenth's sample size/power applets, http://www.stat.uiowa.edu/~rlenth/Power/. Lenth's sample size applets (they are free, good, and easy to use) cover many situations, such as numeric outcome variables (emphasis on means and variances), categorical outcome variables (emphasis on proportions), slope coefficients in regression models, and factorial experiments. Commercial statistical computer software such as Minitab and JMP also include useful programs for sample size calculations.

3 The Advantage of Multi-Factor Designs: Sample Size and Power in 2-Level Factorial Experiments

Consider a factorial experiment with two factors at two levels each and suppose that the same number of independent observations is taken at all factor-level combinations. The response may be weekly store sales of a product, and the factors may be the price of the product (low and high price) and the presence of a store display (no display and display at the store entrance). The grocery chain administering 40 stores in a metropolitan area may assign each of the four factor-level combinations (that is, low price and no display, high price and no display, low price and display, and high price and display) to ten of their stores at random. For a fair comparison, the sales response at a store would need to be adjusted for overall store sales.

The 2^2 factorial experiment with k independent replicates at each of the four factor-level combinations requires a total sample size of $4k$ observations. Two main effects can be estimated. The main effect of each factor compares the average of the $2k$ observations at the low level of the factor with the average of the $2k$ observations at the high level; see Ledolter and Swersey (2007). For example, the main effect of price compares the average sales from the 20 stores where the price was set high with the average sales from the 20 stores where the price was set low. The main effect of display compares the average sales from the 20 stores with a store display with the average sales from the 20 stores without a display.

In Section 2.2 we compared two levels of a single factor and we determined – for given significance, power, and detectable difference – the required total sample

size for the experiment, $N = 2n = 4[(z_\alpha + z_\beta)/\delta]^2 \sigma^2$; see Result 3. A 2^2 factorial experiment with $k = N/4$ replicates requires the same overall sample size, but has important advantages over the one-factor experiment: the factorial experiment provides estimates of not only one but two main effects, and it does so with the same precision as the one-factor experiment. Furthermore, the factorial experiment allows for the estimation of an interaction which expresses how the effect of one factor changes with the level of the other. One may want to know whether the effect of price depends on the presence of a display; price may matter less if people don't know about the product.

Factorial experiments are very useful. With the same number of observations we can study the effects of two factors, and we can do so with the same precision that is achieved by a one-factor experiment. However, the advantage of the factorial experiment diminishes when there is a strong interaction between the two factors, as in such a case we should not estimate an "overall" effect of one factor by averaging the observations over the second factor. A main effect (that is, an effect that is not "conditioned" on the other factor) has no meaning if the effect of one factor changes with the level of the other factor. While the effect of one factor at a given level of the second factor can always be estimated from the factorial experiment, this comparison involves two averages of $k = N/4$ observations and carries less statistical power than a single factor comparison that puts $N/2$ observations into each group.

Multi-factor experiments are more efficient in terms of their sample size than several one-factor experiments pieced together. Consider the un-replicated 2^3 factorial experiment; it contrasts two averages of four observations each when estimating each of three main effects. In a sequence of several one-factor experiments pieced together we would need 8 observations for estimating the main effect of the first factor (contrasting 4 observations at one level with 4 observations at the other; here the second and third factors are fixed at levels that are deemed to be "best"); 4 more observations for estimating the main effect of the second factor (we already have 4 observations from the previous comparison that had fixed the second factor at one level, but we need 4 more observations that set the second factor at its second level), and 4 more for estimating the main effect of the third factor. A sequence of three one-factor experiments requires 16 observations to

estimate the main effects with the very same precision that is achieved by the 2^3 factorial. This amounts to a doubling of the effective sample size, and the experiment still does not tell us anything about the presence/absence of interactions. Ledolter and Swersey (2007; page 98) show that a sequence of one-factor experiments for p factors requires more observations than a 2^p factorial experiment, increasing the number of observations by a factor of $(1+(p-1)/2)$; for $p=3$, this factor is 2.

In field experiments one usually tests two levels of a single factor (the targeted factor) and keeps fixed several other important factors. From economic theory, one often knows that the fixed factors do have an impact. Their impact on the response can be through just their main effects or through their interactions with the targeted factor. If interactions between the targeted and the fixed factors are absent, the approach of fixing the other factors is wasteful as a multi-factor experiment with the same number of observations can estimate all main effects, and not just the one that has been targeted for study. Furthermore, a multi-factor experiment reveals interactions.

If the response effect of a targeted factor at certain fixed levels of some other factors is the only interest, then the analysis should be "powered" for just this particular comparison. But one should keep in mind that one will not learn about the effects of factors that have been fixed and any interactions among the targeted and fixed factors. Running such a narrow experiment is inefficient. It is preferable to cast the "design net" wider so that one can learn about the effects of all factors (that is, also the factors that have been fixed). The factorial experiment is preferable if one can assume that the interactions between the factor of interest and the factors that have been fixed are negligible (and it makes no difference whether or not there are interactions among the factors that have been fixed). Using the multi-factor factorial design in this situation, we achieve the same power for the comparison on the targeted factor, plus we get an additional opportunity to learn about the effects of the factors that had been fixed.

The benefits of the multi-factor plan are greatest if there are no interactions between the factor of interest and the factors that have been fixed. The full benefits of the multi-factor plan are not realized and some of its advantages are lost if there are interactions among the targeted and the fixed factors, as in this case it is not

meaningful to talk about an unconditional main effect and collapse the analysis over the factors that have been fixed. The results of the factorial experiment can always be used to compare the mean responses at the two levels of the targeted factor at specified levels of the other factors, and in the worst case (presence of interactions among the targeted factor and all fixed factors) this comparison contrasts k observations at one level with k observations at the second level of the targeted factor. But for each (formerly) fixed factor that is free of interaction with the targeted factor, we can get a better estimate of the effect of the targeted factor by collapsing the results over the (formerly) fixed factor and comparing the averages of $2k$ observations at each of the two levels of the targeted variable. If two (formerly) fixed factors have no interaction with the targeted variable, we get a comparison of $4k$ observations at each of the two levels of the targeted factor, and so on. Of course, at the outset of a study one usually does not know whether interactions are present and it is difficult to make general statements about the expected benefits of multi-factor experiments. This brings us back to our earlier comment about the sequential nature of experimentation and the importance of leaving resources for follow-up studies. Sir Ronald Fisher once said that "the best time to design an experiment is after you've done it" (George Box (1993), who furthermore writes: "One manifestation of that seeming paradox is that after a preliminary experimental design has been run, questions are often raised about the results with an acuity of hindsight which is quite extraordinary").

References

Abraham, B. and J. Ledolter, (2006), *Introduction to Regression Modeling*, Belmont, CA: Duxbury Press.

Box, G. E. P., (1993), "Sequential Experimentation and Sequential Assembly of Designs," *Quality Engineering*, 5, 321-330.

Box, G. E. P., J. S. Hunter, and W. G. Hunter, (2005), *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd edition, New York: Wiley.

Deming, W. E., (1982), *Out of the Crisis*, Cambridge, MA: MIT Press.

Fisher, R. A., (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

Fisher, R. A., (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

Ledolter, J. and A. J. Swersey, (2007), *Testing 1 – 2 – 3: Experimental Design with Applications in Marketing and Service Operations*, Stanford University Press.

Lenth, R. V., (2006-9), Java Applets for Power and Sample Size [Computer Software], from http://www.stat.uiowa.edu/~rlenth/Power, Accessed 2013 May 31.

List, J., http://www.fieldexperiments.com/, Accessed 2013 May 31, A Useful Website Containing (Economic) Field Experiments.

List, J., S. Sadoff, and M. Wagner, (2011), "So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design," *Experimental Economics*, 14, 439-457.

Sidak, Z., (1967), "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions," *Journal of the American Statistical Association*, 62, 626-633.

Van Belle, G. and D. C. Martin, (1993), "Sample Size as a Function of Coefficient of Variation and Ratio of Means," *The American Statistician*, 47, 165-167.