S4. Sampling strategy

Sample size calculation

To determine how many households to be surveyed, a minimum required number (sample size) is calculated so that we can study meaningful associations between waste disposal practices and some households' characteristics (denoted as T in this document) such as income levels, awareness of rules and regulations, and their perception about waste issues.

Three key elements for sample size determination are (1) the significance level (denoted as a), (2) the power (denoted as $1-b$), and (3) minimum detectable effect size (denoted as d) of testing whether the association exists or not.

Outcome variables (disposal practices) are binomial random variables taking value of one if a household follows certain disposal practice (e.g. recycling), or zero otherwise. Suppose mean outcome of households with type zero ($T=0$) is estimated as p_0, and that of type one household ($T=1$) as p_1.

Assuming the variances of outcomes are the same for both household types (i.e. $\sigma_0^2 = \sigma_1^2 = \sigma^2$), the required minimum sample size (N^*) is

$$N^* = \frac{4\left(z_{a/2} + z_{1-b}\right)^2 \sigma^2}{d^2}. \tag{1}$$

where $z_{a/2}$ is the percentage point of a normal distribution of two-sided hypothesis test with significance level at a, z_{1-b} is the percentage point with statistical power of $1-b$, and d is difference in the outcomes $p_1 - p_0$. (Step by step derivation of equation (1) is included at the end of this document.)

We set the statistical significance level at 0.05 ($z_{0.025} = 1.96$), statistical power at 80% ($z_{0.2} = 0.84$) [1], and minimum effect size at one standard deviation, $d = \sigma$. Resulting required total sample size (N^*) is

$$N^* = 4(1.96 + 0.84)^2 \cong 124.$$

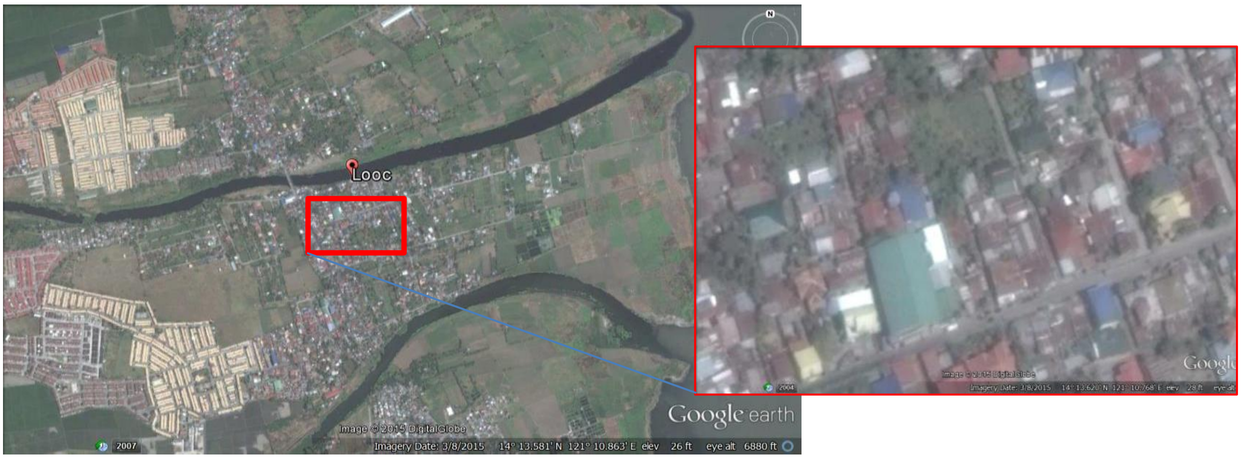
Selection of respondents

Sample households are selected randomly in a systematic manner. Since complete list of

[1] This is equivalent of saying that a possible false positive finding (Type II error of wrongly rejecting null hypothesis $H_0: p_0 = p_1$) is 20%.

all households in the Barangay was not obtainable at the time of survey, following steps were used. (i) We obtained a map of the barangay from Google Earth and assign serial numbers to the houses on the map (as in sample in Figure S4-1). (ii) Equidistant houses are marked with an appropriate interval to meet the required sample size. (iii) The coordinates of the marked houses were transferred to a GPS device for a site visit. (iv) If the chosen household at the marked location is not available for survey on the day of visit, the next household on the map confirm the willingness of the household to participate in the survey was approached.

Figure S4-1 Sampling map and numbering houses


Looc
Google earth
Image © 2015 DigitalGlobe
Imagery Date: 3/8/2015 14° 13.581' N 121° 10.863' E elev 26 ft eye alt 6880 ft
2007
Google

Barangay Looc divides its jurisdiction into 15 “areas” to which barangay health workers (BHWs) are assigned and the waste collection schedule is based. 15 BHWs were asked to become the enumerators of the assigned area because of their familiarity with the area. Focus group discussions and enumerator training session were conducted to refine questionnaires and ensure that the BHWs will be able to elicit adequate information. Each BHW interviewed eight households to yield total of 120 respondents. Each survey form is checked by Baltazar for the consistency before coding.

Step by step derivation of equation (1)

Let outcome Y_{iT} be binomial random variables conditional on T. Using the normal approximation of the binomial distribution, it is

$$Y_{iT}|X_i{\sim}N(p_T,\sigma_T^2)$$

where X_i represents observable individual characteristics, p_T and σ_T^2 are mean and variance of outcome of type T .

Assuming we can model individual outcome as a function of the observable variables

X_i, unobserved person specific characteristic α_i, average effect $\bar{\tau}$ of type *T*, and a person specific effect τ_i of type *T*, we write

$$Y_{iT} = \alpha_i + X_i\beta + \bar{\tau}T + \tau_i T + \varepsilon_i.$$

Given that *T* is assigned to individual randomly and independent of any of the unobserved variables, the correlation between outcomes $Y_{iT}$and types T can be estimated by taking differences of between the observed averages, i.e.

$$\hat{\tau} = E(Y_{i1}) - E(Y_{i0}) = p_1 - p_0.$$

We define the detectable effect size d and a standardized effect, adopting List et.al (2011), as

$$d = E(Y_{i1}) - E(Y_{i0}), \text{ and}$$

$$\text{stadardized effect} = \frac{d}{\sqrt{\frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}}}$$

where n_T is the numbers of observations of type *T*.

The null hypothesis is that there is no difference between types, i.e.

$$H_0: d = 0.$$

In order to reject the null hypothesis at the significance level of 0.05 with a two-sided test ($\frac{a}{2}$ =0.025), the observations must satisfy

$$\frac{p_1 - p_0}{\sqrt{\frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}}} \geq z_{\frac{a}{2}}. \qquad (2)$$

On the other hand, a statistical power of 80% (i.e. $b = 0.2$) with detectable effect size *d*, requires margin of false positive to be less than 20%, i.e.

$$\frac{p_1 - p_0 - d}{\sqrt{\frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}}} \leq -z_b \qquad (3)$$

$$\Rightarrow \frac{-(p_1 - p_0) + d}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} \geq z_b.$$

Combining (2) and (3) yields

$$z_{\frac{a}{2}} + z_b \leq \frac{d}{\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}} \Rightarrow \sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} \leq \frac{d}{z_{\frac{a}{2}} + z_b}.$$

Thus optimal sampling strategy is to choose n_0, n_1 that satisfy

$$\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} = \frac{d}{z_{\frac{a}{2}} + z_b}. \tag{4}$$

(i) When the variances of outcomes are the same for both household types, i.e., $\sigma_0^2 = \sigma_1^2 = \sigma^2, a = 0.05, and\ b = 0.2$, the minimum required sample size of each type (n^*) becomes.

$$\frac{2\sigma^2}{n^*} = \frac{d^2}{(z_{0.025}+z_{0.2})^2}.$$

$$n^* = n_0 = n_1 = \frac{2(z_{0.025}+z_{0.2})^2\sigma^2}{d^2}.$$

Substituting $z_{0.025}, z_{0.2}$ with 1.96 and 0.84,

$$n^* = n_0 = n_1 = 2(1.96 + 0.84)^2 \frac{\sigma^2}{d^2}.$$

If we set the minimum detectable standardized effect size as $\frac{\sigma^2}{d^2}$ =1, i.e. one standard deviation, the required total sample size is

$$N^* = 2 \times n^* = 2 \times 2(1.96 + 0.84)^2 \cong 32.$$

If we set the minimum detectable standardized effect size as 0.5, i.e. half the standard deviation, the required total sample size becomes

$$N^* = 2 \times n^* = 2 \times 2(1.96 + 0.84)^2 \left(\frac{1}{0.5}\right)^2 \cong 125.$$

(ii) Assuming variances are not necessarily identical between the types, i.e., $\sigma_0^2 \neq \sigma_1^2$, we now demonstrate the sample size obtained assuming identical variance in (i) is sufficient.

The equality (4) can be rewritten as

$$\sqrt{\frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}} = \sqrt{\frac{\frac{\sigma_0^2}{n_0+n_1}}{\frac{n_0}{n_0+n_1}}+\frac{\frac{\sigma_1^2}{n_0+n_1}}{\frac{n_1}{n_0+n_1}}} = \sqrt{\frac{1}{n_0+n_1}\left(\frac{\sigma_0^2}{\frac{n_0}{n_0+n_1}}+\frac{\sigma_1^2}{\frac{n_1}{n_0+n_1}}\right)} = \frac{d}{z_{\frac{a}{2}}+z_b}.$$

$$N^* = n_0 + n_1 = \left(\frac{z_{\frac{a}{2}}+z_b}{d}\right)^2 \left(\frac{\sigma_0^2}{\frac{n_0}{n_0+n_1}}+\frac{\sigma_1^2}{\frac{n_1}{n_0+n_1}}\right)$$

To find the minimum sample size N*. let $\frac{n_0}{n_0+n_1} = G$. The optimal G solves

$$\min_G \frac{\sigma_0^2}{G}+\frac{\sigma_1^2}{1-G} = \frac{(1-G)\sigma_0^2 + G\sigma_1^2}{G(1-G)}.$$

The optimal proportion $G^* = \frac{\sigma_0}{\sigma_0 \ +\sigma_1}$ minimizes total sample size, that is

$$N^* = \left(\frac{z_{\frac{a}{2}}+z_b}{d}\right)^2 \left(\frac{\sigma_0^2(\sigma_0^2+\sigma_1^2)}{\sigma_1^2}+\frac{\sigma_1^2(\sigma_0^2+\sigma_1^2)}{\sigma_0^2}\right) = \left(\frac{z_{\frac{a}{2}}+z_b}{d}\right)^2 (\sigma_0^2+\sigma_1^2). \quad (5)$$

We now show that balanced sample size assuming equal variance will be greater than the minimum sample size.

By letting $n^* = n_0 = n_1, N^* = n_0 + n_1 = \left(\frac{z_{\frac{a}{2}}+z_b}{d}\right)^2 2(\sigma_1^2+\sigma_0^2)$, while by letting $\frac{n_0}{n_0+n_1} = \frac{\sigma_0}{\sigma_1 \ +\sigma_0}$, $N^{**} = n_0 + n_1 = \left(\frac{z_{\frac{a}{2}}+z_b}{d}\right)^2 \left(\sigma_1 \ +\sigma_0 \ \right)^2$.

Since $2(\sigma_1^2+\sigma_0^2) - \left(\sigma_1 \ +\sigma_0 \ \right)^2 = \sigma_1^2+\sigma_0^2 - 2\sigma_0\sigma_1 = \left(\sigma_1 \ -\sigma_0 \ \right)^2 > 0$,

$$N^* > N^{**}.$$

Therefore the minimum sample size assuming equal variance is the sufficient minimum total sample size.

References:

List, John A., Sally Sadoff, and Mathis Wagner, 2011, “So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design”, *Experimental Economics*, Vol.14, pp.439-457.