Dissertation Workshop: Exercise 3 Solutions
Johns Hopkins Bloomberg School of Public Health OpenCourseWare

Determination of Sample Size - Simple Random Sample

Purpose of Analysis	Source of Error	Type of Error	General Formula for N	Special Case
Estimate Universal Mean	I	I	$(ZS/D)^2$	$4(S^2/D^2)$
Decide whether Universal Mean Conforms to Defined Standard	I	2	$((Z_1 + Z_2)S/D)^2$	$10.9(S^2/D^2)$
Estimate Differences between Two Universal Means	2	I	$2((ZS)/D)^2$	$8(S^2/D^2)$
Decide Whether Real (non-zero) Differences Exist between Two Universal Means	2	2	$2((Z_1 + Z_2)S/D)^2$	$21.8(S^2/D^2)$

N = Cases; S = Estimate of standard deviation of the sample; D = Distance from mean to one side of the range; Z = 2.0 (95% confidence); Z_1 = 2.0 (5% risk of Type I Error); Z_2 = 1.3 (10% risk of Type II Error)

In the cases below asking for population estimates, these should be obtained with 95% assurance that the true situation is within the specified range. Where the problem involves hypothesis-testing, a significance level of 5% and power of 80% are needed.

We want to estimate the age of the mother at her first marriage in a population to within three months with 95% confidence. An estimated standard deviation of the population of women at first marriage is 2 years.

What formula would you use? **$N = (ZS/D)^2$**

What is the sample size? **$N = (ZS/D)^2 = 4(S^2/D^2) = 4(2/.25)^2 = 256$**

What if we want to estimate to within 1/2 year? **$N = 4(2/.5)^2 = 64$**

What is the formula and the sample size in each sample needed if we want to determine the impact of a program encouraging women to delay their first marriage for one year when compared with the group above? **$N = 2((Z_1 + Z_2)S/D)^2 = 21.8(2/1)^2 = 87.2$**

What is the formula and sample size if we want to estimate the difference in the length of time breast-feeding in two dissimilar countries. In country A, women breast-feed for only 5 months but in country B women breast-feed for 8 months. The approximate standard deviation of the time spent breast-feeding is 2 months. **$N= 2((ZS)/D)^2 = 8 (3/2)^2 = 18$**

A certain district has three million population, 15 percent of whom are children under the age of five. The number of household members averages five and 70 percent of the dwellings have at least one pre-school child. Only one-fourth of the children are from homes that have utilized health center services within the past twelve months. Those households can be identified from records kept in the 150 health centers serving the district.

Because diarrhea is common among the children, a video on the benefits of oral rehydration has been prepared for presentation in selected villages served by community health workers (CHW). The villages vary in size and together have one million people. The remaining two million population is made up of villages that have neither CHWs nor videos available.

After the CHS/video intervention has been in place for six months, a sample survey is to determine whether the use of oral rehydration therapy (ORT) in the intervention area is significantly more common than in the remaining (control) area. The survey among pre-school children is to identify cases of diarrhea occurring within the preceding two weeks and for these cases will ascertain whether ORT was received. It is estimated that the two-week prevalence of diarrhea among pre-school children is about 20 percent, and in the absence of the contemplated intervention about 30 percent of these children receive ORT. An intervention effect of 5 percent is considered meaningful. Thus, if utilization in the intervention area is truly 5 percent higher than in the control area, the survey should produce statistically significant findings with 90 percent assurance. In case the intervention is ineffective, the survey should give 95 percent assurance of producing results that are declared non-significant.

If a stratified survey design are employed to assess separately health center users and non-users in the intervention and control areas, how large a simple random sample of diarrhea cases would be needed in each of the four groups? Although the analysis is to be based on children who had diarrhea within the past two weeks, the unit of observation for sample selection is the household. How many households would be needed to generate the requisite number of cases of diarrhea? What if two cases of diarrhea are encountered in a single household?

As a practical matter, a multistage sampling design is to be used in place of simple random sampling. Specifically, 5 villages are to be selected in each of 10 health center areas in the intervention region and similar numbers in the control area. A constant number of cases is to be identified in each village chosen. Assuming that the multistage design is expected to increase sampling variance by 50 percent (design effect 1.5), by how much would the sample size have to be increased? How many cases would be needed in each village chosen for study? Give the mathematical expression that defines the probability that a child with the following characteristics will be chosen: has diarrhea; family has not used the local health center that serves 28,000 population; lives in an intervention village of 2,000 population?

Review the preceding considerations and calculations in terms of their feasibility. In the interests of realism, make any modifications to the survey design and decision criteria that you think are appropriate.

	Intervention	**Control**	**Total**
Population	1,000,000	2,000,000	3,000,000
Underfives	150,000	300,000	450,000
Households	200,000	400,000	600,000
Households under five	140,000	280,000	420,000
Households with users	50,000	100,000	150,000
Health Centers	50	100	150

The correct answer is:

Simple Random Sample

$N= 2(Z_1 + Z_2)^2 pq /E^2 = 2(2+1.3)^2 (.3)(.7)/(.05)^2 = 1,830$

Need 1,830 cases in 9,150 households with underfives in each area (intervention and control)

9,150/0.7 = 13,072 households = 9,150 with underfives = 1,830 cases

If 2 cases per household, randomly select one.

Multistage Sampling

N = 1.5(1,830) = 2,745 per group

2,745/50 = 55 per village

Probability

(1/5) = Has diarrhea
(3/4) = Non User
(10 * ((28,000/5).7)/((1,000,000/5).7)) = HC area chosen
(5 * ((2,000/5).7)/((28,000/5).7) = Village Chosen
(55 / ((2,000/5).7) = Child Chosen

$$P = \left[\frac{1}{5}\right]\left[\frac{3}{4}\right]\left[10\left(\frac{\frac{28,000}{5}x.7}{\frac{1,000,000}{5}x.7}\right)\right]\left[5\left(\frac{\frac{2,000}{5}x.7}{\frac{28,000}{5}x.7}\right)\right]\left[\frac{55}{\frac{2,000}{5}x.7}\right]$$

$$P = \left(\frac{1}{5}\right)\left(\frac{3}{4}\right)\left(\frac{10x5x55}{140,000}\right) = \left(\frac{3}{20}\right)\left(\frac{2,750}{140,000}\right) = .00295$$

Feasibility

Very large sample needed overall; stratification by use status may not be practical; select small samples with high power only for detecting somewhat larger effects.

References

Henderson R, Sundaresan T. 1992. Cluster Sampling to Assess Immunization Coverage: A review of Experience with a Simplified Sampling Method. *Bulletin World Health Organization* 60: 253-260.

Lemeshow S, Stroh G. 1988. Sampling Techniques for Evaluating Health Parameters in Developing Countries. Board on Science and Technology for International Development (BOSTID) working paper. Washington, DC: National Academy Press.

Reinke W. 1991. Applicability of Industrial Sampling Techniques to Epidemiologic Investigations: Examination of an Underutilized Resource. *American Journal of Epidemiology 134* (10): 1222-1232.