Why is 1,000 enough?
When is 1,000 not enough?

Technote

Small samples

It seems strange that a sample of 1,000 for Canada has about the same theoretical accuracy as a sample of 1,000 for Toronto or Red Deer when there is such a difference in the populations. It seems more reasonable that a percentage of the population—say 5%—should be used for all situations.

Fortunately, statistical theory and practice show this is not necessary. For Canada, a sample of 5 percent would imply the need to survey over 1 million people—a very expensive proposition.

Sampling experiment

Imagine a pail containing index cards inscribed with numbers between 0 and 120. Assume there are cards for everyone in Canada (about 30,000,000) and that the numbers represent ages.

Now, select any card at random and note the age. Then, select another respondent at random, calculate the age, and then find the average of the two. Repeat this random selection, and compute the cumulative average age of the group selected. Select without replacement, and set these cards aside.

We might visualize the experiment proceeding as shown in the table below.

TABLE 1

Selection	Age	Cummulative Age	Average Age
1st	23	23	23 (23/1)
2nd	34	57	28.5 (57/2)
3rd	75	132	44 (132/3)
4th	48	180	45 (180/4)
5th	18	198	39.6 (198/5)
6th	27	225	37.5 (225/6)
:	:	:	:
10,000th			42.3
:	:	:	:
30,000,000	:	:	Final population average

Initial estimates of the average age are inaccurate, but as more observations are added, the average converges toward a final Canadian average, as shown in the figure below.

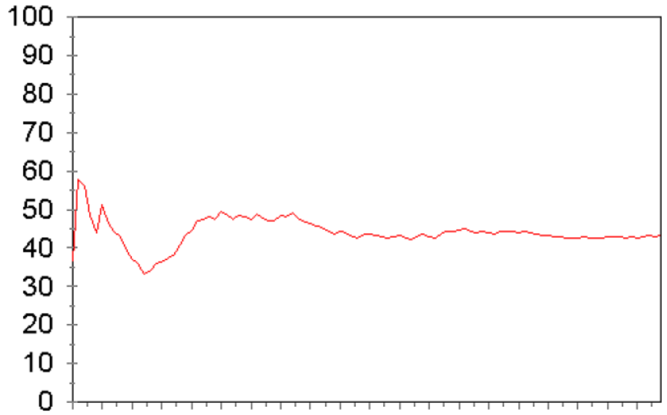

FIGURE 1

Initial estimates of the average fluctuate, but the series eventually converges to some average that is close to the population average. The figure above shows one experiment using random selection from the sample. Another experiment might show a different path to the same final population average. The key point is that the estimate of the average converges at a fairly small sample size.

The approach to the average is not smooth, because the random selection process may draw a number of respondents in a row whose ages are much higher or lower than the final average. These act to temporarily drag the current estimate of the average away from the final value. However, these deviations become increasingly minor as the sample increases. As the figure shows, with about 400 in the sample, they are minor.

What is the correct sample size?

What is a satisfactory estimate of the average age? In the figure depicted above, the estimate stabilizes after about 300 respondents; by 600, the average fluctuates narrowly around the final value. It is reasonable to estimate the average age of the population by using a random selection procedure which draws only 300 to 400 respondents. This shows that a relatively small sample of 300-400 is quite accurate when only a single value (point) of the population is needed, regardless of its size.

The more extreme the values in a population, the higher the sample needed before there is convergence.

This experiment is convincing for samples which only pursue single items (age, income, feelings about disarmament, etc.), but most researchers are interested in the relation between variables. When the data are broken down into these bivariate and multivariate relationships, 1,000 may not be enough sample.

Consider the table below, which presents location by age category in terms of the number of respondents.

TABLE 2

Age by location			
Age	**Urban**	**Non-urban**	**Total**
18-34	211	102	313
35-50	180	108	288
50+	148	136	284
Total	539	346	885

If there were 12 categories of age instead of the three categories as shown in the table above, it is easy to visualize some of the cells becoming sparse. In this situation, some common statistical tests (chi-square, for example), may provide erroneous indicators of association and independence.

As the data are grouped and "cross-classified" the sample size must increase to ensure each category has sufficient information to support inference. For example, the 346 Non-Urban respondents in the table above, should not be spread across 12 categories of age.

Survey biases

In addition to the problems produced by multivariate analysis, surveys are also subject to a number of biases. In the experiment shown in the first table which calculated average age on the basis of the sample, the underlying assumption was that the sample frame (list from which the sample is drawn) represents the population. If the sample frame misrepresents the population, then random sampling will not produce an accurate measure of the population parameters, such as the average. For example, a random selection of driver licence records is unrepresentative of the general population.

Do not assume that all of the cells in a crosstabulation must have a minimum size. It is more important that each variable has a minimum sample size. For example, if one is investigating gender difference and political preference, a table such as the one shown below might appear.

TABLE 3

Voter preferences			
	Male	**Female**	
Red	a	b	R1
Blue	c	d	R2
White	e	f	R3
C1 C2			g n

The actual number in any individual cell such as "a" or "f" is not important. It is essential that the total "N" be large enough to prevent several cells from having small entries. A sparse table (30% of the cells with five or fewer entries) will invalidate many statistical tests.

It is also essential to ensure sufficient "N" to compare the pattern of responses of men and women. If the difference among those voting for the "REDS" and those voting for the "BLUES" is important, then R1, R2, and R3 must have sufficient numbers.

The variation in the population is important to determining the required sample size. In large scale projects, it is wise to undertake pilot sampling to understand the degree of variability in the key variables. This may allow one to reduce the sample size in the larger study and thereby reduce cost, or will ensure a sufficient sample is drawn to answer the relevant questions.

Some general guidelines

It is frustrating to try to answer questions with an insufficient sample, since money has been spent without result.

The following table presents general guidelines for determining sample size. These rules should not replace a pilot study on large projects or a formal power analysis with population data.

Minimum sample sizes (general guidelines)

	Sample N
Single Variable (Income, age, attitude)	300
2 Way Table (less than 4 categories/variables)	750
3 Way Table (less than 4 categories/variables)	1,000
3 Way Table (5+ categories)	1,800
Sample with 5 regions, gender, 3 income groups and 3 age groups	5,000

For additional information, please contact admin@pra.ca