PENNSTATE
Cooperative Extension

Program Evaluation

Tipsheet #60

How to Determine a Sample Size

I want to survey a large group of people. What size should my sample be? Twenty percent? Thirty percent?

AVOID There is no set percentage that is accurate for every population. What matters is the actual number or size of the sample, not the percentage of the population. Consider a coin toss: the first few times you flip the coin, the average result may be skewed wildly in one direction (say you got 5 tails in a row), but the more times you flip the coin, the more likely that the average result will be an even split between heads and tails.

So, if you surveyed 20% of a group of 300 program participants to produce a sample of 60 people, you would **under represent** the population, since there is a fairly large chance in a small group that the respondents you choose will vary from the whole population. On the other hand, 20% of 30,000 county residents (a sample of 6,000) would be a wastefully large sample, and not significantly more accurate than a sample of 400.

USE There are 5 steps in deciding a sample size. If you are familiar with them, select a sample size using the tables in Appendices 1 and 2 at the end of the Tipsheet. To use a formula, move to Appendix 3. If you wish to review the five steps before selecting a sample size however, read on.

Steps in Selecting a Sample-Size

An appropriate sample size is based on a number of accuracy factors that you must consider. Together they comprise a five step process:

1. Determine **Goals**
2. Determine desired **Precision** of results
3. Determine **Confidence** level
4. Estimate the degree of **Variability**
5. Estimate the **Response Rate**

Step One: Determine Goals

- First, know the size of the population with which you're dealing. If your population is small (200 people or less), it may be preferable to do a **census** of everyone in the population, rather than a **sample**. For a marginally higher cost than a 134-person sample, you can survey the entire population and gain a 0% sampling error. However, if the population from which you want to gather information is larger, it makes sense to do a sample.
- Second, decide the methods and design of the sample you're going to draw and the specific **attributes or concepts** you're trying to measure.
- Third, know what kind of resources you have available, as they could be a limitation on other steps below such as your level of precision. Once you have this information in-hand, you're ready to go on to the next step.

Step Two: Determine the Desired Precision of Results

The **level of precision** is the closeness with which the sample predicts where the true values in the population lie. The difference between the sample and the real population is called the **sampling error**. If the sampling error is ±3%, this means we add or subtract 3 percentage points from the value in the survey to find out the actual value in the population. For example, if the value in a survey says that 65% of farmers use a particular pesticide, and the sampling error is ±3%, we know that in the real-world population, between 62% and 68% are likely to use this pesticide. This range is also commonly referred to as the **margin of error**.

The level of precision you accept depends on balancing accuracy and resources. High levels of precision require larger sample sizes and higher costs to achieve those samples, but high margins of error can leave you with results that aren't a whole lot more meaningful than human estimation.

The tables in Appendices 1 and 2 at the end of the Tipsheet provide sample sizes for **precision levels** of 5% and 3% respectively.

Step Three: Determine the Confidence Level

The confidence level involves the **risk** you're willing to accept that your sample is within the average or "bell curve" of the population. A confidence level of 90% means that, were the population sampled 100 times in the same manner, 90 of these samples would have the true population value within the **range of precision** specified earlier, and 10 would be unrepresentative samples. Higher confidence levels require larger sample sizes.

The tables at the end of this Tipsheet assume a 95% confidence level. This level is standard for most social-science applications, though higher levels can be used. If the confidence level that is chosen is too low, results will be "statistically insignificant".

Step Four: Estimate the Degree of Variability

Variability is the degree to which the attributes or concepts being measured in the questions are distributed throughout the population. A heterogeneous population, divided more or less 50%-50% on an attribute or a concept, will be harder to measure precisely than a homogeneous population, divided say 80%-20%. Therefore, the higher the degree of variability you expect the distribution of a concept to be in your target audience, the larger the sample size must be to obtain the same level of precision.

To come up with an estimate of variability, simply take a reasonable guess of the size of the smaller attribute or concept you're trying to measure, rounding **up** if necessary. If you estimate that 25% of the population in your county farms organically and 75% does not, then your variability would be .25 (which rounds up to 30% on the table provided at the end of this Tipsheet). If variability is too difficult to estimate, it is best to use the conservative figure of 50%.

Note: when the population is *extremely* heterogeneous (i.e., greater than 90-10), a larger sample may be needed for an accurate result, because the population with the minority attribute is so low.

At this point, using the **level of precision** and **estimate of variability** you've selected, you can use either the table or the equation provided at the bottom of this Tipsheet to determine the **base sample size** for your project.

Step Five: Estimate the Response Rate

The base sample size is the number of responses you must get back when you conduct your survey. However, since not everyone will respond, you will need to increase your sample size, and perhaps the number of contacts you attempt to account for these non-responses. To estimate response rate that you are likely to get, you should take into consideration the method of your survey and the population involved. Direct contact and multiple contacts increase response, as does a population which is interested in the issues, involved, or connected to the institution doing the surveying, or, limited or specialized in character. You can also look at the rates of response that may have occurred in similar, previous surveys.

When you've come up with an estimate of the percentage you expect to respond, then divide the base sample size by the percentage of response. For example, if you estimated a response rate of 70% and had a base sample size of 220, then your final sample size would be 315 (220/0.7).

Once you have this, you're ready to begin your sampling!

One final note about response rates: the past thirty years of research have demonstrated that the characteristics of non-respondents may differ *significantly* from those of respondents. Follow-up samples may need to be taken of the *non*-respondent population to determine what differences, if any, may exist.

Appendix 1 Example: 5% Error and Qualification.

Appendix 2 Example: 3% Error and Qualification.

Appendix 3 Example: An Equation for Determining Final Sample Size.

References:

Blalock, Hubert M. (1972). Social Statistics. New York: McGraw-Hill Book Company.

Israel, Glen D. 1992. "Determining Sample Size." Program Evaluation and Organizational Development, IFAS, University of Florida. PEOD-6.

National Science Foundation, Research and Development in Industry: 1992, NSF 95-324. Arlington, VA.

Smith, M.F. 1983. "Sampling Considerations in evaluating Cooperative Extension Programs." Cooperative Extension Service, IFAS, University of Florida. DRAFT.

Taylor-Powell, Ellen. May 1998. "Sampling." Program Development and Evaluation, University of Wisconsin Extension. G3658-3.

Sudman, Seymour (1976). Applied Sampling. New York: Academic Press.

Warmbrod, J. Robert (1965). "The Sampling Problem in Research Design." Agriculture Education Magazine. pp 106-107, 114-115.

Yamane, Taro (1973). "Statistics: an introductory analysis." New York: Harper & Row.

Jeff Watson, Research Assistant, Cooperative Extension & Outreach
The reference citation for this Tipsheet is: Watson, Jeff (2001). How to Determine a Sample Size: Tipsheet #60, University Park, PA: Penn State Cooperative Extension.
Available at: http://www.extension.psu.edu/evaluation/pdf/TS60.pdf

Appendix 1: Tables[a] for Finding a Base Sample Size[b]

+/- 5% Margin of Error [c]

Sample Size

	Variability				
Population	**50%**	40%	30%	20%	10% [d]
100 [e]	**81**	79	63	50	37
125	**96**	93	72	56	40
150	**110**	107	80	60	42
175	**122**	119	87	64	44
200	**134**	130	93	67	45
225	**144**	140	98	70	46
250	**154**	149	102	72	47
275	**163**	158	106	74	48
300	**172**	165	109	76	49
325	**180**	173	113	77	50
350	**187**	180	115	79	50
375	**194**	186	118	80	51
400	**201**	192	120	81	51
425	**207**	197	122	82	51
450	**212**	203	124	83	52
500	**222**	212	128	84	52
600	**240**	228	134	87	53
700	**255**	242	138	88	54
800	**267**	252	142	90	54
900	**277**	262	144	91	55
1,000	**286**	269	147	92	55
2,000	**333**	311	158	96	57
3,000	**353**	328	163	98	57
4,000	**364**	338	165	99	58
5,000	**370**	343	166	99	58
6,000	**375**	347	167	100	58
7,000	**378**	350	168	100	58
8,000	**381**	353	168	100	58
9,000	**383**	354	169	100	58
10,000	**385**	356	169	100	58
15,000	**390**	360	170	101	58
20,000	**392**	362	171	101	58
25,000	**394**	363	171	101	58
50,000	**397**	366	172	101	58
100,000	**398**	367	172	101	58

Qualifications

a) This table **assumes a 95% confidence level,** identifying a risk of 1 in 20 that actual error is larger than the margin of error (greater than 5%).

b) Base sample size should be **increased** to take into consideration potential non-response.

c) A **five percent margin of error** indicates willingness to accept an estimate within +/- 5 of the given value.

d) When the estimated population with the smaller attribute or concept is less than 10 percent, the sample may need to be increased.

e) The assumption of normal population is poor for 5% precision levels when the population is 100 or less. The entire population should be sampled, or a lesser precision accepted.

Appendix 2: Tables[a] for Finding a Base Sample Size[b]

+/- 3% Margin of Error [c]

Sample Size

	Variability				
Population	**50%**	40%	30%	20%	10% [d]
2,000 [e]	**714**	677	619	509	322
3,000	**811**	764	690	556	341
4,000	**870**	816	732	583	350
5,000	**909**	850	760	601	357
6,000	**938**	875	780	613	361
7,000	**959**	892	795	622	364
8,000	**976**	908	806	629	367
9,000	**989**	920	815	635	368
10,000	**1000**	929	823	639	370
15,000	**1034**	959	846	653	375
20,000	**1053**	975	858	660	377
25,000	**1064**	984	865	665	378
50,000	**1087**	1004	881	674	381
100,000	**1099**	1014	888	678	383

Qualifications

a) This table **assumes a 95% confidence level,** identifying a risk of 1 in 20 that actual error is larger than the margin of error (greater than 3%).
b) Base sample size should be **increased** to take into consideration potential non-response.
c) A **three percent margin of error** indicates willingness to accept an estimate within +/- 3 of the given value.
d) When the estimated population with the smaller attribute or concept is less than 10 percent, the sample may need to be increased.
e) The assumption of normal population is poor for 3% precision levels when the population is 2,000 or less. The entire population should be sampled, or a lesser precision accepted.

Appendix 3: An Equation for Determining Final Sample Size

$$n = \frac{\left(\dfrac{P[1-P]}{\dfrac{A^2}{Z^2} + \dfrac{P[1-P]}{N}} \right)}{R}$$

Where:

n = sample size required
N = number of people in the population
P = estimated variance in population, as a decimal: (0.5 for 50-50, 0.3 for 70-30)
A = Precision desired, expressed as a decimal (i.e., 0.03, 0.05, 0.1 for 3%, 5%, 10%)
Z = Based on confidence level: <u>1.96</u> for 95% confidence, <u>1.6449</u> for 90% and <u>2.5758</u> for 99%
R = Estimated Response rate, as a decimal