Determining the best sample size for studying your informal (science) education program

Rebecca Cors. May 2022

What's important to consider when selecting a sample size, or participant group size, for studying your informal (science) education program? What value does a small sample size have? If program managers have ample resources for collecting a lot of data, what factors are important to consider?

Here is a collection of guidelines and examples for use in making an informed decision about determining the sample size for studying your informal learning program. To design a study whose findings are more authentic and useful, make the sample group as representative as possible of the population that you want to know about. Then, look at what museum researchers and social scientists recommend.

If you have a question, or want to share an additional tip that I could add to this guide, please do not hesitate to contact me at rcors@wisc.edu .

Contents

Sampling for authenticity 2
Sampling sizes recommended by museum researchers 4
Sampling sizes that support generalizing study results 5
Appendix: About Purposive Sampling 6
Appendix: Conway's Rough Guide to Sample Size Calculation 8
References 9

Sampling for authenticity

Carrying out a study that will produce results that are most useful for making improvements to your program or exhibit depends in large part of how relevant results are to all program participants. Before even selecting a sample size, paying attention to the composition of your sample is important, along with the reliability and validity of your data collection approaches. Using an unrepresentative sample will give you data and results that support biased conclusions. This bias cannot be eliminated by taking a larger sample.

Are small samples useless?

Is it worthwhile to collect feedback about a museum exhibit from just five visitors? To informally ask teachers what they think about how their class visit to your science camp seems to be going for youth participants? The answers are yes and definitely.

For the former, even five visitors can give you clues about what works and what could be improved for your exhibit. Of course, keep in mind what they brought with them, besides their keen interest in your museum. What education or literacy level do these five people have? Do they come from the gender groups, age groups, and communities (zip codes) that represent all of the people who visit your museum (or other out of school learning place)? If they don't, could you arrange the sixth interview or survey with someone who balances out the group?

Either way, be sure that the description of the results of your inquiry include a detailed description of who responded and how the sample (respondent group) compares with the population (everyone who participates in the program, visits that exhibit). Asking about education and demographic factors can sometime feel awkward or take too much time. Get creative. If asking for survey takers' reading level feels strange or off-putting, what about asking about how many books they have read in the last six months?

For the latter, informal conversations with teachers are golden opportunities to get insights about the effectiveness of your program. Such informal conversations are considered bona-fide parts of research studies, called "informal conversations with key informants," so long as you have identified yourself to the interviewees and so long as you have identified your study focus. Not only can teachers can provide great insights into their students' experiences with your program, they can also offer ideas about sampling strategies to learn more (should it be a survey, an interview, a response board?). If they offer you advice, find ways to return the favor with for their classroom: SWAG, a virtual or in-person visit to their classroom, or a link to the most current content.

Sample composition

It is worth noting that in museum studies, the sample is often who shows up that day, or who attends the program. For example, if a school in a predominantly wealthy suburb has a day off, perhaps more of them will attend the museum. An example has to do with field trip programs, which can often only include the participants in the program. A teacher who decides to participate in a museum field trip is likely to be more enthusiastic and perhaps have pupils who are more curious. This kind of purposive sample is a non-probability sample that is selected based on characteristics of a population and the objective of the study.

These kinds of differences are worth considering during study design. At a minimum, it is important to describe your sample composition in any reporting. Ideally your study would include a comparison group (other class group from a different zip code visiting the same museum exhibit). If no comparison class is available, try to identify a similar class that could be a proxy comparison class group.

For more information, contact Rebecca Cors rcors@wisc.edu

Reliability and Validity

To collect relevant, useful data, social scientists design data collection instruments support reliability and validity. Below some definitions for reliability and variability from Scribbr (2022), followed by strategies for improving them.

	Reliability	**Validity**
What does it tell you?	The extent to which the results can be reproduced when the research is repeated under the same conditions.	The extent to which the results really measure what they are supposed to measure.
How is it assessed?	By checking the consistency of results across time, across different observers, and across parts of the test itself.	By checking how well the results correspond to established theories and other measures of the same concept.
How do they relate?	A reliable measurement is not always valid: the results might be reproducible, but they're not necessarily correct.	A valid measurement is generally reliable: if a test produces accurate results, they should be reproducible.

Reliability

An example strategy for examining the reliability of an observation instrument is to have a second person collect observations, which allows one to calculate an interrater reliability statistic. Another strategy is to administer the same survey to the same people on two different occasions and calculating the correlation between the two scores obtained. High test-retest correlations indicate a more reliable survey item.

Improving Validity

Do you mean to ask youth about self-efficacy, but are they instead responding about their self-confidence? Maybe you understand that self-confidence is a term that refers to the strength of belief, whereas self-efficacy refers to the confidence in one's ability to deal with a situation without being overwhelmed. But do your survey takers?

Strategies for improving valid data collection:

Instrument development	– Co-develop with program manager and teachers. – For example, work with stakeholders such as middle school teachers and librarians, to make sure language of surveys and interview guides is appropriate for these students.
Test instruments	– Test survey instruments and interview guides with people similar to the intended study participants. – NOTE: each time I develop a survey, I wonder if it is worth the time to test it. The testers *always* find important improvements!
Triangulate data collection	– Triangulate means to collect data about a factor of interest in several different ways. – For example, collect data from students through an online survey and an end-of-program post-it response board, and also collect data from their teacher during a two minute interview.

For more information, contact Rebecca Cors rcors@wisc.edu

Sampling sizes recommended by museum researchers

The Visitor Research National Network (Kelly, 2014) refers to research to describe appropriate sample sizes for different phases of a study and emphasizes that, given that budgets and programs limit sampling, "any slice of the pie is better than none." Below are a few rules of thumb for sample sizes, based on suggestions from various experts.

Sampling size suggestions from the Visitor Research National Network (Kelly, 2014). More detail here.

Exploratory evaluation	– 5-10 subjects – 15-20 in focus groups – 40-60 in quantitative analysis
Formative/ front end and summative	– 100 onsite surveys – 150 on-line surveys
Main study	– 96 visitor surveys for a museum with 1 million visitors (10 sampling error) – Up to 40 qualitative surveys (open-ended items) – 300-600 surveys for a small museum – 600-700 survey respondents for a national museum

For more information, contact Rebecca Cors rcors@wisc.edu

Sampling sizes that support generalizing study results

If your goal is to carry out research that can contribute to a body of knowledge, such as how people learn about science, you will want to collect enough data to make your results generalizable to a population. The extent to which results of a study are generalizable depend on how confidently research findings and conclusions based on data collected from your sample (eg, the 50 museum visitor who will complete the iPad survey tomorrow) can be used to describe the larger population you want to know about (museum visitors). The larger the sample population, the smaller the margin of error, and the more confidence one can have in generalizing the results. Of course, a representative sample composition also adds to how confidently one can speak about how generalizable study results are.

Below is an example of how to calculate a sample size, based on statisticians' rules of thumb and an explanation about why a larger sample reduces the margin for error, making study results more generalizable.

Calculating sample size

Conway (2018) and others explain how, when a study compares subgroups, you should check that the sample size will have enough power to give you an acceptable margin of error and account for variability within the smallest subgroup of interest.

Two rules of thumb for size of group	– For some statistical tests, some say the sample size should be at least **10** per subgroup especially for ANOVA (comparing averages) http://www.real-statistics.com/one-way-analysis-of-variance-anova/assumptions-anova/ – Another well-accepted sampling per group, especially if you want to conduct a rigorous study, comes from Tabachnick & Fidell (2013), who say a minimal sample size of **30** is necessary for every group.
Example sample calculation	If the goal is to compare males with females, and also responses from the morning with responses in the afternoon, there are four different subgroups (shown in the figure below). If you take the recommendation to have a minimum of 30 subjects per subgroup, the total minimum sample, after eliminating outliers, etc., is 120.

	morning	afternoon
female	30	30
male	30	30

TOTAL SAMPLE = 120

Why are larger samples better?

Ronan Conway's Rough Guide to Sample Size (2018) explains how larger sample sizes offer more precision. They show several examples of how a sample size of less than 100 will result in at least 10% margin of error. The larger the sample, the smaller the margin of uncertainty (confidence interval, or confidence that you can use your findings to describe a visitor population) around the results. There is another factor that also affects precision: the variability of the thing being measured. The more something varies from person to person, for example height varies more than number of fingers in people, the bigger your sample needs to be to achieve the same degree of certainty about your results.

For more information, contact Rebecca Cors rcors@wisc.edu

Appendix: About Purposive Sampling

Purposive sampling, also known as judgmental, selective, or subjective sampling, is a form of non-probability sampling in which researchers rely on their own judgment when choosing members of the population to participate in their study. Here's a summary (italic text below) from the Alchemer Survey Tools group (Alchemer, 2018).

This sampling method requires researchers to have prior knowledge about the purpose of their studies so that they can properly choose and approach eligible participants.

Researchers use purposive sampling when they want to access a particular subset of people, as all participants of a study are selected because they fit a particular profile.

Purposive Sampling vs. Convenience Sampling

The terms purposive sampling and convenience sampling are often used interchangeably, but they do not mean the same thing.

Convenience sampling is when researchers leverage individuals that can be identified and approached with as little effort as possible. These are often individuals that are geographically close to the researchers.

Purposive sampling is when researchers thoroughly think through how they will establish a sample population, even if it is not statistically representative of the greater population at hand. As the name suggests, researchers went to this community on purpose because they think that these individuals fit the profile of the people that they need to reach.

While the findings from purposive sampling do not always have to be statistically representative of the greater population of interest, they are qualitatively generalizable.

The more prior information that researchers have about their particular communities of interest, the better the sample that they're going to select.

How is Purposive Sampling Conducted?

The method for performing purposive sampling is fairly straightforward. All a researcher must do is reject the individuals who do not fit a particular profile when creating the sample.

However, researchers can use various techniques during purposive sampling, depending on the goal of their studies.

Technique Options Used in Purposive Sampling

Technique options include, but are not limited to, the following.

Typical

Typical case sampling is a type of purposive sampling that's useful when a researcher is looking to investigate a phenomenon or trend as it compares to what is considered typical or average for members of the a population.

Extreme or Deviant

Extreme or deviant case sampling is the opposite of typical case sampling. It is used when researchers want to investigate the outliers from the "norm" when it comes to a particular trend. By looking into these outliers, researchers are able to develop a stronger understanding of behavior patterns in the population.

For more information, contact Rebecca Cors rcors@wisc.edu

Critical

Critical case sampling is a type of purposive sampling in which one case is chosen for investigation because researchers believe that by investigating it, insights into other similar cases will be revealed.

Maximum Variation

A maximum variation purposive sample is also referred to as a heterogeneous purposive sample. Researchers use this technique when they are looking to examine a diverse range of cases that are all relevant to a particular phenomenon or event. This allows researchers to gain as much insight from as many angles as possible.

Homogenous

A homogenous purposive sample is the opposite of a maximum variation purposive sample, as it is selected because members of the sample have a shared characteristic or a shared set of characteristics.

Benefits of Purposive Sampling

Purposive sampling enables researchers to squeeze a lot of information out of the data that they have collected. This allows researchers to describe the major impact their findings have on the population.

Purposive sampling is a popular method used by researchers due to the fact that it is extremely time and cost effective when compared to other sampling methods.

Further, the numerous technique options outlined above make purposive sampling a versatile research method that can be tailored to enhance a study's effectiveness.

Sometimes purposive sampling may be the only appropriate method available if there are a limited number of primary data sources that can contribute to the study.

Drawbacks of Purposive Sampling

The primary downside to purposive sampling is that it is prone to researcher bias, due to the fact that researchers are making subjective or generalized assumptions when choosing participants.

When researchers need to ensure that they are eliminating as much bias as possible, they are better off using a form of probability sampling.

However, researcher bias is only a real threat to a study's credibility when the researcher's judgements are poorly considered, or when they have not been based on clear criteria.

For a similar reason, it can be difficult for researchers to convince others that their study has significant representativeness of the larger population of interest.

Due to the fact that researchers are using their personal judgement to select participants and units of measurement, it can be a challenge to convince an audience that if different options were used, the overall findings would still hold true.

For more information, contact Rebecca Cors rcors@wisc.edu

Appendix: Conway's Rough Guide to Sample Size Calculation

This section from Conway's Rough Guide to Sample Size Calculation (Conway, 2018) gives guidelines for sample sizes for studies which measure the proportion or percentage of people who have some characteristic, and for studies which compare this proportion with either a known population or with another group. This characteristic can be a disease, and opinion, a behaviour, anything that can be measured as present or absent. Prevalence is the technical term for the proportion of people who have some feature. You should note that for a prevalence to be measured accurately, the study sample should be a valid sample. That is, it should not contain any significant source of bias.

1.1 Sample size for simple prevalence studies. The sample size needed for a prevalence study depends on how precisely you want to measure the prevalence (precision is the amount of error in a measurement). The bigger your sample, the less error you are likely to make in measuring the prevalence, and therefore the better the chance that the prevalence you find in your sample will be close to the real prevalence in the population. You can calculate the margin of uncertainty around the findings of your study using confidence intervals. A confidence interval gives you a maximum and minimum plausible estimate for the true value you were trying to measure.

Step 1: decide on an acceptable margin of error. The larger your sample, the less uncertainty you will have about the true prevalence. However, you do not necessarily need a tiny margin of uncertainty. For an exploratory study, for example, a margin of error of ±10% might be perfectly acceptable. A 10% margin of uncertainty can be achieved with a sample of only 100. However, to get to a 5% margin of error will require a sample of 384 (four times as large).

Step 2: Is your population finite? Are you sampling a population which has a defined number of members? Such populations might include all the physiotherapists in private practice in Ireland, or all the pharmacies in Ireland. If you have a finite population, the sample size you need can be significantly smaller.

Step 3: Simply read off your required sample size from table 1.1.

Acceptable margin of error	Size of population					
	Large	5000	2500	1000	500	200
±20%	24	24	24	23	23	22
±15%	43	42	42	41	39	35
±10%	96	94	93	88	81	65
±7.5%	171	165	160	146	127	92
±5%	384	357	333	278	217	132
±3%	1067	880	748	516	341	169

Table 1.1

Sample sizes for prevalence studies

For more information, contact Rebecca Cors rcors@wisc.edu

References

Alchemer (formerly SurveyGizmo). (2018, October 16). Purpose sampling web page from survey software provider. https://www.alchemer.com/resources/blog/purposive-sampling-101/#:~:text=The%20primary%20downside%20to%20purposive,participants%20for%20their%20online%20survey. .

Cors, R. (2016). Informal science learning: An investigation of how novelty and motivation affect interest development at a mobile laboratory (Dissertation, no.5019). University of Geneva. Available online https://archive-ouverte.unige.ch/unige:91515.

Kelly, Linda. (2014). Do You Need to Eat a Whole Pie to Taste the Pie? An October 2 #THROWBACKTHURSDAY entry in the EVALUATION AND VISITOR RESEARCH NATIONAL NETWORK (EVRNN) blog. https://musdigi.wordpress.com/2014/10/02/do-you-need-to-eat-a-whole-pie-to-taste-the-pie-throwbackthursday/.

Conroy, R.M. (2018). Sample size. A rough guide. Online resource downloaded on October 16 from Semantic Scholar website.

Field, A. (2013). Discovering Statistics Using IBM SPSS Statistics. London: SAGE publications.Pallant, J. (2013). SPSS Survival Manual: A step by step guide to data analysis using SPSS. New York: Open Universities Press.

Pallant, J. (2013). SPSS Survival Manual: A step by step guide to data analysis using SPSS. New York: Open Universities Press.

Scribbr (2022, February 24). Reliability vs Validity in Research | Differences, Types and Examples. https://www.scribbr.com/methodology/reliability-vs-validity/.

Tabachnick, B. G., & Fidell, L. S. (2013). Using Multivariate Statistics (Vol. Sixth Edition). New Jersey: Pearson Education, Inc.