Beginner's Guide to Measuring Educational Outcomes in CEhp

Concepts Involved in Sampling Data

By Melanie D. Bird, PhD, American Academy of Family Physicians; Erik D. Brady, PhD, CHCP, Clinical Care Options

Appropriate and accurate research design is essential for successful measurement of outcomes in any educational activity. A major component of the design is deciding who will be included in the study or the sample. Gathering data from all learners and potentially affected patients is time and resource prohibitive, and as discussed in previous articles, the burden of data collection has increased. Therefore, most researchers use discreet and manageable sample populations.

Scope of this Article

The goal of this article is to provide a review on foundational concepts in sampling, discussing sampling terminology, external validity, sampling methods, errors and bias, group formation, and a brief overview of statistics relevant to sample size. It will also focus on collection of baseline and post-activity outcomes in regard to sample composition and positioning of questions to assess change in knowledge or competence.

Importance of Sampling

Sampling is the process of selecting participants from a particular population to represent that population as a whole. You can select a subset of your overall targeted learners, allowing for extrapolation of the results to the entire population of interest. It is an important tool for research in CEhp outcomes, as it is much easier to work with a smaller group instead of a large population and, arguably, saves time and money. It also allows for more control of the study and less risk of human error in data entry and analysis.

In general, sampling requires a statement of who the targeted learners are (e.g., degree type, board certification or specialty, experience) and/or who the learners' patients are (e.g., practice setting, type of patients seen, geography of the practice, etc.). It is important to understand the concept of external validity or the extent that the study results can be generalized back to the population at large. In other words, would all the relevant providers show an increase in knowledge, competence or performance following your educational activity, or only those directly similar to your sample population?

Probability Sampling

Most studies employ a sampling plan[1-2] to create the sample population. Examples of these plans can be found in **Table 1.** Probability sampling involves a deliberate and unbiased plan that allows for every sample unit to have an equal chance of being included in the sample.

Simple random sampling would involve selecting participants in such a way that every possible person has an equal chance of being selected, which is the equivalent of drawing names from a hat. This can be a challenging approach in CEhp activities, particularly in instances where a significant percentage of your learners are not members of the activity's targeted audience. Examples include an in-person activity targeted specifically at physicians held in a setting open to many provider types or an online activity that is targeted to health care teams, but does not restrict patients from accessing the content and participating in the outcomes measurement. Both examples allow for non-targeted learners to participate in the outcomes study, making a completely random sampling subject to analysis that leans on data that would otherwise be dropped.

Other categories include systematic sampling, which often involves random, computer-generated numbers used to select participants for the sample population.

Stratified random sampling involves placing participants into mutually exclusive sets, clusters or strata, and then randomly selecting from each set. Examples of strata might include age, sex, practice setting, geographic parameters, etc. By ensuring randomness into the sample selection, we can limit sampling error and subsequent bias in our data and increase the external validity of the study. Random sampling is used in CEhp programs, not only to limit bias but for two other tangible reasons: It requires the least amount of forethought in the design of

Table 1. Examples of Sampling Methods for CE Activities.

Random Sampling	Simple	Participant names are placed in a pool and then are selected one at a time at random to receive a survey following a CE activity.
	Systematic	Participants are assigned a computer-generated number at time of enrollment in a CE activity. Those with particular numbers are sent a survey.
	Stratified	Participants attending a CE activity on atrial fibrillation management are divided into specialty areas: family medicine, cardiology, gerontology, surgeons. Participants from each group (strata) are picked at random to be in the sample.
Nonrandom Sampling	Convenience	A portion of participants in a CE activity are asked to complete a survey based on proximity to the host of the activity.
	Consecutive	All participants at the CE activity complete the survey, but only target learners are included in the sample.
	Snowballing	Participants in a CE activity complete a survey and submit additional names to be contacted for inclusion in the sample.

the outcomes tool, and it allows the analyst to report the highest participation possible in the outcomes study.

Nonprobability sampling

In cases where time, money, or other issues are constraining, investigators may use nonprobability (nonrandom) sampling. In these cases sample units are not selected randomly but are selected based on accessibility or judgment of the researcher. Nonrandom sampling methods are less stringent and widely used; however, there is a greater chance of bias in the sample, decreasing the external validity.

A method for nonrandom sampling includes convenience sampling, which uses easily accessible subjects. If all accessible subjects are included in the group, we call it consecutive sampling. Consecutive sampling is often used in CEhp outcomes analysis. This method allows the analyst to efficiently eliminate non-target audience members, like non-health care providers or specific provider types, when the activity measured is not intended to address their educational needs, while maximizing the number of outcomes participants. Other methods include quota sampling, in which individuals are included in equal numbers in each group based on a specific trait (age, sex, type of practice) or snowball sampling, where recruited subjects are asked to identify others to include in the sample.

Sampling Errors

Two types of error can result from using a sample population: sampling error and non-sampling error. Sampling

> "The lower the standard deviation and the larger the sample size, the smaller the sample error becomes."

error, also called random error, results from differences in the sample compared to the population of interest. For example, even with a random sample, we might end up with too many providers from one geographical region, practice type or specialty. Sampling error is random and out of our control, but can be limited through increased sample size. Sampling error refers to the level of precision and can be expressed in percentage points. For example, if sampling error is low and our level of precision is ± 5 percent then we can expect our results to fall within that range.

Non-sampling error results from a systematic error that can lead to bias in the study. Usually, this error occurs due to mistakes in data entry or acquisition or inappropriate sampling methodology from poor planning and inattention to detail. Non-sampling errors can result from three major areas: errors in data acquisition, non-response errors and selection bias. Errors in data acquisition occur when the recording of responses is incorrect,

due to mistakes made from transcription of primary information, equipment error or faults, inaccurate responses resulting from incorrectly written or ambiguous questions (responder misinterpretation) and more. Non-response error or bias occurs when responses are not obtained from some members of the sample. This type of error results in either a substantially smaller sample size that may no longer be representative of the population, or if the responders' answers are extrapolated to the non-responders, investigators may reach an incorrect conclusion. For example, we may incur non-responder bias to a survey on satisfaction for examination preparation. Participants who scored well may be more motivated to respond than those that did not, thereby skewing the results. Similar to non-response bias, selection bias occurs when some members of a target population cannot be selected for inclusion in the sample. Increasing the sample size will not alleviate non-sampling errors.

Sample Size

Some statistical knowledge is needed in order to understand the importance of sample size.[1-4] The sample size is dependent on several parameters used in inferential statistics related to the data collected. As outlined in the first three articles of this series, the type and amount of data will vary depending on the outcomes being measured and the number and types of questions used for the assessment. Each response is a data unit, and for each data unit a sampling statistic can be calculated (mean, median, mode). To translate the sampling statistic back to the population of interest, we need to understand the distribution of our sample. The sampling distribution is the spread, or possible values of a statistic, across an infinite number of samples and resembles a bell shaped curve when graphed,[1] as shown in **Figure 1.**

The statistic or parameter observed represents just one of infinite possibilities. The spread of scores around the parameter for our population is called the standard deviation (often abbreviated to SD, or denoted by the Greek letter σ). The spread of scores across the sampling distribution is the standard error (sampling error, or SE).[1] The standard error is calculated using the standard deviation and sample size. The lower the standard deviation and the larger the sample size, the smaller the sample error becomes. The sample size needed for measurement of outcomes for a particular CE activity is dependent on the amount of acceptable error related to how big a difference you would like to find. A small difference would require a large sample size. If you are looking for larger differences, then a smaller sample will be sufficient. **Table 2** outlines how different levels of error require varying num-

Table 2. Sample size required varies on population and precision level (level of error).

Population size	5%	10%
10	10	n/a
50	44	n/a
100	81	51
500	222	83
1,000	286	91
2,000	323	92
10,000	385	99
100,000	398	100

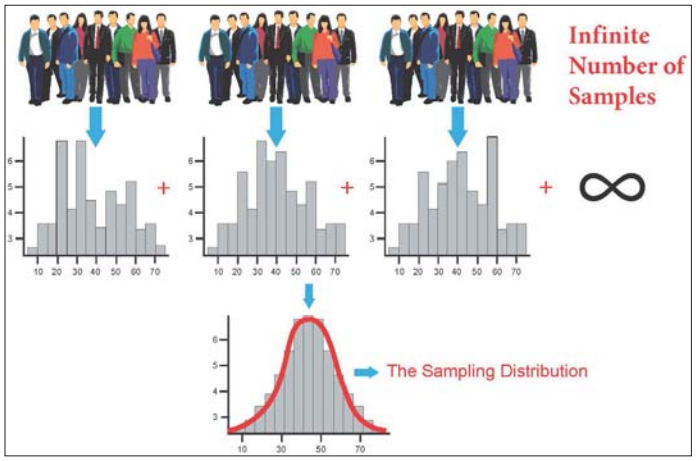

Figure 1: Sampling Distribution: mean, mode, median from an infinite number of samples.

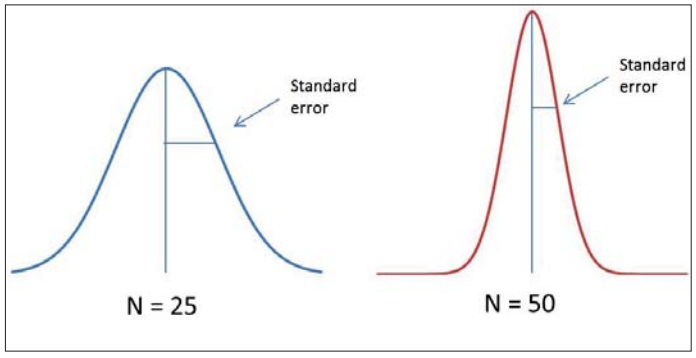

Figure 2: Increased sample size results in a narrower sample distribution and decreased standard error.

bers of participants or respondents.[3] For more precise outcome measurement, the error level should be lower, requiring a larger sample size. As shown in **Figure 2,** by increasing the sample size from 25 to 50 participants, the standard error becomes smaller. For a visual explanation of the impact of sample size on sampling error, see the first three minutes of R. Backman's video **"Sampling Error and Sample Size"**

As discussed in Articles 1 and 2 of this series, in outcome levels 3-7, changes in knowledge, performance and patient/community health are assessed. In order to demonstrate change following an intervention, researchers must create the appropriate groups within their sample: a control group and an experimental group. The experimental group receives the intervention (new educational activity) and the control group receives the standard educational activity or no education depending on the question being evaluated. In order to know how many participants should be included in each group, power analyses can be performed to provide guidance. The precision rate (sampling error), the confidence intervals and the degree of variability all impact the sample size, as seen in **Figure 3.**[4] The precision rate and confidence intervals reflect how sure one can be of the mean result where the degree of variability depends on the heterogeneity of the sample. The more diverse the sample is, the larger it will need to be to account for the variability and the more confident we are in the result. The actual calculation of sample size can be done using any number of websites and software, such as **www.VassarStats.net.** After knowing the size of the control and intervention groups, you must divide sample members between them.

Research Design and Sampling

To assess a change in knowledge, competence or performance, multiple research designs with varying number and composition of groups are available. Designs may include one measurement at the end of an activity, or intervention (posttest only), or a pretest and posttest, and may include repeated measures or tests over a specified duration. Activities that use only a single measurement post activity are particularly subject to the considerations previously described, outlining the need for a demographically matched control for the learner set involved in the outcomes study.

A more simple research design involves each participant serving as their own control, thus receiving the pretest, the educational intervention and then the posttest. This design is easy to set up and requires a more limited

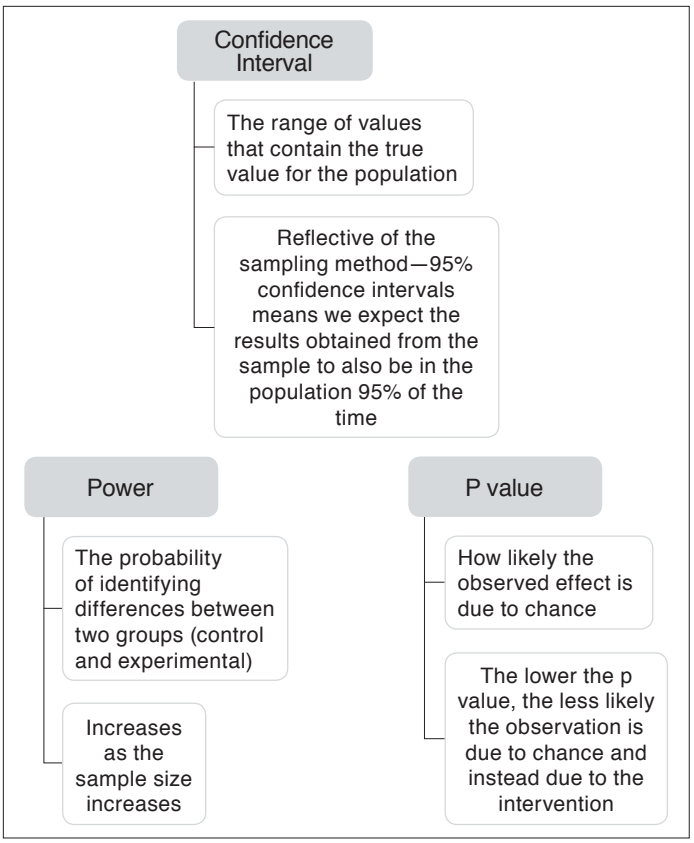

Figure 3: Factors affecting sample size include confidence interval, p value and power of a study.

sample size; however, it can lead to confounding variables with participants performing poorly on the posttest due to fatigue or performing better than expected due to practice. Even so, this design is likely the most typical design found in the CEhp space as the demographic matching of different sample populations is inherently more difficult to manage than item fatigue.

Another simple study set up is the two group posttest only design. In this case, sampling is performed to create a control group and experimental group. Both groups are given a test following the intervention (or control) and the results are compared. The issue with this design is that we cannot be sure if the two groups were similar at baseline without a pretest, which impacts the internal validity of the study.

To determine baseline, both the control and intervention group receive a pretest to determine baseline knowledge and then both groups take a posttest following a specified interval. During the interval, only the experimental group will receive the intervention. A benefit of the pretest-posttest design is that we increase internal validity be-

cause we know the two samples are similar at baseline and can effectively measure a change due to our intervention. A concern for this design is a potential loss of external validity as the pretest may influence the results. For example, participants in the control group previously unconcerned with a topic may do some self-study or get outside information leading to an increase in posttest scores similar to the experimental group. This may hinder our ability to generalize results back to the rest of the population (no pretest) and resulting in decreased external validity. A potential solution for this issue is to use the Solomon Four Group Design[5], where additional control and experimental groups are added that receive the posttest only.

Positioning of Outcomes Items within CEhp Activities

There are several designs in the collection of data that could be considered for CEhp activities and have a bearing on the sample size. One common design involves placing outcomes question items outside of the educational content, either on printed forms, or if collected via audience response system technology, preceding any informing content (for baseline items) and following all informing content (for post items). A drawback of this approach is that, for live activities, late arrivals and early departures may significantly truncate the potential sample size. There is also risk that the facilitator may not appropriately coach the learners to participate in the outcomes study, failing to adequately address a specific point that is focused on within one of the outcomes items. In addition, question fatigue is a risk in this scenario as the design places a potentially large set of questions in front of the learner at two points in time within the overall activity.

A more subtle design involves framing outcomes items tightly around the informing content, i.e., the content that should impact a learner's answer to the outcomes question item. Using this methodology addresses several issues with the placement of items outside of the content altogether. First, question fatigue is generally decreased. Items are spread out across an activity and are central to the content, making facilitators more apt to speak specifically to the outcomes items. Likewise, learners are more likely to offer a matched response to the items. Late arrivals and early departures are less likely to compromise your sample size as well, since learners are more likely to engage in the "heart" of the content.

> "The more diverse the sample is, the larger it will need to be to account for the variability and the more confident we are in the result."

For additional illustrated modules on the concepts of sampling data discussed here, check out **Khan Academy** (free login required). In the subject box, write "inferential statistics," and check out the modules on sampling distribution and confidence intervals.

Forecast of Next Article

In the next article, Gary Bird, PhD, and Sandra Binford, MAEd, will build on the basic points of sampling featured above and focus on the impact of sampling at various time points after an educational intervention.

For Further Reading:

1. Trochim, William M. The Research Methods Knowledge Base, 2nd Edition. http://www.socialresearchmethods.net/kb/sampstat.php (version current as of October 20, 2006). Accessed 4/1/15.
2. Lewis-Beck, MS. 2004. The SAGE Encyclopedia of Social Science Research Methods. Sage Publications.
3. Isaac, S. and Michael, WB. 1981. Recommended sample sizes for two different precision levels. Handbook in Research and Evaluation. 2nd Ed. San Diego, EdITS Publishers.
4. Suresh K, Chandrashekara S. Sample size estimation and power analysis for clinical research studies. Journal of Human Reproductive Sciences. 2012;5(1):7-13.
5. Solomon, RL. 1949. An extension of control group design. Psychol Bull. 46: 137-50.