Pre Data Analysis Activities

Once you obtain survey results the fun begins. Note however that analyzing and reporting survey data can be challenging in that it can be a time-consuming process. If results are not processed, analyzed and reported properly the results may be misleading and possibly inaccurate. There are several issues to consider and a few things to do before you get into the actual data analysis, interpretation, and reporting.

Cleaning and Organizing the data

Before data analysis can take place the information you obtained must be cleaned and organized. Cleaning data refers to the process of removing irrelevant data (as in the case where online surveys add variables to facilitate the survey's function), possibly deidentifying the responses (as required by IRB protocols), or coding open responses (see allowing OTHER responses). Cleaning data is needed prior to examining response patterns and identifying incomplete surveys. The data may also require reorganizing (e.g., collapsing categories or performing summary calculations). For example, you may wish to report the percentage of those *agreeing* or *strongly agreeing* to a statement in comparison to those who *disagree* or *strongly disagree*. The precise cleaning needs will be determined by the purposes and reporting needs of the survey.

Systematic Nonresponse Review

Once the data is cleaned a review of the data is needed to ensure no systematic survey nonresponse patterns have occurred (see Response Rate Issues). Even when random sampling is used, you may have a problem if response refusal patterns indicate that an important group of potential respondents failed to complete the survey. Key indicators may include age, gender or race patterns that do not match the expected demographics. However, other response refusal problems may cause a lack of generalizability; for example, location overrepresentation (e.g., urban vrs rural) or group underrepresentation (e.g., freshman vrs seniors). The degree to which this will be a problem depends on the degree to which the sample provides a reasonable representation of the population. Without pertinent information about respondents you will not be able to conduct this investigation. Therefore, it is important to identify key factors and plan to obtain these details prior to administering the survey. If an issue is identified, you may need to get additional responses from underrepresented individuals to solve the problem.

Determining the Response rate

One of the first things you should calculate once a survey has been administered and the results obtained is the response rate. The calculation is simple. The response rate is an indication of the number of invited participants who complete the survey. This is reported as a percentage but should always be accompanied by the size of the sample. A 66% response rate obtained from a sample comprised of 3 individuals wouldn't provide very compelling evidence. However, a response rate of this magnitude obtained from a much larger sample would be much more impressive. You should also indicate how the sample was obtained when reporting the sample size (the population size as well if it is known). The sampling procedures will have been detailed in the methods section but you should briefly reported this

(as a reminder) when presenting results (e.g., a random sample of 3 individuals).

Completed Survey Decisions

The only consideration (possible controversy) surrounding the response rate calculation is deciding what it means to complete the survey and whether information obtained from partially completed surveys might be usable. In order to make this determination, a careful examination of the survey results is needed. While there are no hard and fast rules, there are some principles that may help you make this determination. In situations where a returned survey is unusable it should not be included in the response rate calculation. This will also affect the margin of error calculation.

Requisite Data. A guiding principle for making an inclusion decision is to determine what information is absolutely essential in order to answer the research question. Sometimes a partially completed survey can be used to answer some of the research questions. Other times, missing a single item on the survey will render the information unusable. For example, if one section of the survey was completed and not another, part of the information provided might be usable. However, the information provided in a partially completed survey would likely need to be excluded if, for example, the respondent only completed the demographics section but nothing else, or they failed to provide vital grouping information required to disaggregate the data and answer the research question.

Accurate Data. Another factor that should be considered regarding the completeness of the survey is that of accuracy. Unfortunately, there are times when participants are not completely honest in the way they answer questions on a survey. You may not know how accurate the information provided will be but you can get indicators that the information is inaccurate. For example, suppose that while

pilot testing the survey you determine that a survey typically takes 10 to 15 minutes to read, reflect and answer all the questions. Then suppose a participant completed the entire survey in only 2 minutes. You might suspect the results to be inaccurate and quite possible unusable.

Inaccurate (unusable) data might also be identified by examining a participant's response pattern. For example, the results might be suspect if a respondent provided the same response for every question, even though it would be extremely unlikely that an honest respondent would actually answer in that way. Random response bias like this is more likely when incentive are provided to individuals for taking the survey. If trigger items where added to identify suspect response patterns, these need to be examined (see random response bias).

Response Rate Example

Returning to the counseling services example, suppose you administered the survey to 400 first-year undergraduates (n=400). Let's say the sample was selected using a simple random sampling procedure and was taken from the 5000 students enrolled (N=5000). However, only 114 students returned the survey. The response rate would be 28.5%.

While this may be considered adequate for the purposes of the study, suppose you realize that 20 of the surveys (for a variety of reasons) were incomplete to the extent that you could not use the information provided. This means the actual number of usable surveys is 94 bringing the actual response rate to 23.5%. In an attempt to increase the return rate, you decide to resend the invitation to those who declined participation with the promise of a $10 gift card redeemable at the university cafeteria. With this incentive you receive an addition 268 surveys. However, upon inspection you realize that hungry university students will do anything for a free meal. Most of the surveys (182) showed signs of random response bias or were incomplete. Removing these surveys, due to the fact that the information could not be used, the actual number of usable responses is 180. This means the response rate for the survey is really 45% even though 95.5% of those invited to take the survey returned the survey.

Acceptable Response Rates

Knowing the response rate is an important piece of information. Opinions about what constitutes an acceptable response rate is debatable. Ultimately, you need to feel confident that the information you obtained will adequately serve the purposes of the survey and

answer the research questions. For some research purposes, only near perfect response rates with no percievable systematic response refusal patterns would be acceptable. In other situations, a much smaller response rate might suffice. Response rates for published social science research can range from 25 to 75 percent and a 30% response rate is typical (Baruch & Holtom, 2008). It is up to the researcher to persuasively argue that the response rate obtained is sufficient.

Estimating the Margin of Error

For some surveys you may need to calculate the margin of error. You can calculate a margin of error for each of the values obtain from the survey. The margin of error is an estimate of the amount of error we might expect for each outcome. In practice, the margin of error is a confidence interval. Any statistics we obtain from a survey is an estimate that includes some amount of error. We don't actually know what the real (true) value is but we can be somewhat confident that the true value will fall within a specific range based on an estimate of the standard error (SE) and a specified confidence level (z).

Margin of Error Calculation (proportion)

In order to calculate the margin of error for a result represented by a proportion we need three values: the confidence level, the sample size (i.e., number of completed surveys with usable data), and the sample proportion. If the population size is known (i.e., you have a finite population), the formula can be adjusted to account for any error that might occur from using a sample instead of taking a census. The modified formula (using the finite population correction, or fpc) assumes you know the population size.

- p = the sample proportion
- n = sample size (number of usable surveys)
- N = population size
- Z = z-value representing the desired confidence level

  2.576 for 99% level of confidence

  1.96 for 95% level of confidence

  1.645 for 90% level of confidence

Margin of Error Example for a proportion

Continuing with the counseling services example, let's say you obtain a result where 84% of respondents selected a specific option on one item (p=.84). This means 16% selected a different option (1-p or .16). Suppose we decided to use a 95% confidence level which would make Z=1.96. Given a sample size of n=180 that would make the margin of error equal to 2.73. Given this margin of error, we can say the result is assumed accurate within plus or minus 2.73 percentage points with a 95% confidence level. However, adjusting for the fact that this is a finite population (N=5000), an adjusted estimate suggests the margin of error might actually be 2.68. These estimates are quite close and both round to 2.7 percentage points. This means the statistic obtained might reasonably be anywhere between 82.3 and 86.7 percent (84 ± 2.7%).

Margin of Error Calculation (mean values)

In order to calculate the margin of error for a result represented by a mean we need three values: the confidence level, the sample size (i.e., number of completed surveys with usable data), and the standard deviation of the sample mean. If the population size is known (i.e., you have a finite population). The formula can be modified using the finite population correction (or fpc).

- σ = standard deviation of the sample mean
- n = sample size (number of usable surveys)
- N = population size
- Z = z-value representing the desired confidence level

Example

When the survey result is a mean rather than a proportion the standard error calculation uses the standard deviation of the sample mean. Suppose you asked people how often (in days/week) they experience feeling of depression and you determine that the average response was 2 days with a standard deviation of 2.8 ($\sigma = 2.8$). Using a 95% confidence level and given a sample size of n=180, the margin of error would be .409. Given this margin of error, we can say the result is assumed accurate within plus or minus .409 days at a 95% confidence level. However, adjusting for the fact that this is a finite population (N=5000), an adjusted estimate suggests the margin of error might actually be .402. These estimates again are quite close, around 0.4 days. This means the statistic obtained might reasonably be anywhere between 1.6 and 2.4 days each week (2 ± 0.4).

Chapter Summary

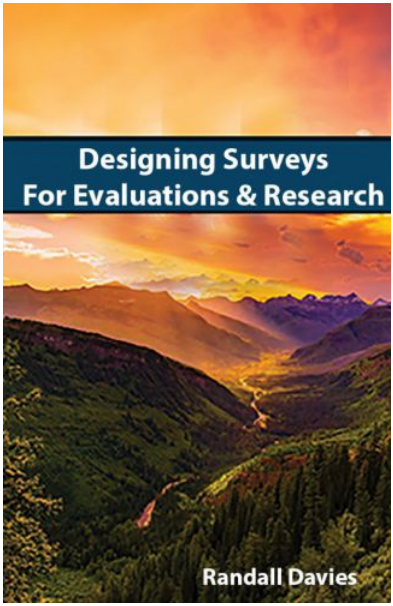
- Data cleaning is required prior to starting the data analysis process.
- If results are not processed, analyzed and reported properly the results may be misleading and possibly inaccurate.
- Data should be reviewed to identify unusable surveys and explore the possibility of any systematic response refusal pattern.
- Response rates along with sample size and sampling methods should be reported.
- An adequate response rate is needed to obtain a representative sample.
- Each statistic obtain from a survey is only an estimate of the true population parameter.
- A margin of error calculation can be used to provide a confidence interval for each sample statistic.

Discussion Questions

1. What impact would you expect if you found systemic response refusal pattern had occured? What step might you consider taking to eliviate the problem?
2. How does the response rate affect the sampliing process?

References

Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human relations*, *61*(8), 1139-1160.



Davies, R. S. (2020). *Designing Surveys for Evaluations and Research*. EdTech Books. Retrieved from https://edtechbooks.org/designing_surveys