What is program evaluation?

A beginners guide

Last revised 3/10/2020

Gene Shackman, Ph.D.
The Global Social Change Research Project
Free Resources for Methods in Program Evaluation
https://ssrn.com/author=1754255

With contributions, suggestions, reviews by Michael Quinn Patton, David Fetterman

Table of Contents

Introduction

In this guide, we focus on one particular kind of evaluation: *program evaluation*.

Basically, when beginning an evaluation, program people will often want the answer to this question:

- Does the program work? And how can it be improved?

There are many equally important questions

- Is the program worthwhile?
- Are there alternatives that would be better?
- Are there unintended consequences?
- Are the program goals appropriate and useful?

Evaluations, and those who request them, may often benefit from a consideration of all of these questions. *This guide,* however, focuses on the first of these issues: does the program "work", and how program evaluation can contribute to improving program services.

How does program evaluation answer questions about whether a program works, or how to improve it.

Basically, program evaluations **systematically collect and analyze data** about program activities and outcomes.

The purpose of this guide is to briefly describe the methods used in the systematic collection and use of data.

Additional Resources

“Program evaluation answers questions like: To what extent does the program achieve its goals? How can it be improved? Should it continue? Are the results worth what the program costs? Program evaluators gather and analyze data about what programs are doing and accomplishing to answer these kinds of questions.”

American Evaluation Association blog, January 2014. https://www.eval.org/p/bl/et/blogid=2&blogaid=4

Norms and Standards for Evaluation. United Nations Evaluation Group. 2016
http://www.unevaluation.org/document/detail/1914
Definition of Evaluation

1. An evaluation is an assessment, conducted as systematically and impartially as possible, of an activity, project, programme, strategy, policy, topic, theme, sector, operational area or institutional performance. It analyses the level of achievement of both expected and unexpected results by examining the results chain, processes, contextual factors and causality using appropriate criteria such as relevance, effectiveness, efficiency, impact and sustainability. An evaluation should provide credible, useful evidence-based information that enables the timely incorporation of its findings, recommendations and lessons into the decision-making processes of organizations and stakeholders.
2. The purposes of evaluation are to promote accountability and learning. Evaluation aims to understand why — and to what extent — intended and unintended results were achieved and to analyse the implications of the results. Evaluation can inform planning, programming, budgeting, implementation and reporting and can contribute to evidence-based policymaking, development effectiveness and organizational effectiveness.

Planning

What is Program Evaluation

The first step is planning. Evaluations should follow a systematic and mutually agreed on plan. Plans typically include the following:

- Determining the purpose of the evaluation (eg accountability, improvement, learning):
- Understanding the program or intervention to be evaluated.

The purpose of the evaluation and understanding the program will determine the next three parts of the plan:

- What is the evaluation question, what is the evaluation to find out.
- How will the evaluation answer the question: what is the evaluation design, how will the data be collected.
- Resources needed: e.g., personnel, resources, skills, time

Finally, the plan should include:

- How to engage the stakeholders. Helps determine the questions and that the results will be used.
- Making the results useful, how will the results be reported so that they can be used by the organization to make improvements.
- Who will own the data.

Sources and additional resources about planning evaluations:

Better Evaluation. Manage Evaluation. https://www.betterevaluation.org/en/rainbow_framework/manage_evaluation

The Program Manager's Guide to Evaluation, Second Edition, 2010. https://www.acf.hhs.gov/opre/resource/the-program-managers-guide-to-evaluation-second-edition

Centers for Disease Control and Prevention. Framework for Program Evaluation in Public Health. MMWR 1999;48(No. RR-11). https://www.cdc.gov/eval/framework/index.htm

A key part of evaluation is to determine the question.

Evaluations can generally answer two types of questions:

1. Outcome evaluation: What is the outcome of the program? Did the program have any impact, was there any improvement in people's lives?

2. Process evaluation: How did the program get to that outcome? Did the program have some set of procedures? Were these procedures followed, were the procedures reasonable, was there a better way to get to the outcomes?

The next section focuses on how to determine the question for your evaluation, describing a few tools which might be useful.

Additional resources:

Approaching An Evaluation-- Ten Issues to Consider
Brad Rose Consulting, Inc.
https://bradroseconsulting.com/approaching-an-evaluation/

Logic Model

One way to figure out the questions is for the evaluator and program people to develop a very good description of:

- What the program/intervention is doing
- What the outcomes should be,
- How the program/intervention is supposed to get there, and
- Why the program leads to the outcome.

This description helps to identify how the program should lead to the outcome, why the program activities should lead to the outcomes, and where to evaluate the program to check whether it does.

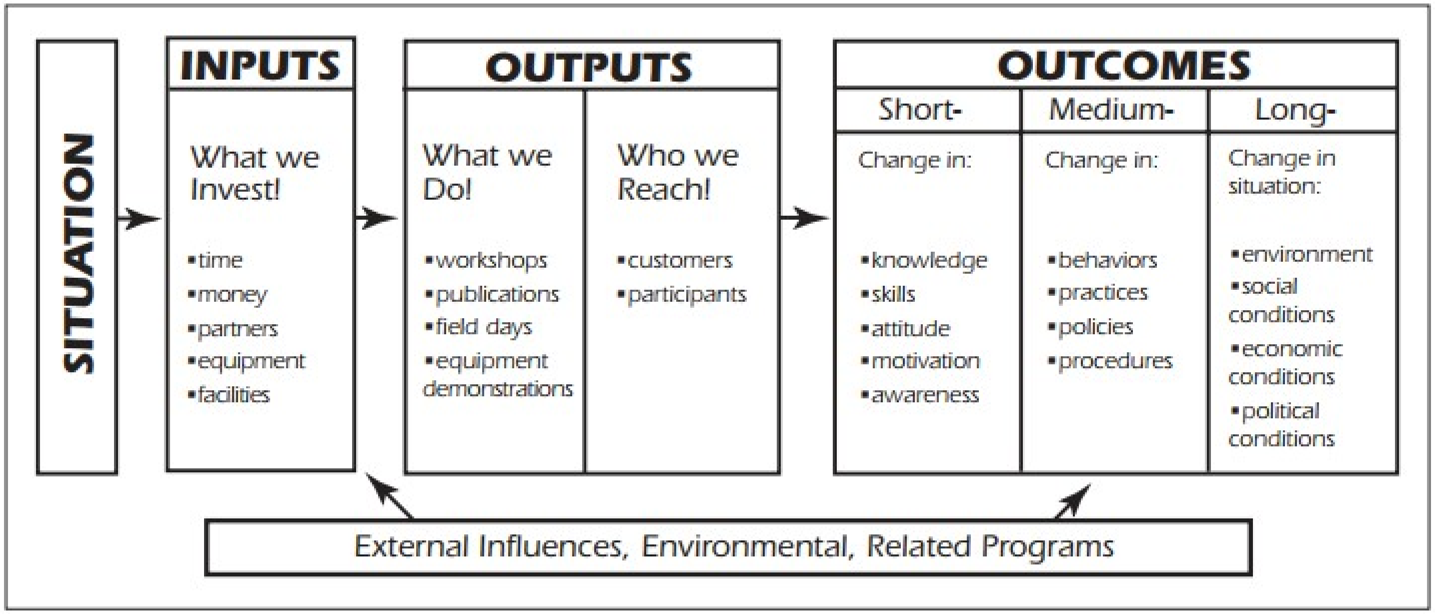
This method is called a *program theory*.

“A program theory explains how and why a program is supposed to work. ... It provides a logical and reasonable description of why the things you do – your program activities – should lead to the intended results or benefits.”

From Program Evaluation Tip Sheets from Wilder Research, Issue 4, October 2005 - Program Theory.
https://dps.mn.gov/divisions/ojp/forms-documents/Documents/Wilder_Program_Evaluation_4.pdf

A useful tool to help work with the program theory is a *logic model*, which visually shows the program theory, how all the program goals, activities, and expected outcomes link together.

Logic model example:



from
The Logic Model for Program Planning and Evaluation
Paul F. McCawley, Associate Director, University of Idaho Extension. 2001.
https://www.extension.uidaho.edu/publishing/pdf/CIS/CIS1097.pdf

Use the program theory or logic model to come up with evaluation questions, for example:

- Does the program have a positive outcome?
- Are people satisfied?
- How could the program be improved?
- How well is the program working?
- Is the program working the way it was intended to work?

Additional Resources

Asking Program Evaluation Questions
Ally Krupar, American University, Online Programs. No date given.
https://programs.online.american.edu/online-graduate-certificates/project-monitoring/resources/asking-program-evaluation-questions

Developing Process Evaluation Questions. At the National Center for Chronic Disease Prevention and Health Promotion. Healthy Youth. Program Evaluation Resources
https://www.cdc.gov/healthyyouth/evaluation/index.htm

However, program theory and logic models can have limits, if used improperly:

- Models are linear, programs are complex, interactive
- Models are static, programs may change over time.
- Models may not take unexpected consequences into account
- Models may not account for conflict, power, control issues
- Theory or model assumes the model is correct.
- Model may under-emphasize processes or external influences.

To reduce the limits, use program theory and logic models, but be flexible, and open to change and feedback. Constantly review and revise them, as necessary.

Additional Resources about logic models.

Enhancing Program Performance with Logic Models. Program Development and Evaluation. Ellen Taylor-Powell, Ph.D. and Ellen Henert. 2008 https://fyi.extension.wisc.edu/programdevelopment/logic-models/

Introduction to logic models, Sarah Morgan-Trimmer, Jane Smith, Krystal Warmoth and Charles Abraham
Public Health England. 7 August, 2018
https://www.gov.uk/government/publications/evaluation-in-health-and-well-being-overview/introduction-to-logic-models

Dana Petersen, Ph.D., Erin F Taylor, Ph.D., and Deborah Peikes, Ph.D. The Logic Model: The Foundation to Implement, Study, and Refine Patient-Centered Medical Home Models. March 2013. AHRQ Publication No. 13-0029-EF .
https://pcmh.ahrq.gov/page/logic-model-foundation-implement-study-and-refine-patient-centered-medical-home-models

Theory of Change

A tool that can be used in addition to or instead of the logic model is the Theory of Change (TOC). The TOC basically includes an in depth explanation of how the program will achieve it's goals. While the logic model illustrates the various resources, activities, tasks and outcomes involved in the program, the TOC may have more detail about how these parts relate to each other, assumptions about the program, and necessary preconditions for the intervention or program.

Having more details spelled out may provide more information about how the program is supposed to work, and how it is supposed to achieve it's goal, and so may provide more information about what to evaluation.

Additional Resources about Theory of Change

Lovely Dhillo and Sara Vaca, Refining Theories of Change. Journal of MultiDisciplinary Evaluation. Volume14, Issue30, 2018. http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/496/444

Theory of Change. Better Evaluation. 2014 https://www.betterevaluation.org/en/resources/guide/theory_of_change

Center for Theory of Change. What is Theory of Change. No date given. http://www.theoryofchange.org/what-is-theory-of-change/

Developing a theory of change
Jessica Sellick, Rural Services Network, 3 November 2019
https://www.rsnonline.org.uk/developing-a-theory-of-change-mapping-out-the-missing-rural-middle-or-redescribing-a-process

Design

What is Program Evaluation

As described in the introduction, the key focus for this guide is how to answer the question of whether the program or intervention “worked”, usually meaning did the program have an effect, did it cause an outcome.

That is, the ultimate goal of a program is to improve people's lives. How do you know whether it did?

One commonly used way to find out whether the program improved people's lives is to ask whether the program **caused the outcome.** If the program caused the outcome, then one could argue that the program improved people's lives.

On the other hand, if the program **did not cause** the outcome, then one would argue that, since the program did not cause the outcome then the program did not improve people's lives.

How to figure this out?

Determining whether a program caused the outcome is one of the most difficult problems in evaluation, and not everyone agrees on how to do it. Some say that randomized experiments are the best way to establish causality. Others advocate in-depth case studies as best. The approach you take depends on how the evaluation will be used, who it is for, what the evaluation users will accept as credible evidence of causality, what resources are available for the evaluation, and how important it is to establish causality with considerable confidence. (This paragraph suggested by Michael Quinn Patton.).

The design of the evaluation is determined in large part by the question which the evaluation is trying to answer. In turn, the design largely determines the methods used to collect the data. Thus, the design is a key part of the plan of the evaluation.

There are three approaches frequently used to establishing whether a program or intervention causes an outcome.

- Experimental and quasi experimental design – for example using comparison groups, comparing people in the program to people not in the program
- Multiple evaluation methods – comparing results from several evaluations, each using different methods
- In depth case studies of programs and outcomes – showing that the links between what program participants experience and the outcomes attained are reasonable, empirically validated, and based on multiple sources and data. The linkages between program and outcomes are direct and observable. No alternative possible causes offer a better explanation. (See the in depth case study section later.)

The particular method that is used should reflect a careful discussion and understanding of the pros and cons of each method, and agreement among all parties involved.

Experimental Design:

Comparison groups and random assignment

The main idea of experimental design is that units of study (typically people) are randomly assigned either to a treatment or to a control group. The idea is this:

Randomly assign people to either be in the program (the 'treatment' group) or to be not in the program (the 'control' group).

Measure people in the control and in the treatment groups. Since people in both groups were randomly assigned, then before the program the two groups of people should be pretty much the same. Test to make sure they are.

Expose the people in the treatment group to the treatment or program.

Then measure the treatment group after the program and the control group at the same time.

After the program, if the 'treatment' group is better off than is the control group, then the difference should be from being in the program, and so it is reasonable to argue that the program caused, or contributed to the cause, of that outcome.

Additional Resources:

Why randomize. Institution for Social and Policy Studies. https://isps.yale.edu/node/16697

Randomized Control Trial. Angela Ambroz, and others. Better Evaluation. https://www.betterevaluation.org/en/plan/approach/rct

Understanding and misunderstanding randomized controlled trials. Angus Deaton and Nancy Cartwright. Social Science and Medicine, August 2018, Vol 2010. https://www.sciencedirect.com/science/article/pii/S0277953617307359

Advantages and disadvantages of random assignment to treatment and comparison groups.

Advantages:

- Results provide clearest demonstration of whether a program causes an outcome.
- Provides results that are easiest to explain.

Challenges:

- Often not practical to do. Often can't randomly assign people to program or not program, and may be unethical to randomly assign someone to no treatment.
- Randomly assigning people to be in the program is not how programs really work, so results of the evaluation may not apply to the program as it really exists.
- Very difficult to apply to causes that operate over the long term or to programs that are very complex.
- Can tell whether a program caused outcome, but doesn't give much in depth information about why or how. Qualitative research may help with the why and how.
- In true experimental design, people shouldn’t know whether they are in the experimental or control group (getting treatment or not getting treatment). But in most programs, people know whether they are getting treatment so outcome may be influenced by knowledge that they are getting treatment.
- In many programs, people in “control” group often don’t have no treatment, but may seek out alternative treatments.

These are summary of points from:
Munck and Verkuilen. Research Designs, 2005
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2480797
and
Michael Scriven. A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research. Journal of MultiDisciplinary Evaluation, Volume 5, Number 9, 2008. http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/160

Quasi-experimental design

Comparison groups and non random assignment When random assignment is not possible, quasi-experimental design can be used. There are many reasons why random assignment can't be used. In some cases, random assignment may be unethical. If "no treatment" is unsafe, then cannot randomly assign people to treatment or no treatment. Random assignment may not be practical, for example, if an existing program is open to everyone, so cannot control who participates and who does not.

There are many types of quasi-experimental design. These are a few.

- Comparing people already on the program to those who are not on the program. One example is to observe (O) people before they join the program or there is an intervention (X), then observe both groups after :

Pretest-posttest design

-Intervention group	O_1	X	O_2
-Comparison group	O_1		O_2

- Measuring the client many times before they join the program (or before a new intervention) and many times afterward, them compare before to after. One example is:

Time series design

-Intervention group	O_1	O_2	X	O_3	O_4

- Combination of the two above

Time series design

-Intervention group	O_1	O_2	X	O_3	O_4
-Control group	O_1	O_2		O_3	O_4

A major challenge to non random assignment approaches is that people on the program may start off being very different from the people not on the program.

That is, only some people choose to be on the program. Something made these people different and it may be the something which caused the better outcome, not the program.

One way to deal with this is to collect as much information as possible on characteristics of the people and program that relate to the program outcome (what the program is supposed to do), and use this information in statistical analysis to “control” for the differences between people on the program vs people not on the program.

The problem is that there may be differences, some critical, that are not observed, and for which the evaluator has no data.

Additional Resources:

AllPsych On Line. By Dr. Christopher L. Heffner
Section 5.3 Quasi-Experimental Design
http://allpsych.com/researchmethods/quasiexperimentaldesign.html

Study Design 101
https://himmelfarb.gwu.edu/tutorials/studydesign101/

Diagrams on previous page from:
Measuring the Difference: Guide to Planning and Evaluating Health Information Outreach. Stage 4, Planning Evaluation. National Network of Libraries of Medicine, 2000
https://collections.nlm.nih.gov/catalog/nlm:nlmuid-100961143-bk

Quasi-Experimental Evaluation, Kristin Anderson Moore, Ph.D. 2008
https://www.childtrends.org/?publications=quasi-experimental-evaluations

Howard White and Shagun Sabarwal. Quasi-Experimental Design and Methods. 2014
https://www.unicef-irc.org/publications/pdf/brief_8_quasi-experimental%20design_eng.pdf

Multiple evaluation methods

Multiple evaluation methods could support the idea that the program causes the outcome if different sources agree.

For example, collect information from:

- Program participants
- Program staff
- Community members
- Subject experts
- Published research and reports

Collect data through many methods, for example:

- Surveys
- Interviews
- Observations
- Program administrative data

If data from different sources don't agree, it doesn't necessarily mean the results from any of the sources are not valid. However, the more agreement there is from different sources, the more confident you can be about your conclusions.

Additional Resources:

Encyclopedia of Social Science Research Methods: Triangulation
Contributors: Alan Bryman, 2004.
https://edge.sagepub.com/system/files/Triangulation.pdf

Triangulation In Social Work Reserach: The Theory and Examples of its Practical Applications. Mike Weyers, Herman Strydom, Arnel Huisamen. Social Work/Maatskaplike Werk 2008:44(2) https://socialwork.journals.ac.za/pub/article/view/251

In depth case study

An in depth case study can be used to demonstrate the connection between the intervention and the outcome.

An in-depth case study documents in detail what a group of participants experienced in a program and any ways in which they have changed so that the evaluator and users of the evaluation can make a judgment about the likelihood that the program led to the observed changes. For example, a group of chronic alcoholics go through a residential chemical dependency program. Their participation is fully documented. They return home maintaining their sobriety. They attribute their sobriety to the program as do their families, friends, and program staff. These multiple sources agree on the documented causal chain. The links between what they experienced and the outcomes attained are reasonable, empirically validated, and based on multiple sources and data. The linkages between program and outcomes are direct and observable. No alternative possible causes offer a better explanation. The preponderance of evidence points to a conclusion about causality. Establishing causality involves both data and reasoning about the findings. (This paragraph contributed by Michael Quinn Patton.)

Cause - Summary:

Random assignment is often seen as a very clear way to show whether a program causes an outcome. However, random assignment is often not practical or reasonable.

Other methods such as non random assignment, multiple evaluation methods, or in depth case studies may be more practical and can be used to give reasonable arguments about whether a program caused an outcome.

However, these methods are less certain in establishing that the program is the cause of the outcome. There may be other things going on that are unknown and these other things might really be the cause of the outcome. It is more difficult for these other methods to rule out other possible causes, although the other methods can, again, establish reasonable likelihood. If a cause cannot be established, the evaluation can be used to describe what happened.

For example, the evaluation could say, “After the program, people were better off.” This doesn't necessarily mean it was the program which made the people better off, but the program may be one reasonable cause. Gathering more evidence, from multiple methods, driven by a very clear understanding of the program, can help determine the most reasonable explanation of causes of the outcomes.

Additional Resources about design:

Steps in Program Evaluation
Gathering Credible Evidence
https://www.cdc.gov/eval/steps/gaterhingcredibleevidence.pdf

Program Evaluation: A Variety of Rigorous Methods Can Help Identify Effective Interventions
GAO-10-30 November 23, 2009
https://www.gao.gov/products/GAO-10-30

What Constitutes Credible Evidence in Evaluation and Applied Research? Claremont Graduate University Stauffer Symposium, 2006
https://research.cgu.edu/dbos-events-and-conferences/home/what-constitutes-credible-evidence-in-evaluation-and-applied-research

A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research. Michael Scriven. 2008. Journal of MultiDisciplinary Evaluation, Volume 5, Number 9.
http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/160

A Research Primer: Basic Guidelines for the Novice Researcher. Grace D. Brannan, PhD; Jane Z. Dumsha, PhD; David P. Yens, PhD. The Journal of the American Osteopathic Association, July 2013, Vol. 113, 556-563. https://jaoa.org/article.aspx?articleid=2094672

A Research Primer, Part 2: Guidelines for Developing a Research Project. The Journal of the American Osteopathic Association, January 2014, Vol. 114, 41-51. https://jaoa.org/article.aspx?articleid=2094639

Evidence and Equity: Challenges for Research Design. February 20, 2018. U.S. Department of Health & Human Services.
https://www.acf.hhs.gov/opre/resource/evidence-and-equity-challenges-for-research-design
The purposes of this brief are to Discuss research disparities between easier-to-study populations and harder-to-study, more marginalized groups, and to Present four strategies to address these research disparities

Methods for collecting data

What is Program Evaluation

There are <u>many methods</u>, each with their own uses, advantages and difficulties. Methods include:

Surveys

Analysis of Administrative Data

Key Informant Interviews

Observation

Focus Groups

Evaluations could use <u>any, not necessarily all,</u> of these methods, depending on the question and goal of the evaluation.

In this section, I will describe these methods. This section also includes brief discussions of some of the challenges with these methods.

Additional Resources

Overview of Basic Methods to Collect Information
Carter McNamara, MBA, PhD, Authenticity Consulting, LLC
https://managementhelp.org/businessresearch/methods.htm

Surveys

Surveys are a set of questions that are asked of everyone in the same way.

Surveys can answer question about how many and how often. For example:

- How many clients are satisfied with services?
- How often do people have difficulties using the services?

Typical questions might be like this:

How satisfied are you with the program?

very satisfied	satisfied	neither	dissatisfied	very dissatisfied

How did you hear about the program? Check all that apply.

□ Radio
□ TV
□ friends
□ other ______________________

Surveys might be used to describe the entire client population, if respondents were chosen randomly or systematically (see next page) and if the sample is sufficiently large.

Survey Administration:

Surveys can be cross sectional.

Cross sectional means the surveys are given to many people in the same time frame, for example in the same month or year. This is a snapshot of the population during a single time period.

Cross sectional surveys can be used to compare different groups. For example, men said xyz while women said abc. National opinion polls are typically cross sectional surveys.

It is difficult for cross sectional surveys to establish cause, because they often cannot establish event order. Surveys can ask people about past events, but usually the further back or less important past events are, the less accurate the memory (recall bias).

Surveys can also be longitudinal. These kinds of surveys can follow the same people over time. Longitudinal surveys can study how people change over time and the order of events associated with the changes.

Longitudinal surveys can help to reduce recall bias. Also, because these surveys follow people over time, they can show whether changes came before or after events, and so can help in establishing causality.

Potential problems with longitudinal surveys have to do with how people leave the survey group. People may move, just want to quit, or leave the survey for many other reasons. If the pattern of people leaving is not random, then the remaining sample may be bias. That is, if those who leave are different from everyone else, then those who stay may no longer be a random sample, which could affect the validity of the results.

Survey Administration: Generally, the goal is to administer a survey so the results can describe a larger population. That means the survey should be administered to people using some clearly defined method of selection, so that participants have a known probability of being in the sample. For example:

- Random selection – generate a list of random numbers, assign each person a random number, sort the people by the random number and take the people listed first. They were put on top of the list randomly.

- Systematic selection – a typical method is to start with the 5th person and then select every 7th person after that. The numbers, the 5th and the 7th are also chosen randomly. This should mean that the people are chosen randomly.

- Cluster sampling - Randomly select locations to be in the sample (e.g., randomly select neighborhoods in a city, randomly select schools in a district) and then use systematic sampling on the people in the clusters.

Random or systematic selection means that the group of people you select are more likely to be similar to your clients, in general. You aren't excluding or only including any particular group. You are avoiding bias, in sampling terms.

If you **do** use random or systematic selection, then it is more likely that the sample represents the population and you **can** use the results of your survey to make conclusions about your clients.

If you **do not** use random or systematic selection, your sample is more likely to be biased, or not representative of the population, so you can **NOT** use the results of your survey to make conclusions about your client population. That is, you cannot generalize from your study to your client population. You can only describe the people who took the survey, so for example, you could say “The people who took this survey said ...”

National surveys or polls typically include about 1,000 to 2,000 people. If the surveys follow good, well defined, sampling practices, surveys with these sample sizes can represent the nation.

National Research Center, March2017 https://www.n-r-c.com/survey-sampling-work/
Polling Fundamentals, Roper Center. https://ropercenter.cornell.edu/support/polling-fundamentals/
About Washington Post polls https://www.washingtonpost.com/wp-srv/politics/polls/poll_faq.htm

When using surveys, a critical step is to check whether the surveys were, in fact, administered, appropriately.

A major step in checking data based on a sample is to describe the demographic characteristics of the sample, e.g., age, race/ethnicity, sex, education, and so on. Then, if possible, compare the sample and population characteristics. If the sample was based on random selection, then the sample **should be** similar to the population. If the sample **is** similar to the population, then you are in a better position to say that the results from the sample can be used to generalize to the population.

Similarly, when you have treatment and control groups, compare the demographic characteristics of the treatment and control groups. If people were randomly assigned to either treatment or control groups, then the two groups **should be** similar. If they **are** similar, then it is much more likely that any outcome differences are due to the treatment, because, again, the two groups started off with similar characteristics.

However, even if the sample does appear to be similar to the population using the demographic characteristics available, it is still possible that there are other characteristics, for which data are not available, that show the sample to be different from the population. That is, it may be difficult to make sure, definitively, that the sample is representative of the population. However, the first basic step is always to check whether they are similar using the data that are available.

A second step, for surveys, is to check the response rate. Generally, in most surveys, there are people who do not participate. That is, they were selected to participate in the survey, but for some reason, did not. They might have refused, moved, couldn't be located, been unable to participate, etc. Compare the characteristics of those who participated with those who did not. Again, if the characteristics of those who did and who did not participate are similar, you are in a better position to generalize from the sample results to the population.

In sum, if the compositions of the sample and population are similar, and if the participants and non-participants are similar, then there would be a better chance that the sample is representative, and the results from the sample can be generalized to the population.

Louis M. Rea, Richard A. Parker. Designing and Conducting Survey Research: A Comprehensive Guide. Jossey-Bass. 2014. https://books.google.com/books?hl=en&lr=&id=Ub8BBAAAQBAJ&oi=fnd&pg=PA201&dq=analyzing+survey+data&ots=ixEtw-KixJ&sig=unGo88jsYXZcoIbwiskkt879iFs#v=onepage&q=inferential&f=false

A few notes about **response rate**:

Often, current surveys may have very low response rates, as low as 7% to 9%.

However, “It's important to note that response rates themselves are not a measure of survey quality. So long as the final sample of respondents is representative of the entire population, relatively low response rates do not pose a risk to data integrity. Researchers have demonstrated that telephone surveys with low response rates are still able to represent the entire population accurately.” (Marken, 2018).

Kamel Mellahi and Lloyd C. Harris. Response Rates in Business and Management Research: An Overview of Current Practice and Suggestions for Future Direction. British Journal of Management, Vol. 27, 426–437 (2016)
https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8551.12154
Full paper available here http://wrap.warwick.ac.uk/73828/

Scott Keeter, Nick Hatley, Courtney Kennedy and Arnold Lau. What Low Response Rates Mean for Telephone Surveys. Pew Research Center. 5/15/2017
https://www.pewresearch.org/methods/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/

Stephanie Marken. Methodology Blog: Still Listening: The State of Telephone Surveys, 1/11/2018.
https://news.gallup.com/opinion/methodology/225143/listening-state-telephone-surveys.aspx

Additional Resources about surveys

What researchers mean by... cross-sectional vs. longitudinal studies. Institute for Work and Health. At Work, Issue 81, Summer 2015.
https://www.iwh.on.ca/what-researchers-mean-by/cross-sectional-vs-longitudinal-studies

Introduction to study designs - cross-sectional studies
https://www.healthknowledge.org.uk/e-learning/epidemiology/practitioners/introduction-study-design-css

Evaluating Survey Quality in Today's Complex Environment, June 2016
American Association for Public Opinion Researcher
https://www.aapor.org/Education-Resources/Reports/Evaluating-Survey-Quality.aspx

Sampling. Pew Research Center. https://www.pewresearch.org/methods/u-s-survey-research/sampling/

Anol Bhattacherjee, University of South Florida. Social Science Research: Principles, Methods, and Practices
Chapter 8 is on Sampling. Chapter 9 is on Surveys
https://scholarcommons.usf.edu/oa_textbooks/3/

Five Tips for Designing an Effective Survey. January 22, 2018.
Duke Global Health Institute, https://globalhealth.duke.edu/media/news/five-tips-designing-effective-survey

The Qualtrics Hand Book of Question Design, and other resources,
K-State Survey, https://survey.k-state.edu/help/index.html

Administrative data

Analysis of administrative data is just using statistical analysis on program data that is already collected.

Administrative data has advantages:

- No new data collection is required
- Many databases are relatively large
- Data may be available electronically

and disadvantages:

- Data were gathered for another purpose, so may not have necessary variables.
- In all administrative data sets, some fields are likely to be more accurate than others.

Additional Resources

David J. Hand. Statistical challenges of administrative and transaction data. First published: 09 February 2018. Journal of the Royal Statistical Society.
https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssa.12315

Data collection: Types of data collection – Administrative Data. 2013
Statistics Canada. (now archived, and no longer updated, but still available)
https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch2/types/5214777-eng.htm

Ethical Use of Administrative Data for Research Purposes. Paul G. Stiles, Ph.D., J.D., Roger A. Boothroyd, Ph.D., Department of Mental Health, Law and Policy, University of South Florida
https://www.aisp.upenn.edu/case-studies/ethics/

Focus groups

Focus groups are structured discussions among small groups of people.

Generally, a facilitator leads a group of 8-10 people in a discussion about selected topics with planned questions, while allowing for interesting, new or unplanned follow up questions.

Typical focus group questions are like these:

- What is your overall impression of the program?
- What are the things you like or dislike about the program?
- What have you gained in this program?
- If you have not noticed any changes in yourself, what do you think are the reasons?

Additional Resources about focus groups

Basics of Conducting Focus Groups
Carter McNamara, MBA, PhD, Authenticity Consulting, LLC
https://managementhelp.org/businessresearch/focus-groups.htm

Community Toolbox: Conducting Focus Groups.
https://ctb.ku.edu/en/table-of-contents/assessment/assessing-community-needs-and-resources/conduct-focus-groups/main

Key informant interviews

Key informant interviews are qualitative, in-depth interviews of 15 to 35 people selected for their first-hand knowledge about a topic of interest.

Although key informant interviews are more informal than other forms of data collection, they still require a structure to be effective. Your respondent is more likely to take you seriously (and provide better information) if you are prepared and the conversation has direction.

Begin by introducing your project and purpose. Start with an easy question. Ask your most important questions first

Additional Resources

UCLA Center for Health Policy Research
Section 4: Key Informant Interviews
http://healthpolicy.ucla.edu/programs/health-data/trainings/Documents/tw_cba23.pdf

Community Toolbox, Conducting Interviews.
https://ctb.ku.edu/en/table-of-contents/assessment/assessing-community-needs-and-resources/conduct-interviews/main

6 tips for planning successful key informant interviews
https://upstream.consulting/programs/6-tips-for-planning-successful-key-informant-interviews
and
How to conduct a successful key informant interview
https://upstream.consulting/evaluation/how-to-conduct-a-successful-key-informant-interview
Breyana Davis, July and August 2019, Upstream Consulting.

Key Informant Interviews
University of Illinois Extension
http://ppa.aces.uiuc.edu/KeyInform.htm

Observations

Observation is a method of data collection, involving observing and recording in a specified setting.

Observations can be structured, that is, there are defined things to observe. For example, one set of starting points is to describe the situation by identifying the people present, their titles, roles, relationships to each other, the activities which are going on, describing the physical space, and so on.

From Observation Methods, Elon’s Program for Ethnographic Research and Community Studies (PERCS)
https://www.elon.edu/u/academics/percs/resources/observation-methods/

On the other hand, observations can be unstructured, which usually means recording as much as possible, usually without there being predefined things of interest to observe. Unstructured observation might be useful if the evaluator is going into the situation with no specific and predetermined questions.

Observation need not be entirely one or the other, of course, but could be a combination of structured and unstructured, or start with structured and move onto less structured.

Observation can also vary by level of participation. At the one end is just observing, without becoming involved (non-participant observation). Examples might be watching people at check out lines in supermarkets, watching traffic patterns, or watching behaviors at social events. In this case, the observer just watches.

Evaluators can also conduct participant observation, in which the observer becomes actively involved in the group or situation they are studying. Examples might be joining a group activity, as a member of the group. Someone may do participant observation to get information or understanding that only an insider could have.

The premise underlying participant observation “is that the researcher becomes a more effective observer by taking an active role in the performance of regular activities. In other words, knowledge gained through doing is of a higher quality than what is obtained only through observation. In many cases, involvement with ordinary chores will not only enhance the researcher's understanding of the processes, techniques, and words associated with these activities, but will also result in better rapport with informants.”

Above quote from: Documenting Maritime Folklife: An Introductory Guide, Part 2: How to Document. Participant Observation. American Folklife Center. Library of Congress. http://www.loc.gov/folklife/maritime/twopo.html

Potential issues with observation:

Likely to be very resource intensive

Presence of observer may change behavior of observed.

Observer can’t observe and record **everything**. So they have to select what to record, which means they do **not** record some things. Address this by having multiple observers, observing multiple occasions.

Difficult to know whether the situation/event observed is representative.

There might be questions about whether the observe or observed are biased in reporting or recording or interpreting.

Ethical considerations:

- Do those being observed know they are being observed.
- Whether observed need to give informed consent.
- What if you observe something that those being observed don’t want others to know.
- Trust between observed and observer.
- Keeping observations of people confidential or anonymous

Also see this:
Ethical Practices in Fieldwork
Elon’s Program for Ethnographic Research and Community Studies (PERCS)
https://www.elon.edu/u/academics/percs/resources/ethics/

Observation: Additional Resources

Conducting Observational Research. Associate Prof. Melanie Bryant
https://www.deakin.edu.au/__data/assets/pdf_file/0004/681025/Participant-observation.pdf
https://video.deakin.edu.au/media/t/1_ezwptl1k/36493582

Better Eval, non-Participant observations
https://www.betterevaluation.org/en/rainbow_framework/describe/collect_retrieve_data#info_observation

Observational Researcher. in Open Text Research Methods in Psychology
Carrie Cuttler, Washington State University, 2017
https://opentext.wsu.edu/carriecuttler/chapter/observational-research/

Observation Research: A Methodological Discourse in Communication Research
Esiri, Ajasa, Okido and Edomi. Research on Humanities and Social Sciences www.iiste.org Vol.7, No.20, 2017
https://www.iiste.org/Journals/index.php/RHSS/article/viewFile/39471/40579

Michael Dunn, Research Methods on Psychology, Fall 2015.
Lab 3: Applying Observational Methods - Systematic Observation
https://webcourses.ucf.edu/courses/1140056/assignments/4055510

James Sawusch, Psychology 250, Scientific Inquiry, Fall 2014
http://www.acsu.buffalo.edu/~jsawusch/PSY250/250f14.html

Qualitative Research methods

Focus groups, interviews and observation are qualitative research methods, that is, methods that are less likely to rely on statistical analysis.

Advantages

- Can help explain the “how” and “why”, through gathering greater detail, and more in depth responses from participants.
- Because these methods are less structured and more flexible, can discover information that might be unexpected or not planned for.
- Can be used when little is known about the topic/situation. Surveys, for example, are hard to use when don’t know enough to know what questions to ask.
- Can help to generate further questions, suggestions, recommendations.

Disadvantages

- The evaluator's subjective views can introduce error. Difficult to validate.
- The focus of the evaluator is only on what is observed at one time in one place, and samples are usually small and non-representative, so usually cannot generalize to the client population.
- Information from observations/ interviews/ groups can be time consuming and difficult to interpret.
- Focus groups could be dominated by one individual and their point of view.

Qualitative methods and establishing cause

Typically, it is more difficult to establish causal relationships using qualitative methods for a number of reasons: more often small and non-random samples, and the method is often less “standardized” or controlled than are quantitative methods and so more difficult to isolate specific causes.

However, when various methods (eg., fieldwork, case studies, interviews) are combined for long periods of time, patterns may emerge, which may help establish both validity and reliability – if we see the same behaviors and patterns over time. (Suggested by David Fetterman.) Also, see the discussion about in depth case study described earlier.

Additional Resource

Community Toolbox. Section 15. Qualitative Methods to Assess Community Issues
https://ctb.ku.edu/en/table-of-contents/assessment/assessing-community-needs-and-resources/qualitative-methods/main

Methods: Additional Resources

Simply Psychology, Saul McLeod. Chapter on Research Methods. Updated 2017
https://simplypsychology.org/research-methods.html

e-Source http://www.esourceresearch.org/ From the US NIH. This site says "Inside you will find 20 interactive chapters with authoritative answers to methodological questions on behavioral and social science research." Basically, how they say research should be done.

Anol Bhattacherjee Social Science Research: Principles, Methods, and Practices. 2012 https://scholarcommons.usf.edu/oa_textbooks/3/

Data Analysis

A major part of evaluation is to collect data, described in the methods sections, above. The next step is to analyze the data. This is a very brief introduction to some very basic ideas of data analysis.

Often, data analysis is complex, and best conducted by someone with professional training in statistics. However, understanding some of the basics of statistics is very useful for those involved in conducting evaluations.

For example, it is important for those involved to know what data analysis can do, and what it cannot do. You can know some of the limitations, how it can be useful, how to interpret statistical reports, and whether conclusions in the statistical analyses are appropriate.

Understanding the basic ideas will help you work with the statistician to design the evaluation and the analysis, and to understand what the results mean.

In this section, I will explain a few of the basic ideas and issues of data analysis

Reviewing data

The first step of data analysis is always to review or check the data. Almost all data sets will have some errors, such as missing values, data entry errors, values that seem way out of line with other values, invalid values, duplicate records, misunderstandings between those who collected the data, entered the data and those who analyze the data, and so on. Thus the first step of data analysis is always to identify any errors.

Generally, reviewing data starts with preparing a list all the variables in the data set, and what values the variables can have. For example, people's age can be 0 to less than 120. This list is a *data dictionary*. The data dictionary would also show what values mean. For example, 1=male and 2=female. If the data set is plain text, the dictionary also lists the columns for each variable.

After preparing the data dictionary, researchers read the data set into statistical software, and examine the data, through frequency tables, bar charts, means, and graphs. Doing these basic analyses makes it easier to find errors in the data.

Cross checks or logic checks are useful to make sure that two step questions don't have problems. For example, people shouldn't say they don't do some activity, but also say that they were very satisfied with the activity (that they said they didn't do).

After the errors are found, evaluators can fix some of the errors, **and document** the errors and fixes. Describing the errors in the data set and how they are fixed is a critical step so that, 1) researchers can remember later what they did, and 2) so anyone reviewing the research can see what was done. That is, a simple step to keeping the research transparent.

Example of a data dictionary

Variable	Columns	Valid Values
ID	1-4	0001 to 5000
Age	5-7	000 - 120 888 = Refused 999 = Unknown
Degree	8	1 = Less than high school degree 2 = High school degree but not more 4 = Technical / Vocational degree but not more 5 = Bachelors degree but not more 6 = Post bachelors degree 8 = Refused 9 = Don’t know
Drink Coffee	9	0 = None 1 = Less than 5 cups a week. 2 = 5 to 10 cups a week 3 = More than 10 cups a week 8 = Refused 9 = Don’t know

Example of identifying invalid data using a frequency table

Variable: Drink Coffee

0 – None	581
1 - Less than 5 cups a week.	1870
2 - 5 to 10 cups a week	1658
3 - More than 10 cups a week	783
8 – Refused	12
9 – Don’t know	89

According to the data dictionary on the previous page, the variable <u>Drink Coffee</u> should only have values 0, 1, 2, 3 or 8 (Refused) or 9 (Don’t Know). Values of “4” are invalid values, probably data recording errors. Decisions need to be made on what to do about these.

Example of identifying invalid data using a cross tab

Variables Age, Degree

	Age			
Degree	Less than 15	15 to 20	21 to30	31 or more
Less than HS	170	87	10	1
HS graduate	2	642	890	193
College graduate	0	12	1086	1102
Post college degree	3	7	275	520

While it is certainly possible that a few kids younger than 15 might earn high school degrees, it seems unlikely, though possible, that many kids under age 15 to have earned a post college degree. Review these records to identify possible errors.

Additional Resources:

National Emergency Medical Service for Children Data Analysis Resource Center https://www.nedarc.org/tutorials/analyzingData/index.html This resource has a couple of chapters on validating and cleaning data sets.

World Bank, Data cleaning. https://dimewiki.worldbank.org/wiki/Data_Cleaning

Data Quality https://sites.google.com/site/gsocialchange/dataquality

Statistics Canada Quality Guidelines, Guidelines for ensuring data quality. December 2019. https://www150.statcan.gc.ca/n1/pub/12-539-x/2019001/ensuring-assurer-eng.htm

The Ultimate Guide to Data Cleaning, When the data is spewing garbage. Omar Elgabry, 2/28/2019 https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

Descriptive analysis

Generally, there are two kinds of analysis: descriptive and inferential.

Descriptive is just summarizing what you have observed. For example if you surveyed 100 people about their opinions on a recent movie, a summary could be: X% of the respondents liked it, Y% of respondents did not. Of women in the survey, XW% liked it. Of men in the survey, XM% liked it. This is a simple summary of the data you have collected. The analysis **only** describes people for whom there are data. Descriptive statistics is **not** used to infer anything about a larger population.

If there is only one variable (for example, education), commonly used statistics include measures of frequency (e.g., x% with only some college, y% with only college degrees, z% with degrees beyond bachelors), indicators of central tendency (mean, median), and indicators of variability, spread or dispersion (range with high and low, variance, standard deviation).

When there are two or more variables (for example, education and age), we use measures of association, like correlation (for numerical variables) or cross tabulation (for categorical variables). These show how closely related the two variables are. Given one happens, how often does the other to happen. This is only a measure of association, NOT a measure of causation.

Sources

Pat Bartlein, class material for Winter 2018 Geographic Data Analysis. Department of Geography. University of Oregon, https://pjbartlein.github.io/GeogDataAnalysis/

Revealing Patterns Using Descriptive Statistics, The Writing Studio, Colorado State University, 2018. https://writing.colostate.edu/guides/page.cfm?pageid=1394&guideid=67

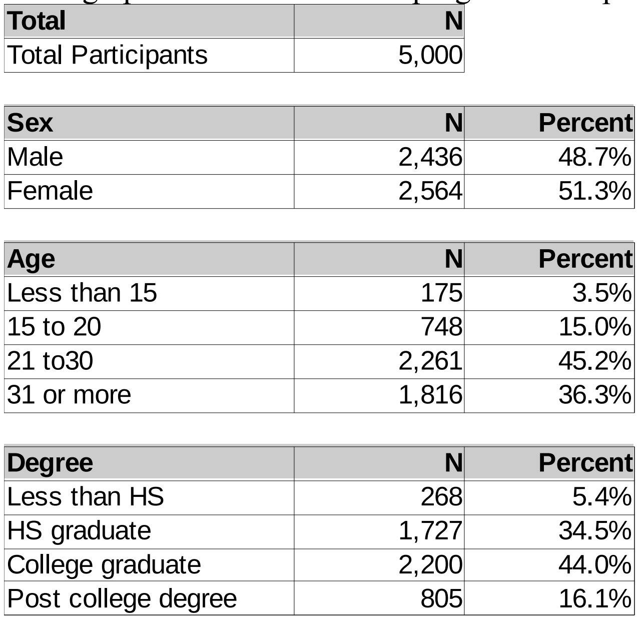
April Klazema, May 2014.Descriptive and Inferential Statistics: How to Analyze Your Data https://blog.udemy.com/descriptive-and-inferential-statistics/

Statistics dictionary, from Harvey Berman, Stat Trek. https://www.stattrek.com/statistics/dictionary.aspx

Generally, descriptive analysis should start with a description of the subjects being studied. Describe who the subjects are and how you got them.

If your subjects are people, a table or chart of demographics would be appropriate. For example, how many and what percent are male or female, various races or ethnicities, what are their ages, where do they live, and so on. Table 1 is a simple example.

Table 1
Demographic characteristics of program XYZ participants



Total	**N**
Total Participants	5,000

Sex	**N**	**Percent**
Male	2,436	48.7%
Female	2,564	51.3%

Age	**N**	**Percent**
Less than 15	175	3.5%
15 to 20	748	15.0%
21 to30	2,261	45.2%
31 or more	1,816	36.3%

Degree	**N**	**Percent**
Less than HS	268	5.4%
HS graduate	1,727	34.5%
College graduate	2,200	44.0%
Post college degree	805	16.1%

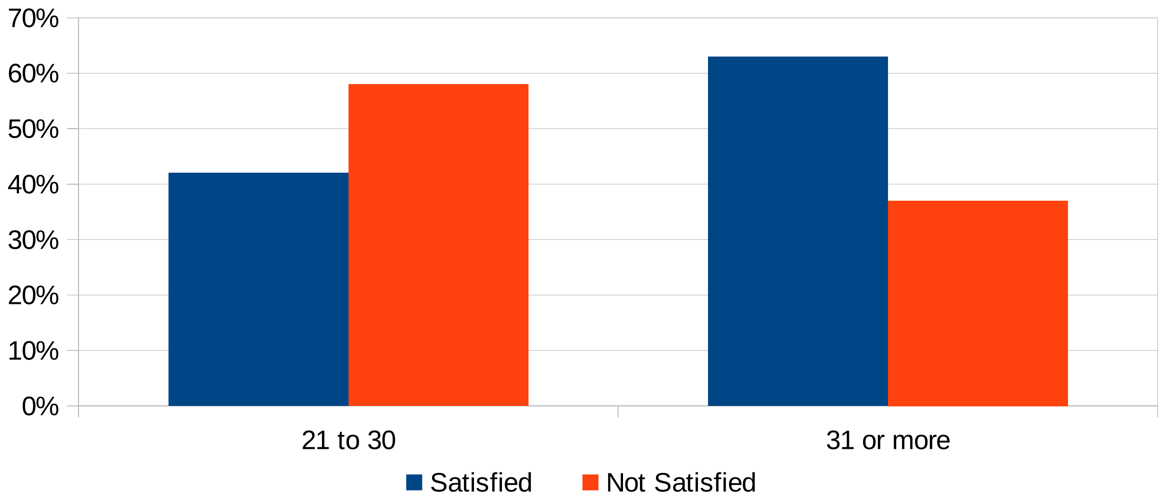
Source: Survey of participants in program XYZ, September 2019

Evaluation reports can also present other data besides basic demographics. For example, people might be asked if they are satisfied with a program in which they participate.

Results can be presented in a table or chart, or some combination. Results might be simple frequencies, but it is often more interesting if results are presented by subgroup.

For example, the chart below shows that, of people asked, among people age 31 or more, a higher percent were satisfied. On the other hand, among people aged 21 to 30, a higher percent were not satisfied.

Chart 1
Percent of Participants Satisfied / Not Satisfied with Program
By Age



Source: Survey of participants in program XYZ, September 2019

Guides for presenting descriptive data are fairly straightforward. Generally, each table or graph should have all the information needed to interpret it, so it can be used on it's own. At a minimum, each table or graph should have:

- An identification number, e.g., Table 1, Figure 1, etc.
- A descriptive title.
- Column and row labels
- Definitions of any symbols or abbreviations used in the presentation.
- Source of data.

Sources

Kids Zone, Create a Graph, US National Center for Educational Statistics, https://nces.ed.gov/nceskids/createagraph/
For example see Building Bar Graphs, https://nces.ed.gov/nceskids/help/user_guide/graph/bar.asp

Designing effective tables, graphs and maps, UK Government Statistical Service, first published February 2017
https://gss.civilservice.gov.uk/policy-store/effective-graphs-and-tables-in-official-statistics/

Charts DOs and DON'Ts. European Environment Agency. Last updated 21 January 2019
https://www.eea.europa.eu/data-and-maps/daviz/learn-more/chart-dos-and-donts

Inferential Analysis

Inferential statistics is used when you cannot or do not collect data from all of the people in the relevant population. It may not be possible to collect data from everyone who is in the program, everyone who gets “treatment”, everyone in the population, and so on. In these cases, you collect data from a sample to try to draw conclusions about a larger population.

The basic idea is to use a sampling method to select certain people to represent the population. If the study and sampling design, and data collection methods are appropriate, we make certain assumptions about the sample, and then conduct statistical tests on the sample, and draw conclusions about the entire population.

Did an evaluation of your program show that people for whom you collected data benefited from the program’s intervention, was the intervention the cause, do the results of your study show that everyone in your program will also benefit from the intervention? Those are the questions for inferential statistics.

It is necessary to mention that statistical testing is enormously complex, involving many assumptions about the data, how it was collected, how it is distributed, and many decisions about what statistical methods to use, and how to interpret the results of statistical testing. This guide is an introduction to some of the basic ideas, but it is usually best to involve a professional statistician.

Inferential statistics is often used in studies which compares people who got some kind of treatment to those who didn't, to determine whether the "treatment" made a difference in the population. If we had data for the entire population, we could say conclusively, for example, that people who got treatment X had a different outcome than did people who did not get treatment X. However, again, we usually do not have data for the entire population, so we use a sample. If the sampling and study design were appropriate then the sample will be representative of the larger population. If so, then we can use results from the sample to make conclusions about the population. For example, if we find a treatment "works" for the sample, then we could conclude it is likely that the treatment would also "work" for people beyond those in the study.

For example, one experiment randomly assigned breast cancer (BC) survivors to either a mindfulness based stress reduction group (treatment), or to a control group (no treatment). The study reported that, after mindfulness training, the treatment group reported lower anxiety, and lower stress, compared to the control group. The researchers concluded that their study provided evidence to support mindfulness based stress reduction "as a treatment with sustained effects" associated with breast cancer treatment. That is, the authors concluded that their treatment made a difference in their study, and that their sample was representative of the larger population of people with breast cancer, so that the effects found in their study would apply to people beyond those included in the study, that is, to all breast cancer survivors.

Actually, the study was rather more complex than just the above. The participants were stratified "by type of surgery (lumpectomy v mastectomy), BC treatment (chemotherapy with or without radiotherapy v radiotherapy alone), and BC stage (stage 0 to I v II to III)." In addition, "The sample size was calculated to compare adjusted mean outcome scores separately at 6 and 12 weeks of follow-up between the two groups" (treatment and control), and the design "Allowed for 10% loss to follow-up of the sample". The analysis was: "Linear mixed models were implemented to assess the interaction between participant assignment (intervention vs control) and time (baseline, 6 weeks, and 12 weeks) in relation to symptom outcomes, testing whether the rate of symptom change varied by study assignment."

As described above, statistical analysis can be complex and it is very useful to have a professional statistician to help in developing a plan to collect the data, and to help plan and conduct the analysis.

Examination of Broad Symptom Improvement Resulting From Mindfulness-Based Stress Reduction in Breast Cancer Survivors: A Randomized Controlled Trial. Lengacher, Reich, et al. Journal of Clinical Oncology, August 20, 2016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5012660/

Inferential statistics are also used in survey or polls. For example, one survey used random digit dialing to survey about 1,000 adults during July 1-8 2018. If the sampling design and survey administration were appropriate, it is assumed that the responses from people in the sample could be used to make conclusions about the entire population (e.g., the adult US population). Thus, the researchers used this data to infer, for example, that “59 percent of Americans support a ban on smoking in public places” and that support for the ban “is relatively steady across all demographics, with slightly higher numbers among women, older adults, college graduates, and those whose political views are moderate or liberal.” Without inferential statistics, they would only be able to describe how people in the sample responded, and would not be able to use the data to describe the population.

Widespread Support for Public Smoking Ban, Finds Gallup Poll, Diane Domina, July 26,2018. Health Central.
https://www.healthcentral.com/article/widespread-support-for-public-smoking-ban-finds-gallup-poll

Statistical testing:

Generally, in inferential statistics, people use statistical tests to examine the relationship among two or more groups. For example, in experiments, this might be looking for significant differences between the “treatment” and “control” groups, or between different types of treatments. In surveys, it could be to find whether there are differences among various demographic groups (e.g, men vs women, White vs Black vs Hispanic, etc.)

There are a very large number of tests. Some you may have heard about include the t-test, analysis of variance, or regression. These tests are called “parametric”, which means the data meets certain assumptions. Other kinds of statistical tests, called “non-parametric” include the Mann-Whitney and the Kruskal-Wallis tests. These tests generally have fewer assumptions, but are less powerful, that is, may be less likely to detect a difference when it really exists.

There are very many different kinds of tests, and usually it is best to consult a professional statistician to know which to use.

Sources:

Choosing a Statistical Test. Salvatore Magniafico. 2016. Summary and Analysis of Extension Program Evaluation in R, version 1.18.1
http://rcompanion.org/handbook/D_03.html

How to choose the right statistical test? Barun K Nayak and Avijit Hazra. Indian J Ophthalmol. 2011 Mar-Apr; 59(2): 85–86.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116565/

Choosing a statistical test. John H. McDonald. Handbook of Biological Statistics.
http://www.biostathandbook.com/testchoice.html

Selection of Appropriate Statistical Methods for Data Analysis. Prabhaker Mishra, Chandra Mani Pandey, Uttam Singh, Amit Keshri, and Mayilvaganan Sabaretnam. Ann Card Anaesth. 2019 Jul-Sep; 22(3): 297–301.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6639881/

The basic ideas of statistical testing.

Often, a conclusion drawn from statistical analysis is whether the results were “significant”. For example, was there a difference between the intervention or treatment group and the control group, or was there no difference?

The breast cancer study described earlier concluded “Breast Cancer subjects randomly assigned to mindfulness based stress reduction demonstrated greater symptom improvement in fatigue (severity and interference; $P < .01$).” and also “Participants assigned to (treatment) tended to experience greater improvement (in psychological symptoms) than those assigned to (control); however, this trend did not reach statistical significance ($P = .06$).”

However:

Statistical analysis is a rather difficult concept (and I am not sure I understand it very well). Greenland (2019) writes "P-values are ... difficult to understand properly". and so I would strongly recommend working with a professionally trained statistician. That person could help understand what conclusions could be drawn from the results.

There are a number of issues to consider.

First, basically, statistical inference proposes a model about the data, and then tests whether the data fit the model. Part of the model is about differences between an intervention and control group. However, the model includes much more: "Every method of statistical inference depends on a complex web of assumptions about how data were collected and analyzed" (Greenland et al, 2016). There are assumptions about sampling, about what the population data actually look like, assumptions about which particular statistical analysis to use and the appropriateness of that method, and so on. As Greenland et al (2016) point out, "the P value tests all of the assumptions about how the data were generated (the entire model)". So it would not be correct to conclude that, if p is some specific value, then the intervention group and control group do not have the same outcome. The more appropriate conclusion would be about whether the proposed model, including all of it's assumptions, fit or did not fit, the data.

Sources:

Sander Greenland (2019) Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values, The American Statistician, 73:sup1, 106-114, DOI: 10.1080/00031305.2018.1529625 https://doi.org/10.1080/00031305.2018.1529625

Greenland, S., Senn, S.J., Rothman, K.J. Carlin, J.B., Poole, C., Goodman, S.N. and Altman, D.G. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol (2016) 31: 337.
https://link.springer.com/article/10.1007/s10654-016-0149-3
Additional references on Statistics.

Second, p values are often considered dichotomous, that is the study results shows or does not show statistical significance.

On the other hand, Amrheim, Greenland and McShane (2019) wrote, “we are calling for a stop to the use of P values in the conventional, dichotomous way — to decide whether a result refutes or supports a scientific hypothesis … we are not advocating a ban on P values, confidence intervals or other statistical measures — only that we should not treat them categorically. This includes dichotomization as statistically significant or not. … One practical way to do so is to rename confidence intervals as ‘compatibility intervals’ and interpret them in a way that avoids overconfidence. Specifically, we recommend that authors describe the practical implications of all values inside the interval, especially the observed effect (or point estimate) and the limits. In doing so, they should remember that all the values between the interval’s limits are reasonably compatible with the data, given the statistical assumptions used to compute the interval”

Exactly how to treat p values may be not entirely clear, partly because, apparently, there is not yet consensus within the statistical community (Wasserstein, Schirm and Lazar, 2019). But Wasserstein, Schirm and Lazar (2019) offer some guidelines, worth reviewing, which they summarize as “Accept uncertainty. Be thoughtful, open, and modest”.

Sources:

Valentin Amrhein, Sander Greenland & Blake McShane. Scientists rise up against statistical significance.
Nature. 20 March, 2019 https://www.nature.com/articles/d41586-019-00857-9

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar. Editorial: Moving to a World Beyond “p < 0.05”
The American Statistician. Volume 73, 2019 - Issue sup1: Statistical Inference in the 21st Century: A World
https://amstat.tandfonline.com/doi/full/10.1080/00031305.2019.1583913#.XUYchhhofIU

Additional sources:

Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108
https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#.XSUpdO1ofIX

Todd A. Kuffner ORCID Icon &Stephen G. Walker. Why are p-Values Controversial? The American Statistician. Volume 73, 2019 - Issue 1. https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1277161#.XemcpD1ofIU

Successful evaluations

One way to increase likelihood of successful evaluation is by using a collaborative, participatory or empowerment approach.

These approaches increasingly involve stakeholders in decisions about the evaluation: how it is going to be conducted, what questions toask, and how the results will be interpreted and used.

Collaborative evaluation: where the evaluator is in charge of the evaluation but they create an ongoing engagement between evaluators and stakeholders. (description by David Fetterman.)

Empowerment evaluation: the program staff members, program participants, and community members are in control of the evaluation and the evaluator serves as a coach and critical friend in the evaluation. (description by David Fetterman.)

Involving stakeholders may:

- reduce suspicion
- increase commitment
- broaden knowledge of evaluation team
- increase the possibility that results will be used.

Additional Resources:

Introduction to Program Evaluation for Public Health Programs: A Self-Study Guide. Step 1: Engage Stakeholders.
https://www.cdc.gov/eval/guide/step1/

Better Evaluation. Understand and engage stakeholders
https://www.betterevaluation.org/en/rainbow_framework/manage/understand_engage_stakeholders

Participatory and Empowerment Evaluations
https://www.betterevaluation.org/en/plan/approach/participatory_evaluation
https://www.betterevaluation.org/en/plan/approach/empowerment_evaluation

Another way is to plan the evaluation from the start of the program or intervention.

“Having a plan helps ensure that future evaluations are feasible and instructive.” (Martin, NIJ Journal, September 2015)

So for example, knowing the questions to be asked can help the program be designed to gather the information needed.

Having a plan in writing also makes the evaluation transparent.

Additional Resources:

Plan for Program Evaluation from the Start
by Alison Brooks Martin. NIJ Journal, #275. September 2015
https://nij.ojp.gov/topics/articles/plan-program-evaluation-start

Putting it all together

In sum, planing a program evaluation includes answering these three key points:

1. What, exactly, is the question?
2. How will you get the information to answer the question?
3. What will you do with the answers to the question?

What else does an evaluator need to know.

- **Evaluation Standards**:

Joint Committee on Standards for Educational Evaluation standards (JCSEE)
https://jcsee.org/program/
Standards about utility, feasibility, fairness and legality, accuracy, and accountability.

Canadian Evaluation Society and CDC adopted the JCSEE standards. 2012
https://evaluationcanada.ca/program-evaluation-standards
https://www.cdc.gov/eval/standards/index.htm

2018 Updated Guiding Principals for Evaluators, American Evaluation Association http://www.eval.org/p/cm/ld/fid=51
Principals about technical methods, competence, integrity/honesty, respect for people, responsibility for general welfare.

United National Evaluation Group, Norms and Standards for Evaluation. 2017. http://www.uneval.org/document/detail/1914

- **Learning evaluation methods**

Links to free online classes
https://sites.google.com/site/gsocialchange/learning
(full disclosure, this is my site.)

Examples:

Northwest Center for Public Health Practice   http://www.nwcphp.org/training

Action Research and Action Learning   http://www.aral.com.au/

EvalPartners e-Learning programme in Development Evaluation   https://elearning.evalpartners.org/elearning

The Global Health Learning Center section on monitoring and evaluation https://www.globalhealthlearning.org/program/monitoring-and-evaluation

Edx   current and archived classes, statistics   https://www.edx.org/courses?search_query=statistics

- **Statistics**

Online statistics books
https://sites.google.com/site/gsocialchange/statbooks

Examples:

CADDIS Volume 4: Data Analysis
https://www.epa.gov/caddis-vol4

Gustman, Burt. Stat Primer. Last revised 2016 http://www.sjsu.edu/faculty/gerstman/StatPrimer/

Lane, David. HyperStat. 2013 http://davidmlane.com/hyperstat/index.html

Stark, Philip B. SticiGui: Statistics Tools for Internet and Classroom Instruction. Last modified 2019.
https://www.stat.berkeley.edu/~stark/SticiGui/

Stockburger, David. Introductory Statistics. 2016 http://dwstockburger.com/Introbook/sbk.htm

UCLA. Probability and statistics Ebook. 2014 http://wiki.stat.ucla.edu/socr/index.php/EBook

- **Journals**, keeping up with methods, research

Online journals about evaluation methods
https://sites.google.com/site/gsocialchange/journals

Examples:

African Evaluation Journal https://aejonline.org/index.php/aej/index

Journal of Methods and Measurement in the Social Sciences (JMM) https://journals.uair.arizona.edu/index.php/jmmss/index

Journal of Official Statistics https://content.sciendo.com/view/journals/jos/jos-overview.xml

Journal of MultiDisciplinary Evaluation http://journals.sfu.ca/jmde/index.php/jmde_1/index

Methods, data, analyses https://mda.gesis.org/index.php/mda

Qualitative Social Research http://www.qualitative-research.net/index.php/fqs/index

- **Presenting statistical data**

Links to sites about presenting data
https://sites.google.com/site/gsocialchange/presenting

Examples:

Making Data Meaningful
http://www.unece.org/stats/documents/writing/

Gallery of Data Visualization, best and worst
http://www.datavis.ca/gallery/

Rougier NP, Droettboom M, Bourne PE (2014) Ten Simple Rules for Better Figures. PLoS Comput Biol
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833

Presenting Statistical Data
https://www.qgso.qld.gov.au/about-statistics/presenting-data

Presenting Numerical Data
https://www2.le.ac.uk/offices/ld/resources/numerical-data/numerical-data

- **Free-to-use Statistical Software**

Links to sites with free to use software
https://sites.google.com/site/gsocialchange/statsoft

http://statpages.info/javasta2.html

http://www.bettycjung.net/Statpgms.htm

Examples:

R https://cran.r-project.org

Epi Info https://www.cdc.gov/epiinfo/index.html

jamovi https://www.jamovi.org

JASP https://jasp-stats.org/

PSPP http://www.gnu.org/software/pspp/

SAS University Edition https://www.sas.com/en_us/software/university-edition.html

Additional Resources

Books/Guides

The Program Manager's Guide to Evaluation, 2010. Office of Planning, Research and Evaluation. Administration for Children and Families. US Department of Health and Human Services
https://www.acf.hhs.gov/opre/research/project/the-program-managers-guide-to-evaluation

Introduction to program evaluation for public health programs: A self-study guide. CDC. https://www.cdc.gov/eval/guide/index.htm

Making evaluations matter: a practical guide for evaluators
Kusters, C.S.L.; Vugt, S.M. van; Wigboldus, S.A.; Williams, B.; Woodhill, A.J. Wageningen University. 2011
https://library.wur.nl/WebQuery/wurpubs/405451

Community Toolbox, Chapters 36-39, Evaluating Community Programs and Initiatives. http://ctb.ku.edu/en/table-of-contents

What makes a good evaluation. Devina Buchkshee. 6/12/2017. 4th Wheel Social Impact.
http://the4thwheel.com/blog/2017/12/06/what-makes-a-good-evaluation/

A Complete Guide to Quantitative Research Method, by Manu Bhatia, June 11, 2018.
https://humansofdata.atlan.com/2018/06/quantitative-research-methods/

Evaluation Building Blocks – A Guide
http://kinnect.co.nz/evaluation-building-blocks-a-guide/

Good Practice Guidelines https://www.evaluation.org.uk/professional-development/good-practice-guideline/

Organizations/Sites about improving evaluation

EvalPartners
https://www.evalpartners.org
foster knowledge sharing and networking among M&E practitioners worldwide

Genuine Evaluation.
http://genuineevaluation.com
Patricia J Rogers and E Jane Davidson blog about real, genuine, authentic, practical evaluation

Better Evaluation
https://www.betterevaluation.org
An international collaboration to improve evaluation practice and theory by sharing and generating information about options (methods or processes) and approaches.

Links to more resources

Betty C. Jung's Evaluation Resources on the Internet
http://www.bettycjung.net/Evaluation.htm

Free Resources on the Web
https://sites.google.com/site/gsocialchange/

Lars Balzer's Evaluation Portal
http://www.evaluation.lars-balzer.name/links/

American Evaluation Association list of online resources
http://www.eval.org/p/cm/ld/fid=53

I got my Ph.D. in sociology from the University at Albany, State University of New York. See more info about me here
https://sites.google.com/site/gsocialchange/thepeople

Some usual legal disclaimers:

This guide can be freely distributed without need for permission, provided it is distributed as is. Distribution for any commercial purpose is strictly forbidden. This guide cannot be sold under any circumstances. I absolutely do NOT give permission for any of this to be quoted without attribution on wikipedia.

This guide is only for education purposes. It does not represent any guidelines, recommendations or requirements about how to do program evaluation. The only purpose is to provide the general public, consumers, students, and evaluators with information about things that may go into evaluations, so that evaluation may be better understood, and evaluators and clients might work better together to get more out of their evaluation.

In my work on this guide, I do not represent or speak for any organization. I prepared this on my own time, at home, and was not supported by any organization.

I also benefited greatly from feedback from folks on various email lists.

Materials on web sites listed in this guide do not necessarily reflect my opinions, nor do I assume any responsibility for the content provided at these web sites. This guide only lists web sites with legal content. Listing a website is not necessarily endorsement of any services or organization. The sites are only listed because they have some freely available information. I also do not have any financial relationships with any site or organization listed on this guide.

This guide does not contain or promote; pornography, hatred, racism, propaganda, warez, nudity, hacking activities, violence, degradation, harm or slander. To the best of my knowledge, I do not list any website that contains or promotes any of that either.