Process: Survey Analysis Workflow

Topics: Analysis workflow from research question to interpretation of results

This video summarizes an analysis workflow. [On the last page of this handout,] the steps and outputs are marked in the center of the diagram, and software tools are marked on the left.

1. Research Question

Research Question
- What factors are associated with [outcome]?
- Does [X] effect [outcome]?

Impactful survey analysis starts with a good research question. A good research question is focused and answerable. Some people use the F.I.N.E.R criteria to evaluate a good research question.

A good research question is:
- **F**easibility
- **I**nteresting
- **N**ovel
- **E**thical
- **R**elevant

F. Can the question be feasibly answered with the available sample size, time, resources, and statistical skills of the team?

I. Is the question interesting to policy-makers, funders, program implementers, and most importantly, you?

N. Is the research question novel; does it extend, confirm, or refute previous findings?

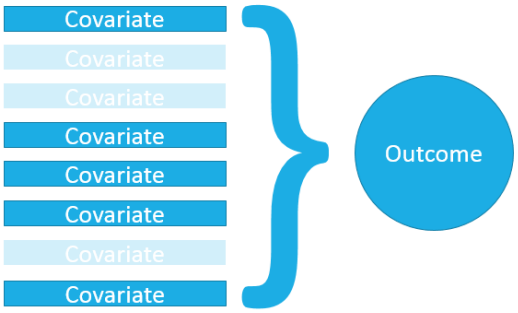
E. Is the study design ethical? Does it ensure minimal risk to participants, as well as informed consent and confidentiality?

R. Is the research question relevant to policy, program management, patient care, or scientific knowledge? If no one cares about the results of the study, it is probably a poor use of your limited time and resources.

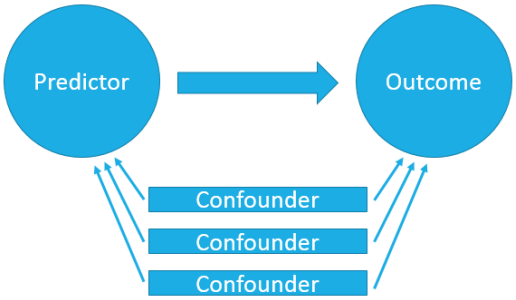
It is important to spend extra time developing a good research question in conjunction with decision-makers, and after becoming familiar with the scientific literature, because a good research question is the foundation of an impactful study.

There are many formats that a quantitative research question can take. Here are two common formats:

General explanatory: What factors are associated with [outcome]?



Hypothesis test: Does [X] effect [outcome]?



Conceptual Framework

1a. General explanatory research question

A general explanatory research question has the format: "What factors are associated with [outcome]?" For example, what social and demographic characteristics are associated with intimate partner violence in Rwanda? The researcher develops a model with multiple factors that are potentially associated with the outcome, and narrows down which key factors are most important based on the magnitude, direction, and statistical significance of their associations.

1b. Hypothesis test research question

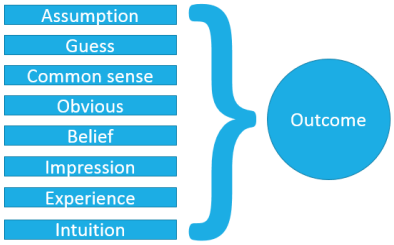
Another common form of research question is a hypothesis test. Hypothesis test questions are formatted: "Does [X] effect [outcome]?" For example, does having comprehensive HIV knowledge reduce risk of HIV infection? To answer this type of question, the researcher tests the effect of one "predictor" on one outcome. (The term "predictor" is used loosely here. To predict something implies causation, and cross-section survey data cannot be used to determine causation of health or social outcomes. But a hypothesis test question *does* provide evidence that a factor is or is not associated with an outcome, after controlling for other potential confounding factors.)

Confounders will be discussed further in a multivariable regression video, but in short, confounders are factors that are associated with both the predictor and the outcome which make the two variables appear to be statistically correlated, even when they are not. If we do not control for confounders by including them in the model, then we might incorrectly estimate the magnitude, direction, and statistical significance of the relationship between the two variables of interest, and ultimately arrive at a wrong answer to our research question.

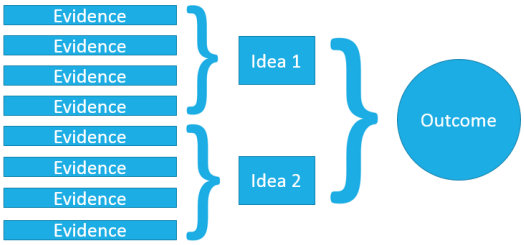
2. Conceptual framework

Once the questions and associated research objectives are clearly stated, the researcher should outline her conceptual framework. A conceptual framework can be stated in words, but I will encourage you to draw a picture of the factors (or

Start:



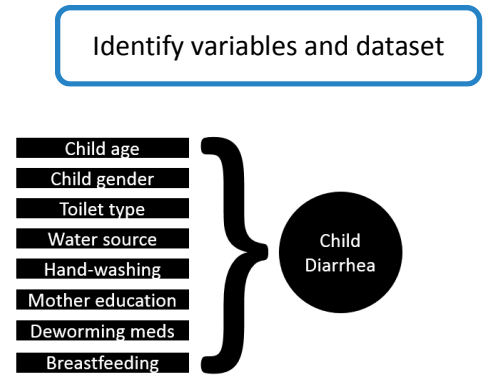
Update:



variables) involved with the research question, with arrows showing how they are related. The conceptual framework can be simple – as simple as the diagrams that we just saw. The video about developing a conceptual framework will provide a few more templates to help you outline a conceptual framework.

Deciding which variables to include in the conceptual framework requires background research. I usually start by documenting my own assumptions and knowledge about the research question, then update the layout and variables in the conceptual framework based on a systematic literature review.

3. Identify variables and datasets



Development of the conceptual framework helps you to identify which variables you need to answer the research question, and to decide which survey dataset(s) you will use for the analysis. Household survey datasets tend to be very large with 1000s of variables, so it is helpful to document the key variables at this early stage before you open the dataset and get severely side tracked. Without a conceptual framework to focus you, you *will* get side tracked!

4. Decide on a dataset

There are hundreds of existing household survey datasets waiting to be analyzed. Many of these surveys are conducted on a regular basis across multiple countries including the Demographic and Health Surveys supported by USAID, the Multiple Indicator Cluster Surveys supported by UNICEF, and the STEPS survey on chronic disease supported by the World Health Organization, to name a few. Once you know which variables you need (and the time frame and geographic coverage of your research question), it will be straight forward to identify which existing survey dataset is best suited to your research.

Create the dataprep.do

5. Data preparation

It is only at this stage that we open the dataset in Stata. The next phase is to prepare the analysis dataset. This involves identifying variables from the survey dataset; summarizing those variables to understand their definitions and distributions; and renaming, re-categorizing, and combining variables as needed for the analysis.

Descriptive analysis_v1.do

6. Descriptive analysis

After the analysis dataset is prepared, we describe the study population. Results of the descriptive analysis are usually summarized in Table 1 of a manuscript. Table 1 might simply provide summary statistics for the overall population, though in some manuscripts, Table 1 describes the characteristics of two or more sub-groups, for example, an intervention group and a comparison group.

Bivariate analysis_v2.do

7. Bivariate analysis

In the bivariate analysis, we look at relationships between each of the covariates and the outcome using statistical tests. This step can serve as a filter, helping us to decide which variables are even worth testing in a multivariable model. We also use bivariate statistics to make comparisons; for example, to test if an intervention group and a comparison group had a similar distribution of age and gender, which provides evidence about whether we can fairly compare the two groups. In a manuscript, we usually report bivariate statistics in Table 2.

Collinearity analysis_v3.do

8. Collinearity

Collinearity occurs when two covariates in a multivariable model are highly related; usually this is because the two variables represent the same *thing*. Basically, one of the variables does a good job explaining variance in the outcome, leaving little left over variance to be explained by the other covariate. As a result, the model becomes unstable. To produce parsimonious (efficient) multivariable models, and to prevent strange, unstable results, we test for strong associations among covariates and remove any collinear covariates from the analysis.

Regression analysis_v4.do

9. Multivariable regression modeling

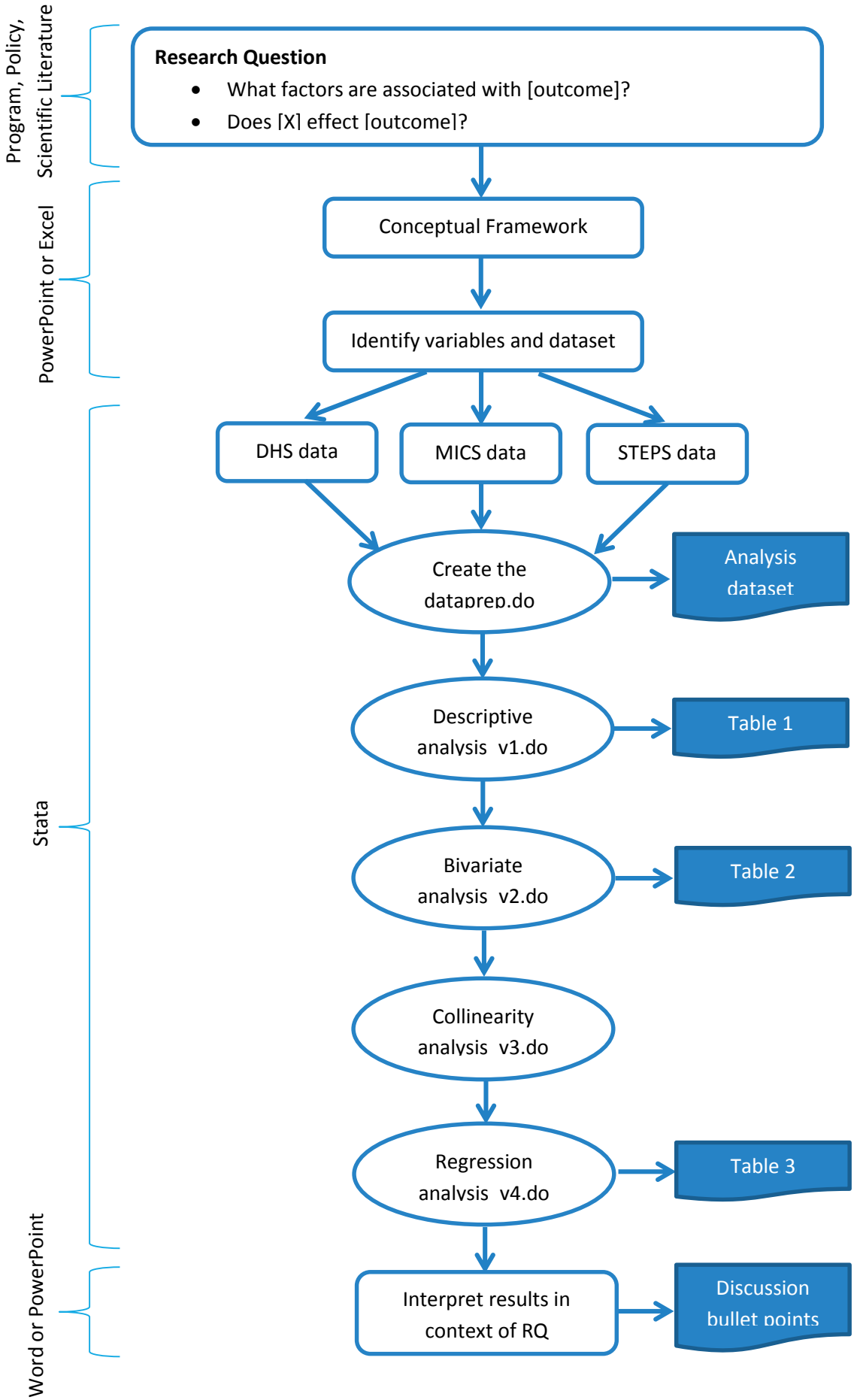
By the time we start building a multivariable model, we have a narrowed list of meaningful covariates that are not collinear. We fit a model with all of these remaining covariates and use manual backward stepwise regression to narrow down which covariates are most important toward explaining the variance in the outcome.

Interpret results in context of RQ

10. Interpreting regression results

Producing a multivariable regression model is somewhat mechanical; we follow a standard process like a recipe to arrive at a final model. How we interpret these final results, and relate them to the real world varies widely. In this part of the analysis you perform statistical interpretation of the model output in terms of the magnitude, direction, and statistical significance of the results. But you should also extend the interpretation to think about the "story" that the data "tells". This "story" is usually comprised of a few key take away messages that are discuss at the end of the manuscript.

This analysis workflow will lead to robust statistical analysis of any existing survey dataset, and will serve as a roadmap for courses taught with populationsurveyanalysis.com materials.


Program, Policy, Scientific Literature
PowerPoint or Excel
Stata
Word or PowerPoint
Research Question
What factors are associated with [outcome]?
Does [X] effect [outcome]?
Conceptual Framework
Identify variables and dataset
DHS data
MICS data
STEPS data
Create the dataprep.do
Analysis dataset
Descriptive analysis v1.do
Table 1
Bivariate analysis v2.do
Table 2
Collinearity analysis v3.do
Regression analysis v4.do
Table 3
Interpret results in context of RQ
Discussion bullet points