

UK Data Service Workshop: From Research Question to Exploratory Data Analysis

Session 1: Clarifying Research Questions-Glossary of Terms

Research Question

The scientific research process always starts by identifying a research question. A research question asks about something we don't know yet or about unresolved problems, e.g. paradoxes, discrepancies in previous research or in observed facts/events. A good research question is relevant: addressing the question is a contribution to the accumulation of knowledge. The research question is a question, not just a topic of research! A good research question is analytical and testable: it can be addressed following the pre-established procedures of scientific research. A good research question is focussed: not too broad (but not too narrow!)

Theory

Set of propositions about how different concepts relate to each other. High level of abstraction: concepts are theoretical and represent an idea, property, characteristic or phenomenon. E.g. power, democracy, political participation, well-being, equality, etc. Derived from, and then tested in, empirical reality and observable patterns. From which empirical forecasts can be derived in turn creating generalisation.

Hypotheses

Theories are transformed into hypotheses to be empirically tested. Hypotheses are more specific, less general, than theory. A hypothesis is a proposition that suggests the existence of a relationship between two concepts. It's a temporary statement that needs to be tested. Hypotheses are explicit (less vague than a research question). Hypotheses are testable: we can measure the concepts and analyse how they relate to each other.

Unit of Analysis

First step in the process of measurement: we need to determine the unit of analysis. Concepts will be applied to social objects. The unit of analysis is the property/nature of this object. We generally find five types of units of analysis: (1) Individuals, (2) Aggregates of individuals, (3) territorial units, Institutions, organisations, entities, (4) Events, (5) Cultural products. Do not mistake "unit of analysis" for "cases" or "observations"

Population

The population of your analysis is imperative. What is the population you are analysing, i.e. is it geographic, socio-economic, demographic, temporal etc.

Levels of Measurement

Also known as the scale of measure it is a classification that describes the nature of information within the numbers assigned to variables. There are three best known levels, or scales, of measurement: nominal, ordinal, and interval

Longitudinal data

Longitudinal data, track the same sample at different points in time. The sample can consist of individuals, households, establishments, and so on. Longitudinal data have a number of

advantages over repeated cross-sectional data. Longitudinal data allow for the measurement of within-sample change over time, enable the measurement of the duration of events, and record the timing of various events.

Cross-sectional data

Cross-sectional data are data that are collected from participants at one point in time. Time is not considered one of the study variables in a cross-sectional research design. It is worth noting that in a cross-sectional study, all participants do not provide data at one exact moment. Even in one session, a participant will complete the questionnaire over some duration of time. Nonetheless, cross-sectional data are usually collected from respondents making up the sample within a relatively short time frame (field period). In a cross-sectional study, time is assumed to have random effect that produces only variance, not bias.

Micro-data

Micro data can generally be described as individual level data. These data have often been collected from each individual through a survey or interview. In such a dataset, each row typically represents an individual person and each column an attribute such as age, gender or job-type.

Macro-data

Macro data is generally a term used to describe mainly two subtypes of data; (1) aggregated data, (2) -system-level data

Aggregated macro data provide information constructed by combining information on the lower level units, which the higher level unit is composed of. Examples of aggregate data include summaries of the properties of individuals, unemployment statistics, demographics, GDP etc. Most often, aggregated macro data imply that the variables are summaries of the properties of lower level units and not measures of inherent higher level properties.

Probability Sample

A probability sampling method is any method of sampling that utilizes some form of random selection. In order to have a random selection method, you must set up some process or procedure that assures that the different units in your population have equal probabilities of being chosen. Humans have long practiced various forms of random selection, such as picking a name out of a hat, or choosing the short straw. These days, we tend to use computers as the mechanism for generating random numbers as the basis for random selection.

Standard classifications/Harmonisation

Standard classifications are those that follow prescribed rules and are generally recommended and accepted. They aim to ensure that information is classified consistently regardless of the collection, source, point of time etc. Standard classifications contain categories that are mutually exclusive and exhaustive i.e. any given response or data can be matched to one category only and every possibility is anticipated and catered for.