Data Acquisition and Management

By

Okoruwa, V. O, E.A. BAMGBOYE
A.E. Oluleye A.E and J.O. Olaomi

University of Ibadan, Ibadan

Training Workshop; Faculty of Science
Wednesday, November 22, 2017
Outline

- Introduction
- Objectives
- Data acquisition (sources of data; concept of research design and study design)
- Data collation (coding, entry, entry program and cleaning)
- Data analysis (required skills, methods and type of analysis, and hypothesis testing)
Introduction

Data acquisition and management (collation and analysis) are very important aspects to consider when conducting researches or studies in science, social sciences and humanities.

The process of acquiring good data hinges on the type of experimental/design of study used. Collation of data takes place immediately after data acquisition and this is preceded by the data analysis.
Objectives

Expected outcome - understanding the following:

- Concepts and types of data acquisition technique used in research and why

- The basic things to do after collection of data through either primary or secondary sources.

- The basic fundamentals/concepts and importance of data management in research
Objectives cont...

- Basic elements of data collation
- Relevant software that can be employed for data collation
- The appropriateness of the analytical techniques to be used based on researcher’s data
Data Acquisition & Management

**Purpose**
- Generate information
- Planning
- Action
Sources of data

• Routine or regular sources (Census, VRS, Institutions): CBN, NBS, FAO, NPC etc.
  - Yields Secondary data

• Non-routine sources (Ad-hoc surveys, experiments)
  - Yields Primary data collected by researcher directly from the field
Sources of data cont...

NON-ROUTINE SOURCES

- QUALITATIVE
- QUANTITATIVE
Sources of data cont...

**Qualitative Methods-types**
- Ethnography/Case studies
- Life Histories- biographies
- Participant Observation- lives and takes part
- Non-Participant observation
- Focus Group Discussion
- Group Interviews
- In-depth Interviews
- Semi-Structured Interviews
Sources of data cont...

Qualitative data collection

- Informal approach including:
  - familiarization or reconnaissance survey,
  - rapid rural appraisal (RRA) and
  - participatory rural appraisal

- Justification for its use:
  - most effective way of learning a good deal about a particular topic in a short time

- Danger of the approach:
  - highly prone to bias or incorrect impression e.g. key informant share a particular prejudice or viewpoint
Sources of data cont...

**Quantitative data collection**

- Most widely used method of obtaining primary data includes:
  - direct observation;
  - interviewing respondents and
  - extracting from records

- Most widely abused

For a comprehensive data collection a combination of both approaches (qualitative and quantitative) is most ideal
Concept of research design

- Basic plan or architecture for empirical research
- Bridge between the research questions and the data
- Shows how the research questions will be connected to the data
- And what tools and procedures will be used in answering the questions
Concept of research design cont…

Components includes:

- Study strategy or study design
- Study subjects
- Instruments for data collection
- Plans for statistical analysis
Study design

Types

- Experiments
- Quasi-experiments
- Ex-Post Facto (EPF) design
  - Case control studies
  - Cohort studies
- Survey sampling design
- Observational studies
  - Cross sectional or analytic
Sources of errors in experiments

- Choice of experimental subjects
- Choice of treatments (Drips or vaccines or surgery or fertilizer)
- Allocation of treatment to experimental units or subjects
- Definition of outcome measures (subjective or objective)
- Sample size.
Study design cont....

**Removal of errors**

a. *Allocation bias*
   - Ensure objectivity in the allocation procedure
   - Use randomization procedure (random table or computer)

b. Design bias (*Random Variation*)
   - Matching if two groups
   - Blocking if more than 2 groups
   - Cross-over design
   - Stratification

c. *Assessment bias*
   - Blocking (to remove subjects' and members' perception)
   - Single - removes subjects psychological reaction
   - Double
   - Triple
Adjustment for drop out of Non-responses

Estimate level of non-response
Inflate sample size accordingly
Multiplying factor: $q = \frac{1}{1-f}$;

f is called estimated non-response rate.
Finite Population Correction
Useful more in descriptive studies

\[ n^1 = \frac{n}{1 \times \frac{n}{N}} \]

- \( N \): total population
- \( n \): sample size
- \( n^1 \): Adjusted sample size
Quasi - Experiments
Similar to experiments but lack at least one of the 3 essential features of a true experiment

The missing feature is usually randomization or the control group or both

Example:
A study to measure the effectiveness of the neem tree extract in the treatment of ring worm in a single group of patients seen before and after treatment.
Study design cont....

**Ex-Post Facto (EPF) designs**

Investigations designed to examine the relationship among variables of factors

- lack active manipulation of the intervention
  - the investigator does not actively apply a treatment
  - the treatment/intervention has occurred in the past (hence Ex-post facto= after the fact)

**Example:** An investigation to study the effect of maternal smoking on birth weight of babies.
Type of Observational Studies

Descriptive Studies
- case reports
- case series
- cross sectional studies

Analytic Studies
- case-control studies
- cohort study
- Ecologic studies
Study design cont....

Sample and Sampling

- Sampling is the process of selecting a sample from a population
- Sampling is necessitated by factors such as feasibility and limited resources

Why Sampling!!!!!

- Avoidance of bias
- Achieve high precision for a given outlay of resources
Avoiding bias in sampling

- take a probability sample (known as a random sample)
- sample has a known chance of being selected

Types of probability sampling

- simple random sample
- systematic sample
- stratified random sample
- cluster sample
- multi-stage random sample
- multi-phase random sample
Importance of probability sampling

1. Ensures findings from sample can be generalized about a population
2. Provides ways of determining the number of respondent needed
3. Specifies the chance (probability) that any unit will be included in the sample
4. Estimate the error resulting from sampling the selected units rather than carrying out total census
5. Assists us to determine the degree of confidence that can be placed in the population estimates derived from the sample
Types of Non-probability sampling

- volunteers
- haphazard
- quota
- purposive
- convenience- accidental sampling
- snowballing sampling
Decision on sample size for a study

- too few subjects makes estimates unreliable, imprecise, and low power
- too many subjects needless waste of resources
- rely on logistical and pragmatic considerations
Other components of study design

- Data collection instruments
  - the questionnaire or case report forms,
  - Need testing and validation, translation
Once data is collected the need arises to bring such data together and presents it in a manageable form- a process termed **data collation**

For easy interpretation and analysis, data collation usually involves summarising and tabulation of collected information.

**Processes in data collation:**

Data Coding; entry/capture; cleaning; while making use of a good data entry progm.
Data collation cont..

**Data coding**: process of grouping responses either from questionnaires or interview schedules into categories.

**What to code for?**

themes, concepts, ideas, name of people, agencies, major projects, dates, stages or steps of a process or anything useful in tying things together.
Example of Coding guide data in spreadsheet

<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>Case #</td>
<td>001-150 (for 150 cases)</td>
</tr>
<tr>
<td>4</td>
<td>Gender</td>
<td>1 = male, 2 = female</td>
</tr>
<tr>
<td>5-6</td>
<td>Month of birth</td>
<td>1 = January, 2 = February, etc.</td>
</tr>
<tr>
<td>7</td>
<td>Type of health</td>
<td>1 = Private, 2 = Public, 3 = Uninsured</td>
</tr>
<tr>
<td>8-9</td>
<td>Insurance</td>
<td>8 = Don’t know, 9 = Refused/Missing</td>
</tr>
<tr>
<td>10</td>
<td>Height in inches</td>
<td>e.g. 64.0 inches = 64, 64.5 = 65, 66.25 = 66</td>
</tr>
<tr>
<td></td>
<td>Marital Status</td>
<td>1 = Married, 2 = Widowed, 3 = Separated,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = Divorced, 5 = Never Married</td>
</tr>
</tbody>
</table>
Data collation cont..

**Data entry**: covers the initial computerized log-in of the completed questionnaire.

The process involves five elements:

- the data entry screens,
- validity checking,
- consistency checking,
- data combination and transformation, and
- data base organization.
Data cleaning (scrubbing): deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data.

**Phases of data cleaning**

- Data analysis
- Definition of transformation workflow and mapping rules
- Verification
- Transformation
- Backflow of cleaned data
Data entry programmes:
Statistical software packages:- SPSS, Microsoft access, Microsoft excel, ISSA

Database packages:- dbase, paradox etc. (allow the user to create screen forms with validity and consistency checks).

Custom programme: write a programme in a language such as FORTRAN, C-language, Visual Basic to perform data entry
Data analysis

What is data analysis?
It is a body of methods that help to describe facts, detect patterns, develop explanations, and test hypotheses,

OR

A process of looking at and summarizing data with the intent to extract useful information and develop conclusions
Data analysis cont...

Involves iterative process of preparing descriptive statistics, frequency distribution tables, selection of variables for cross tabulations, running of inferential statistical analysis and tests of significance.

Broadly divided into exploratory data analysis and confirmatory data analysis.
Data analysis cont...  

Required skills for data analysis:

- conceptual clarity,
- field experience,
- adequate knowledge of statistical techniques and
- ability to use computer to manipulate data
Data analysis cont...

**Types of analysis:**
- Qualitative and
- Quantitative analysis (particular attention)

**Methods of analysing data:**
- statistical,
- mathematical,
- Econometrics/Biometrics
- cartographic,
- Mapping (involves the use of GIS) and
- Others
Data analysis cont...

Decision on what type(s) of method to use depends largely on the hypothesis to be tested and the types of data available.

Descriptive statistics:
- Measure of central tendency
- Measure of dispersion
- Frequency distribution table and graphs
Measure of central tendency: In practice, three different types exist including Mode, Median and Mean.

Examples - If the weights in kg of seven boys are: 61, 63, 63, 64, 68, 71, 73

- The Mode is: 63kg
- The Median is: 64kg
- The Mean is: 66.14kg
Data analysis cont...

**Measure of dispersion:** This captures the deviation of different observation from the mean and they include:

- Range,
- Quartile Range,
- Standard Deviation (SD),
- Variance,
- Coefficient of Variation (CV),
- Coefficient of Skewness (CS), and
- Moving Average (MA)

Examples

Wednesday, November 22, 2017
Data analysis cont...

**Frequency distribution**: This provides tabular and graphical display of data in group form (either as disaggregated values, percentages, cumulative, sum, maximum, minimum etc)

exempl.

**Cross tabulation**: A popular method for analysing data to ascertain measures of association. The procedure forms two or multi way in which tests can be performed

exempl.
Data analysis cont...

**Principal component analysis (PCA):** A way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences

**Advantage of PCA**

- Useful when patterns in data is difficult to find and when the luxury of graphical representation is not available
- Minimum loss of information in the process of finding a pattern and reducing the number of dimensions
- Used for many purposes- e.g. image compression, social capital analysis, multidimensional poverty etc.
Inferential statistics

Often employed in decision making about a population on the basis of sample information.

It involves hypothesis testing and confidence intervals construction and can be grouped into two broad categories-

PARAMEDIC and

NON- PARAMETRIC STATISTICS.
Data analysis cont...

**Parametric statistics:** Underlying assumption is that the population being investigated has normal distribution.

**Some tools of parametric analysis**

- Regression analysis (simple & multiple),
- Correlation analysis (simple & multiple)
- Factorial analysis employing- ANOVA (one-way & two-way), ANCOVA, MANOVA, etc

All these employ the t-test and F-test for individual parameter and overall model test of significance.

exmpl.
Data analysis cont...

Non-parametric analysis: Becomes handy when population distribution is skewed (not normally distributed)

Some variables analysed with this technique include Categorical variables (dichotomous and multi-dichotomous), ordinal values etc.

Types: Chi-square, Binomial test, runs test, Kolmogorov-Smirnov test, Kruskal-Wallis, Kandall, Probit, Logit, Multinomial Logit, Nested Logit, Spearman Correlation analysis etc. examp.}

Choosing the right statistical test
Hypothesis setting and testing:

In attempting to reach decisions, it is useful to make assumptions (or guesses) about the populations involved. Such assumptions, which may or may not be true, are called statistical hypothesis. They are generally statements about the probability distribution of the population.
Data analysis cont... 

When we formulate hypothesis stating that there is no difference between procedures or variables (in terms of relationship), such hypotheses are often termed null hypotheses and are denoted by $H_0$.

Any hypothesis that differs from a given hypothesis is called an alternative hypothesis and are denoted by $H_a$.
**Data analysis cont...**

**Tips for setting good hypothesis**

*Hypothesis must:*

- be formulated in such a manner that their implications and relationships to the problem can be shown logically

- be as simple as possible both in terms of theory and number of variables involved

- be capable of verification or rejection within the limits of the research resources available
Data analysis cont...

- be capable of verification or rejection within the limits of the research resources available

- be adequate relative to providing a meaningful degree of problem resolution
Steps in statistical hypothesis testing

i. Clearly stated the null hypothesis and the alternative
e.g. $H_0: \beta = \mu$

$H_1: \beta \neq \mu$  for two-tail test
$H_1: \beta < \mu$  for one-tail test
$H_1: \beta > \mu$

ii. Determine the level of significance, which could be 1%, 5% or 10% having in mind that 1% is better than 5%, and 5% better than 10%.

iii. Chose the test statistic which could be Z-test, T-test, F-test, etc.

iv. Compute the value of the statistic from a random variable of size $n$

v. Decision rule: Reject the null hypothesis (Ho) if statistics value computed falls within the critical region and accept the alternative (i.e., if calculated statistics is greater than tabulated).
Data analysis cont…

Level of significance

In many instances we formulate a statistical hypothesis for the sole purpose of rejecting or nullifying it.

In practice, a significance level of 0.05 (P< 0.05) or 0.01 (P< 0.01) is customary, although other values are used. For example, if the 0.05 (or 5%) significance level is chosen in designing a decision rule, implications are...
that there are about 5 chances in 100 that we would reject the hypothesis when it should be accepted; or

that is, we are about 95% confident that we have made the right decision.

In such case we say that the hypothesis has been rejected at the 0.05 significance level, which means that the hypothesis has a 0.05 probability of being wrong.
Thanks for Listening!!