# Dr. Norbert Cheung's Lecture Series

## Level 5 Topic no: 37

# Research Data and Method Selection

**Contents**



**Reference**

Engineering Research: Design Methods and Publication, Herman Tang, Wiley, 2021.

# 1. Data in Research

Data are fundamental for much of research because they are raw materials and can provide the connection between the real-world problems and the formalization of a study model, hypothesis, or theory. Without appropriate and reliable data, research outcomes remain unverified regardless how perfect the model, hypothesis, or process of research are.

Most research projects are data-driven. We use data to discover, explain, prove, or disprove new phenomena and principles. For engineering and technology research, almost all tasks are related to data, such as collection, analysis, interpretation, and validation. Without data, it is an opinion.

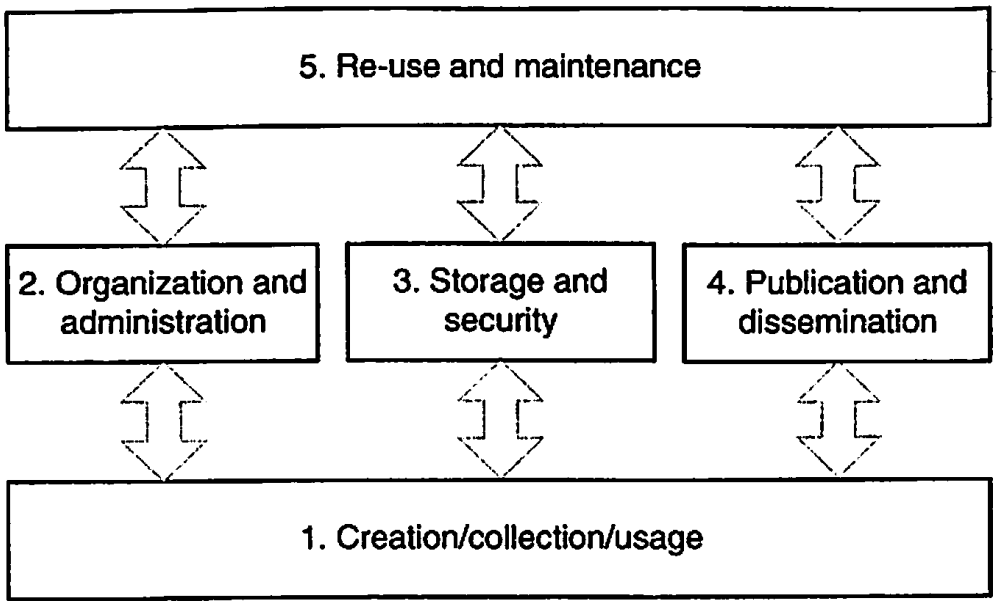



**Figure 4.1** Main elements of research data management.

Data Management is very important. 5 elements in a data management plan.

1. *Data Creation and Usage.* This can be the main effort in a research project.
2. *Data Organization and Administration.* This element is to specify how data are organized and managed.
3. *Storage, Sharing, Security, and Back Up.* Data storage, associated issues with sharing, accessibility, and security controls, must be specifically designed.
4. *Publication and Dissemination.* Research data and results may be disseminated in different ways. The range and timing of dissemination and approval process are defined.
5. *Re-use and Maintenance.* Data generated from a research project can have a long-term reference value, in addition to its importance to research validity. During and after a research project, data maintenance should be planned.

## Characteristics of Data

### *#1: Data Distribution*

Data probability distribution is important as a research assumption. Many data analyses are valid only if the assumptions on the data distribution are true. There are also a lot of research on the characteristics and modeling of data distributions. For continuous data, the basic information of common distributions is shown in Figure 4.2 and Table 4.1.

**Table 4.1** Common data distributions.

| Distribution | Probability density function | Application | Example |
|---|---|---|---|
| Uniform | $f(x) = \frac{1}{b-a}$ | For the cases with a range ($a$, b) of equally likely values | Strength and thermal conductivity of metal matrix composites (Mazloum et al. 2019) |
| Triangle | $f(x) = \begin{cases} \frac{2(x-a)}{(c-a)(b-a)} \\ \frac{2(b-x)}{(b-c)(b-a)} \end{cases}$ | For the analysis of risk and stochastic processes with min ($a$), max ($b$), and mode ($c$) | Dual phase nano-particulate AlN composite (Zhao et al. 2019) |
| Normal (Gaussian) | $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ | Common statistical distribution; typical assumption for unknown random variables | Modeling for fragment-size distribution (Kim and No 2019) |
| Logistic | $f(x) = \frac{e^{-\frac{x-\mu}{\sigma}}}{\sigma\left(1+e^{-\frac{x-\mu}{\sigma}}\right)^2}$ | For the situations with longer tails and higher kurtosis than the normal distribution | Bootstrap confidence intervals of CNpk (Gadde et al. 2019) |
| Exponential | $f(x) = \lambda e^{-\lambda x}$ | To model the time between events in a continuous Poisson process | Modified chain sampling plan (Jeyadurga et al. 2018) |

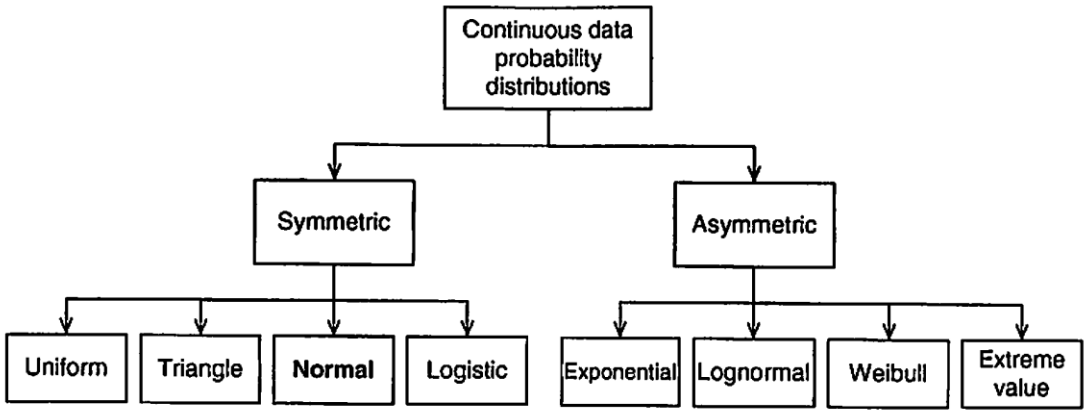



**Figure 4.2** Common distribution types of continuous data.

## *#2 Considerations in Research Data*

*Source.* Data source is also a source of variation. For example, data may be collected from either a single experiment or in multiple settings. If data are collected from (even slightly) different conditions, such as multiple pieces of test equipment and production shifts, the data most likely have different characteristics, say mean values. The characteristics can be both a challenge and an opportunity to study the variation and its causes.

*Elusiveness.* Data are explicitly presented in such a way that the meaning is sometimes obvious. However, in many cases, data need to be thoroughly studied using different approaches to reveal the real meanings. That is often a major challenge in research.

*Conditions.* A key point should be kept in mind is about the conditions of collection when doing research. Data of any phenomenon are limited due to sources, conditions, data analysis, and time. Therefore, research outcomes based on the limited data can be conditional. Again, it can be both a challenge and an opportunity based on limited data to draw a general conclusion.

*Ephemerality.* Data may be available for a short time in a particular place. The observations may be different at different times due to known and unknown factors. For example, observations on a machine's operation are dynamic and transient in nature. If data have such a short-lived characteristic, even if valuable, using them can be controversial because they are difficult to validate later on.

### *#3 Data Preparation*

Before doing a data analysis, there may be a few preparation tasks needed:

- *Data inspection* is the first task after data collection to look for obvious or subtle issues. Visualization of raw data, such as using various charts, is a good way for data inspection. We should have the predefined rules for data inspection. Data inspection also relies on researcher's experience.
- *Data cleaning* is another pretreatment process to remove incomplete, erroneous, and duplicated data, which can be important for following analysis. In addition, for a large amount of data, it is often a good practice to organize or condense the data to make them more manageable for an effective analysis.

- *Data transformation* is a process of converting data from one format, type, or structure into another one. We may consider this process part of either data preparation or an early phase of data analysis. In many cases, we transform qualitative data to quantitative data for further analysis.

### *#4 Overall Data Analysis*

Data analysis is a systematic process. The main objective of data analysis is to understand or discover the messages contained in the data by extracting and summarizing their main characteristics. There are several factors to select methods for data analysis. One factor is about the nature of the data, i.e. quantitative, comparative, or qualitative. The other is the assumptions about the data, such as distribution, independence, sample size, and statistical significance. Therefore, data analysis and analysis methods can be case and discipline dependent.

e.g., Examples in MATLAB

## 2. Types of Data

### *#1 Primary Data*

The types of data to use and the approaches to collect the data are key focuses for researchers. Data sources can be either primary or secondary or combined, as shown in Figure 4.4. Primary data are the first-hand data that are collected by the original researchers. Such examples include the data from experiments, through observations, and interviews. Otherwise, the data are secondary. If shared with other others or to public, the data are also called open data, which are a type of secondary data in nature.

## *#2 Secondary Data*

Secondary data are the data that already exist in the reports or databases generated by other parties. The common sources of secondary data are handbooks, journals, government databases, and commercial databases, and the like. One type of secondary data is called administrative data, which are collected routinely as part of the day-to-day operations of an organization or government agency. Secondary data may be derived from primary data as well. For instance, written sources that interpret or record primary data are secondary.

Using secondary data in research reports, we need to clearly define, assess, and disclose the conditions of derivation and conclusions based on secondary data. For example,

"*Determining a Cut-Off Point for Scores of the Breastfeeding Self-Efficacy Scale–Short Form: Secondary Data Analysis of an Intervention Study in Japan*" (Nanishi et al. 2015)
"*Robust space time processing based on bi-iterative scheme of secondary data selection and PSWF method*" (Du et al. 2016)
"*The State and Prospects for Development of Railway Transport Infrastructure in Eastern Poland – Secondary Data Analysis*" (Jarocka and Glińska 2017)

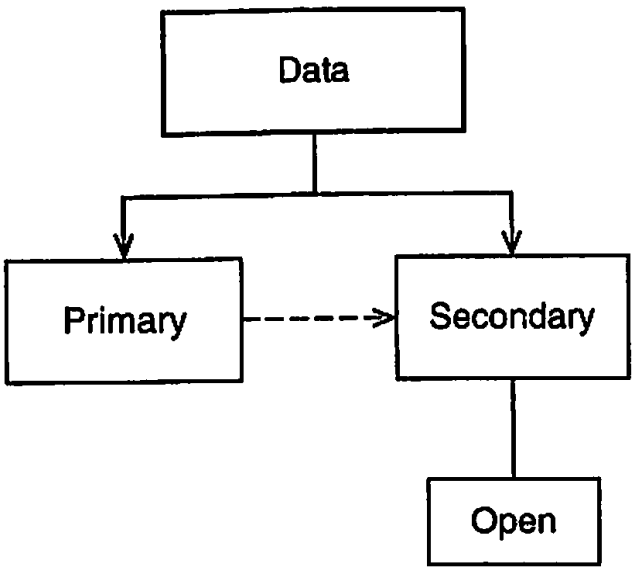



**Figure 4.4** Basic types of data.

## *#3 Open Data*

Traditionally, the data used in applied research and R&D are not open to others. However, some professionals consider that closed data hinders testing the validity and reliability of research results. A trend is to share or "open" research data to external parties. Open data can be available to everyone to use without restrictions from copyright, patents, or other mechanisms of control, which is similar to those of other "open" movements, such as open sources and open access.

**Table 4.2** Some government data websites.

| Country | Website address |
|---|---|
| Canada | https://open.canada.ca |
| China | data.stats.gov.cn |
| France | www.insee.fr |
| Germany | www.destatis.de |
| Italy | https://www.dati.gov.it |
| Japan | https://www.data.go.jp |
| Russia | https://data.gov.ru |
| UK | www.natcen.ac.uk |
| US | https://www.data.gov |

## Quantitative Data vs Qualitative Data

Data can be either quantitative (numerical) or qualitative (descriptive). In technical research, we mostly use quantitative data because we can analyze them using various statistical and/or numerical techniques. Other research projects, e.g. in social sciences, are often qualitative in nature.

| | Qualitative | Quantitative |
|---|---|---|
| Strength | • Detailed and in depth (rich)<br>• Providing a nuanced understanding | • Clear and reliable<br>• Easy to handle and analyze |
| Weakness | • Often small sample size<br>• Subjective<br>• Time for handling and analysis | • Often superficial<br>• Not comprehensive to complex situations |

## *#1 Numerical vs non-numerical*

Quantitative data are presented by numbers. They may be in either a continuous or discrete format. Continuous data are an infinite number of possible values in a range. For example, any number in a range between $a$ and $b$; where, $a$ and $b$ are real numbers and $a \neq b$. It is clear that continuous data are uncountable.

In contrast to quantitative data, qualitative data do not measure the attributes, characteristics, properties of a phenomenon but measure its types. Qualitative data are not expressed numerically but may be described and expressed in descriptive words, such as name, symbol, or a number code. The information from interviews, surveys, field notes, documents, electronic media, etc., is often qualitative. In many complex studies, using qualitative data and methods can make the research results more explicable.

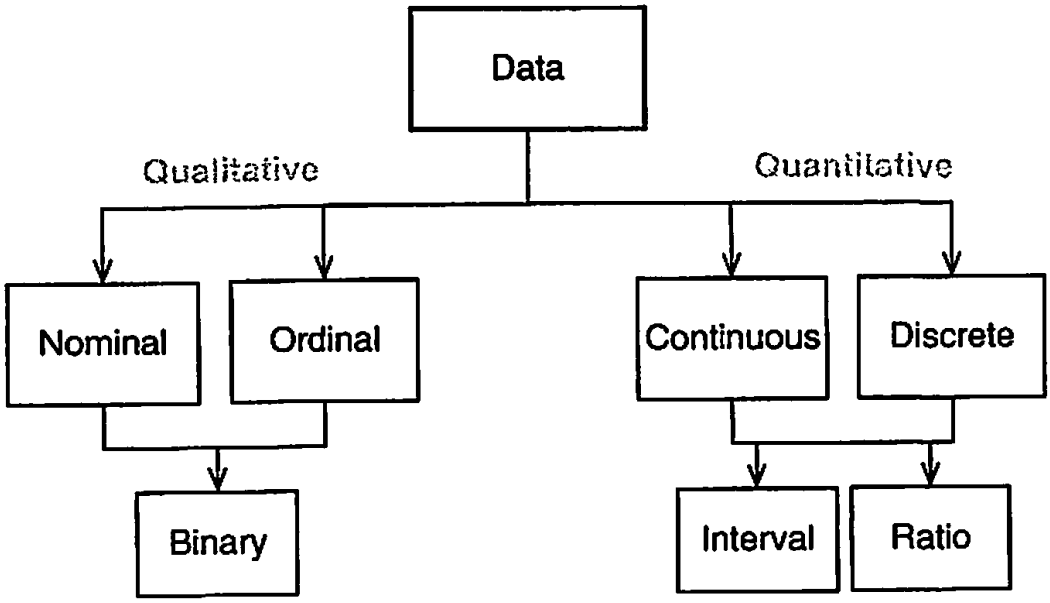
## *#2 Key factors of Qualitative Data*

- *Data Accuracy.* Data accuracy is about the closeness of a measured or collected value to a true value. Data accuracy is normally addressed through the calibration of a measurement system.
- *Data Repeatability.* It is defined that the same object is measured by the same people using a single instrument on different measurement occasions.
- *Data Reproducibility.* It is measured by two or more human individuals on the same product or performance using identical measurement instruments.

## *#3 Reliability of Qualitative Data*

It may be difficult to measure qualitative data accurately. Our interpretation based on qualitative data may be more challengeable than that from quantitative data. However, this does not mean that the qualitative data and the associated findings are less valuable. The observations are normally experience-, attention-, and/or perspective-based. They may be informative and nuanced that lead to great insights into human society.

## Scales of Data




### *#1 Scales of data*

Data scale specifies the categories of measurements and data. Based on a measurement scale, we may consider data nominal, ordinal, etc., refer to Figure 4.5. The scale of data also determines the measurement procedure and following data analysis.

### *#2 Nominal Data*

Nominal data are separate, non-ranked data. Sometimes, they are called categorical data. Each data uniquely belongs to a specific category, which may be coded by a number that has no real numeric meaning. For example, data collected by color are nominal data.

### *#3 Ordinal Data*

Different from (or maybe better than) nominal data, ordinal data can be an order, in terms of relative significance or priority, in either an increasing or a decreasing order. The data fall into categories, and the numbers for the categories may have physical meanings. For example, we may have a rating on a scale from 1 (lowest) to 5 (highest) for an ordinal dataset. We may use ordinal data to measure qualitative concepts.

### *#4 Binary Data*

Binary data are in two possible states, traditionally labeled as the combination of "0" and "1". In computer science and engineering, we call a binary digit a bit.

### *#5 Interval and Ratio Data*

Interval and ratio data are quantitative. Interval data are sorted numeric scales, meaning there are not only the order of data but also the equal difference between the individual values. In interval data, the distance between attributes does have a meaning, which can be useful in carrying out more sophisticated statistical analysis. Two simple examples can be time and temperature data.

The interval scale with a natural origin is called a ratio scale. A length measurement is an example of such a ratio scale. Different from interval data, ratio data have a true zero point. Ratio data are in relation to a zero value (e.g. a distance).

## 3. Data Collection

### Data Collection Sampling

In most cases of conducting research, it can be either impractical or uneconomical to get all the data of a target situation. Thus, a research project is a study of the samples of data collected from the entire population or situation.

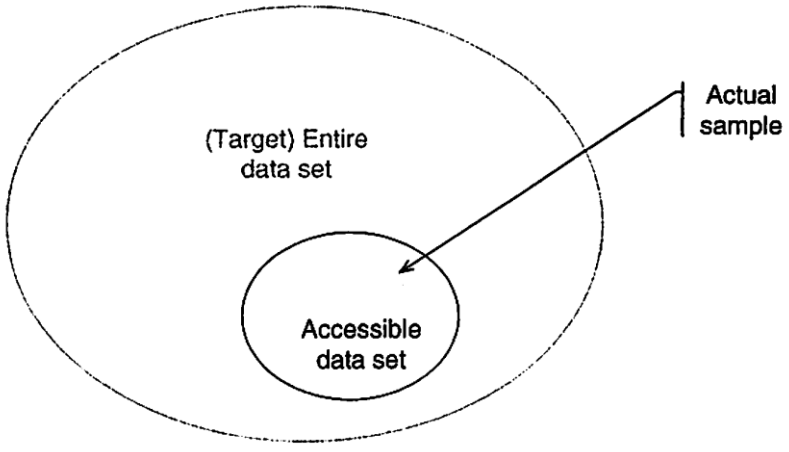



**Figure 4.6** Entire data, accessible data, and samples.

It is very possible that the samples do not accurately represent the status or condition of the entire population. Such an inaccuracy may be called sampling bias.

## Sample Size

An appropriate sample size is very important to all research-based experimental and empirical studies. Sample size should be determined based on the research objective and requirements. Small samples can undermine the internal and external validity of a study. With an adequate sample size, study results may become statistically significant. Based on statistics, sampling error can be estimated. On the other hand, obtaining large samples requests more resources.

- For a statistical analysis based on quantitative data, the minimum sample size should be 25. With the minimum sample size, we may claim the analysis results are statistically significant in many cases.
- For many statistical analyses, there are recommendations for sample size requirements and calculation.
- For the research based on a hypothesis, an appropriate sample size may be determined according to an operating characteristic curve or formula for a type II error ($\beta$). For example, a two-tailed test (Figure 4.7) has

$$\beta = \Phi\left(Z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-Z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right)$$

where, $\Phi$ – standard normal cumulative distribution function, $Z$ – test statistic, $\alpha$ – type I error, $\delta$ – mean deviation, $n$ – sample size, and $\sigma$ – standard deviation.

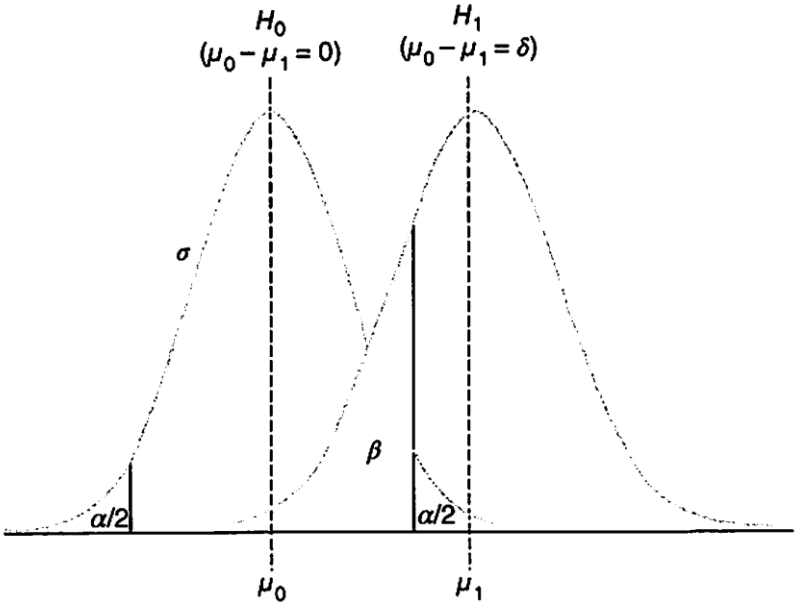



**Figure 4.7** An illustration of types of I and II errors.

- For data collection, such as for a survey, we can decide the sample size based on the entire population, the confidence level (normally set at 95%), confidence interval (or margin of error, e.g. ±3%). For example, if a population is small, say fewer than 100, it is recommended that data be collected from the entire population. If a population is around 500, the sample size may be around 50%. In general, the larger the population, the smaller the sampling percentage can be, considering the feasibility and cost of a study.

## Probability Sampling Methods

A common sampling method is probability or random sampling. Using this method, we assume that the chance of each sample is the same and that the samples statistically approximate to the characteristics of the total population.

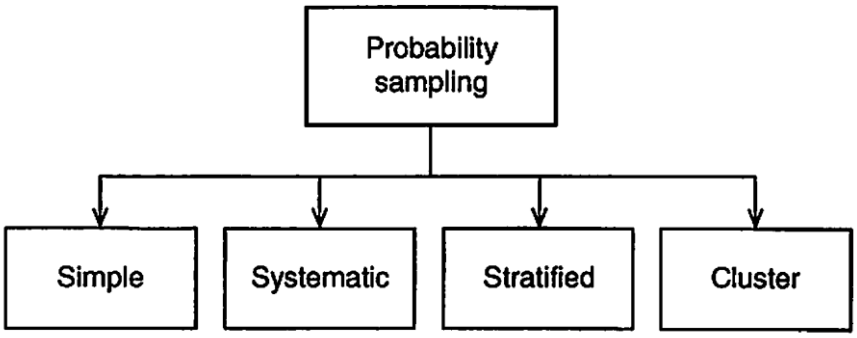



**Figure 4.8** Types of probability sampling.

### *#1 Simple Sampling*

In a simple random sampling, all elements of a population are considered and have an equal chance of being selected at any stage during the sampling process. Randomness is built into a sampling design so that the properties of the population can be assessed probabilistically. Hence, we may conclude that a simple random sampling is unbiased.

### *#2 Systemic Sampling*

We often use systematic sampling for a large list to select elements from an ordered sampling frame. We pick the every $k$th one from a complete set with total $N$ elements in sampling. Thus, the sample size is $\frac{N}{k}$. In other words, the $k$, a fixed interval, is determined by $k = \frac{N}{n}$ if $N$ is known and $n$ is decided. An example of systematic sampling is in a study of "Automated Pre-Seizure Detection for Epileptic Patients Using Machine Learning Methods" (GÜL et al. 2017).

### *#3 Stratified and Cluster Sampling*

Stratification sampling is the process of dividing members of a population into homogeneous subgroups before sampling. For a large population, stratified random sampling is an alternative to systematic sampling. We divide a population into mutually exclusive groups (called strata) and then use simple sampling to collect samples from each group equally. For example, this method was used in the research of "Refined Stratified Sampling for efficient Monte Carlo based uncertainty quantification" (Shields 2015).

Cluster sampling is similar to stratified sampling. Sometimes, researchers divide a population into natural groups based on their existing quantities. The sample size may be different. For example, the sampling probability may be proportional to size. The groups are called clusters in a cluster sampling. For example, an electrical study on "An I/O Efficient Distributed Approximation Framework Using Cluster Sampling" (Zhang et al. 2019) used the method.

**Table 4.4** Stratified sampling vs. cluster sampling.

| | **Stratified sampling** | **Cluster sampling** |
|---|---|---|
| Groups in population | Population divided into groups | Naturally occurring groups |
| Sampling | Individually from all the strata | Collectively from selected group (clusters) |
| Homogeneity | Between groups | Within group |
| Advantages | Precision and representation | Cost and efficiency |

## 4. Method Selection

### Selection Factors

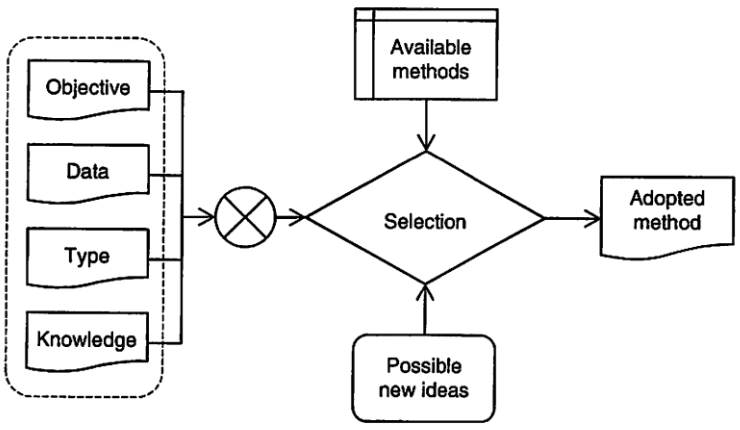



**Figure 4.10** Factors of research method selection.

## *#1 Object Driven*

The purpose and expected accomplishment of a research project is the first question for us to think about when selecting a method. In many cases, objectives of research play a determinative role to select methods, assuming related data are available.

**Table 4.5** Research objectives and methods.

| Objective | Research type | Common method |
|---|---|---|
| (1) To generate understanding or principle | Basic (Scientific) | Qualitative |
| (2) To advance understanding or principle | Basic and Applied | Qualitative (as well as quantitative) |
| (3) To apply understanding or principle | Applied | Quantitative |
| (4) To develop new product or process | R&D | Qualitative (direction) and quantitative (details) |

## *#2 Data Based*

The data play a crucial role in the method selection, as a method is to use for data collection and analysis. In other words, data and methods can go hand in hand in engineering and technical studies. For example, we may use a linear regression and other types of statistical analysis for continuous data. Sometimes, we need to work on their relations and determine the best pair of data and method.

## *#3 Various Process Steps*

The unique characteristics of action research are to go through the simultaneous process of taking informed actions and doing research to seek transformative change. From this point, action research is a kind of comparative investigation on the conditions and effects of various forms of actions and research.

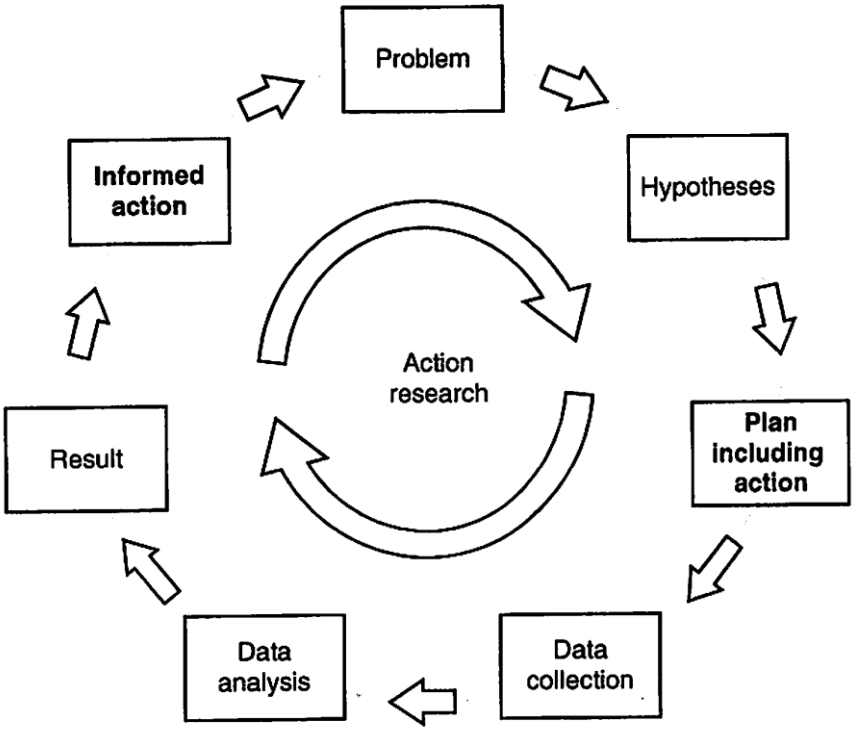



**Figure 4.11** A process of action research.

## *#4 Qualitative vs Quantitative*

**Table 4.6** Characteristics of qualitative and quantitative analyses.

| Characteristics | Qualitative | Quantitative |
|---|---|---|
| Origin | Art and social science | Natural science |
| Researcher's knowledge | Rough idea | Clear understanding |
| Question | What, why, how | How many, when, where |
| Data sampling | Purposeful | Probabilistic/random |
| Format | Words, image, etc. | Numerical data |
| Reasoning (epistemology) | Empiricism/induction | Rationalism/deduction |
| Results | Interpretation | Detailed and structured |
| Extension | Non-generalizable | Maybe generalizable |
| Nature | More subjective | More objective |
| Strength | Observational | Statistical |

## #6 Induction vs Deduction

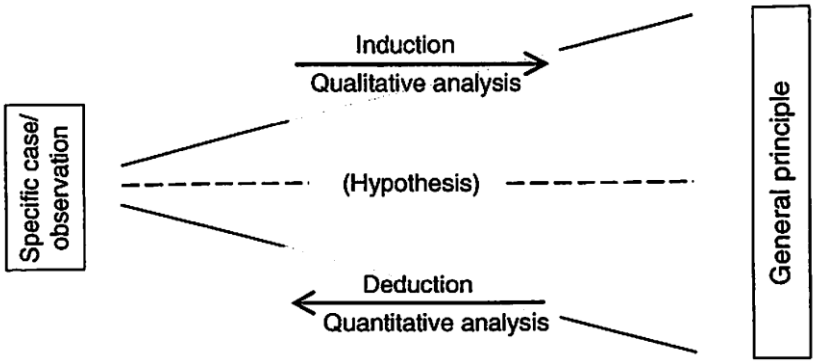



**Figure 4.12** A diagram of induction vs. deduction.

Inductive reasoning starts from specific data and develops a general conclusion, such as a new theory and model. From a problem, we may induct to a systematic observation. This approach is exploratory rather than confirmatory. Using this approach, we may develop a theory or principle rather than use it. In other words, we induct from specific observations to generalize how that thing works.

In engineering applied research and R&D, deductive approaches are widely used with the known general principles and theories. While when approaching a topic with little background on the subject, induction is more likely to be used. In a word, which reasoning approach to use depends on research objective. Readers may refer to relevant literatures, for instant, *How to Get a Ph.D.: Methods and Practical Hints* (Mämmelä 2009) that provides a summary.

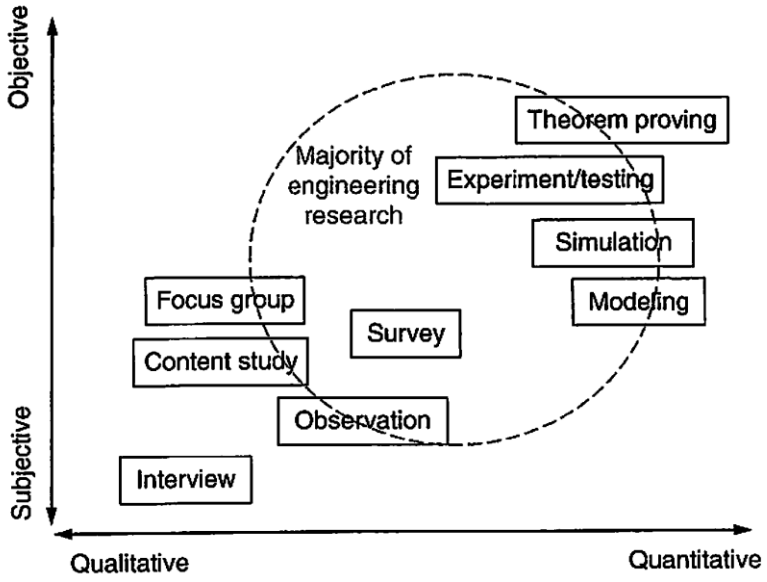



**Figure 4.13** Characteristics of research methods.

# 5. Summary

### Data in Research

1. Data play a critical role in research. Most research is data driven and research outcomes need data to verify and validate.
2. Data management includes the creation, administration, storage and security, publication, and maintenance of data.
3. Data science is based on applied statistics. It can be a research subject and more often the method for other research subjects.
4. There are several data distributions, such as normal, exponential, Weibull, etc., commonly used in engineering research.
5. Considerations for research data include their source, elusiveness, conditions, and ephemerality.
6. Main steps of data handling in research are preparation, analysis, and interpretation.

### Types of Data

7. *There are two types of data in terms of data sources*: primary (to collect new) data and secondary (to use existing) data. Open data is a type of secondary data, which have various sources.
8. The data can be either quantitative (numerical) or qualitative (in formats of words, etc.)
9. The quality of quantitative data can be measured on their accuracy, repeatability, and reproducibility.
10. The scale of data can be categorized into normal, ordinal, binary, interval, and ratio.

### Data Collection

11. Collected data samples are the subset of the accessible data set, which is a subset of the entire data set.
12. There are several methods to determine sample size.
13. Probability (or random) data sampling methods should be used when feasible, which include simple, systematic, stratified, and cluster methods.
14. Non-probability data sampling methods include convenience, quota, purposive, expert, and snowball methods.
15. Non-probability data sampling has some advantages but is lack of representation of the entire population.

## Method Selection

16. Data type is a determinative factor for research method selection. Another factor is the type of research (basic, applied, or R&D).
17. Knowledge, experience, personal preference, and available resources are other factors for research method selection.
18. Based on data types, research methods are categorized as qualitative, quantitative, and mixed methods.
19. Quantitative research analysis is often concerned with the relationship between variables. While qualitative analysis seeks to understand phenomena and explanation.
20. Two basic types of reasoning approaches are induction and deduction. The former is to obtain general principle from special cases primarily based on qualitative methods; the latter is to use general principles for special cases using quantitative analysis.

--- END ---