African Centre for Project Management

Post Graduate Diploma-Monitoring and Evaluation

Assignment 5

By

Kiden Betty Beneya

Adm no: ACPMPGD/124/2019

Submitted on 30th/Nov/2019

Q1. Explain the difference between data collection and data capture (10mrks)

Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. Data can be collected through the use of questionnaires, focused group discussions observation among others. On the other hand, Data capture is the process of collecting data which will be processed and used later to fulfil certain purposes. Ways of capturing data can range from high end technologies (for example synchrotron, sensor networks and computer simulation models) to low end paper instruments used in the field. (Statistics Canada, 2003)

Differences between data collection and data capture

By definition

Data collection is any process whose purpose is to acquire or assist in the acquisition of data. Collection is achieved by requesting and obtaining pertinent data from individuals or organizations via an appropriate vehicle. The data is either provided directly by the respondent (self-enumeration) or via an interviewer. Collection also includes the extraction of information from administrative sources which may require asking the respondent permission to link to administrative records. On the other hand, data capture refers to any process that converts the information provided by a respondent into electronic format. This conversion is either automated or involves staff keying the collected data (keyers).

By Nature

Data collection is active; you are getting the data by asking questions, sending a survey, or having people fill out a questionnaire. While data capture is passive; you are recording data that is being generated by an activity you didn't originate, for instance, data on visiting a particular website or data from registration books at the entrance of institutions.

Q 2: Explain the benefits of correctly interpreting data in an M&E process. (5 mrks)

Data interpretation refers to the implementation of processes through which data is reviewed for the purpose of arriving at an informed conclusion. The interpretation of data assigns a meaning to the information analyzed and determines its signification and implications.

The followings are the benefits of correctly interpreting data in an M&E process;

Informed Decision-making

First and foremost, correct interpretation of data helps project managers and other project team to have informed decision-making. A decision is only as good as the knowledge that formed it. Most decisive actions will arise only after a problem has been identified or a goal defined. Data analysis should include identification, thesis development and data collection followed by data communication. The monitoring of data results will inevitably return the process to the start with new data and sights.

Anticipating needs with trends identification.

Secondly, it helps in anticipating needs with trends identification. Data insights provide knowledge, and knowledge is power. An example of how data analysis can impact trend prediction can be evidenced in the music identification application, Shazam. The application allows users to upload an audio clip of a song they like, but can't seem to identify. Users make 15 million song identifications a day. With this data, Shazam has been instrumental in predicting future popular artist. Data gathering and interpretation processes can allow for industry-wide climate prediction and result in greater revenue streams across the market. For this reason, all institutions should follow the basic data cycle of collection, interpretation, decision making and monitoring.

Cost efficiency.

Furthermore, Proper implementation of data analysis processes have the ability to alert management to cost-reduction opportunities without any significant exertion of effort on the part of human capital. When data is collected and analyzed properly, project current and future problems can be anticipated before it actually happens and managers can take informed decisions to prevent such an

issue from arising. This would have saved the organization a whole lot of money, that would have been wasted in implementing activities that cannot help us reach the organization goals.

Clear foresight.

Furthermore, organizations that collect and analyze their data gain better knowledge about themselves, their processes and performance. They can identify performance challenges when they arise and take action to overcome them. Data interpretation through visual representations lets them process their findings faster and make better-informed decisions on the future of the company.

Q3. Explain the main concerns for a data analyst while undertaking the task of data analysis. (10 mrks)

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data (Peersman, 2014). According to Shamoo & Resnik (2003) various analytic procedures provide a way of drawing inductive inferences from data and distinguishing the signal (the phenomenon of interest) from the noise (statistical fluctuations) present in the data. The followings are the main concerns for data Analyst while undertaking the task of data analysis

Having necessary skills to analyze

A silent assumption of investigators is that they have received training sufficient to demonstrate a high standard of research practice. Unintentional ‘scientific misconduct' is likely the result of poor instruction and follow-up. A number of studies suggest this may be the case more often than believed (Nowak, 1994; Silverman, Manson, 2003). A common practice of investigators is to defer the selection of analytic procedure to a research team ‘statistician’. Ideally, investigators should have substantially more than a basic understanding of the rationale for selecting one method of analysis over another. This can allow investigators to better supervise staff who conduct the data analyses process and make informed decisions

Concurrently selecting data collection methods and appropriate analysis

While methods of analysis may differ by scientific discipline, the optimal stage for determining appropriate analytic procedures occurs early in the research process and should not be an afterthought. According to (Smeeton & Goda, 2003), "Statistical advice should be obtained at the stage of initial planning of an investigation so that, for example, the method of sampling and design of questionnaire are appropriate".

Drawing unbiased inference

The main aim of analysis is to distinguish between an event occurring as either reflecting a true effect versus a false one. Any bias occurring in the collection of the data, or selection of method of analysis, will increase the likelihood of drawing a biased inference. Bias can occur when recruitment of study participants falls below minimum number required to demonstrate statistical power or failure to maintain a sufficient follow-up period needed to demonstrate an effect (Altman, 2001).

Inappropriate subgroup analysis

When failing to demonstrate statistically different levels between treatment groups, investigators may resort to breaking down the analysis to smaller and smaller subgroups in order to find a difference. Although this practice may not inherently be unethical, these analyses should be proposed before beginning the study even if the intent is exploratory in nature. If it the study is exploratory in nature, the investigator should make this explicit so that readers understand that the research is more of a hunting expedition rather than being primarily theory driven.

Following acceptable norms for disciplines

Every field of study has developed its accepted practices for data analysis. (David, 2000) states that it is prudent for investigators to follow these accepted norms. Resnik further states that the norms are based on two factors that is to say, the nature of the variables used (i.e., quantitative, comparative, or qualitative) and assumptions about the population from which the data are drawn (i.e., random distribution, independence, sample size, etc.). If one uses unconventional norms, it is

crucial to clearly state this is being done, and to show how this new and possibly unaccepted method of analysis is being used, as well as how it differs from other more traditional methods.

Determining significance

While the conventional practice is to establish a standard of acceptability for statistical significance, with certain disciplines, it may also be appropriate to discuss whether attaining statistical significance has a true practical meaning, i.e., 'clinical significance'. Jeans (1992) defines 'clinical significance' as "the potential for research findings to make a real and important difference to clients or clinical practice, to health status or to any other problem identified as a relevant priority for the discipline." Kendall and Grove (1988) define clinical significance in terms of what happens when troubled and disordered clients are now, after treatment, not distinguishable from a meaningful and representative non-disturbed reference group." Thompson and Noferi (2002) suggest that readers of counseling literature should expect authors to report either practical or clinical significance indices, or both, within their research reports. Shepard (2003) questions why some authors fail to point out that the magnitude of observed changes may be too small to have any clinical or practical significance, "sometimes, a supposed change may be described in some detail, but the investigator fails to disclose that the trend is not statistically significant".

Lack of clearly defined and objective outcome measurements

No amount of statistical analysis, regardless of the level of the sophistication, will correct poorly defined objective outcome measurements. Whether done unintentionally or by design, this practice increases the likelihood of clouding the interpretation of findings, thus potentially misleading readers.

Provide honest and accurate analysis

The basis for this issue is the urgency of reducing the likelihood of statistical error. Common challenges include the exclusion of outliers, filling in missing data, altering or otherwise changing data, data mining, and developing graphical representations of the data (Shamoo & Resnik, 2003).

Manner of presenting data

At times investigators may enhance the impression of a significant finding by determining how to present derived data (as opposed to data in its raw form), which portion of the data is shown, why, how and to whom (Shamoo & Resnik, 2003). Nowak (1994) notes that even experts do not agree in distinguishing between analyzing and massaging data. Shamoo (1989) recommends that investigators maintain a sufficient and accurate paper trail of how data was manipulated for future review.

Environmental/contextual issues

The integrity of data analysis can be compromised by the environment or context in which data was collected that is to say. face-to face interviews vs. focused group. The interaction occurring within a dyadic relationship (interviewer-interviewee) differs from the group dynamic occurring within a focus group because of the number of participants, and how they react to each other's responses. Since the data collection process could be influenced by the environment/context, researchers should take this into account when conducting data analysis.

Data recording method

Analyses could also be influenced by the method in which data was recorded. For example, research events could be documented by; recording audio and/or video and transcribing later, either a researcher or self-administered survey, either closed ended survey or open ended survey, preparing ethnographic field notes from a participant/observer and requesting that participants themselves take notes, compile and submit them to researchers.

While each methodology employed has rationale and advantages, issues of objectivity and subjectivity may be raised when data is analyzed.

Partitioning the text

During content analysis, staff researchers or 'raters' may use inconsistent strategies in analyzing text material. Some 'raters' may analyze comments as a whole while others may prefer to dissect text material by separating words, phrases, clauses, sentences or groups of sentences. Every effort should be made to reduce or eliminate inconsistencies between "raters" so that data integrity is not compromised.

Training of Staff conducting analyses

A major challenge to data integrity could occur with the unmonitored supervision of inductive techniques. Content analysis requires raters to assign topics to text material (comments). The threat to integrity may arise when raters have received inconsistent training, or may have received previous training experience(s). Previous experience may affect how raters perceive the material or even perceive the nature of the analyses to be conducted. Thus one rater could assign topics or codes to material that is significantly different from another rater. Strategies to address this would include clearly stating a list of analyses procedures in the protocol manual, consistent training, and routine monitoring of raters.

Reliability and Validity

Researchers performing analysis on either quantitative or qualitative analyses should be aware of challenges to reliability and validity. For example, in the area of content analysis, (Louis August, 1995) identifies three factors that can affect the reliability of analyzed data;

stability, or the tendency for coders to consistently re-code the same data in the same way over a period of time

reproducibility, or the tendency for a group of coders to classify categories membership in the same way

accuracy, or the extent to which the classification of a text corresponds to a standard or norm statistically

The potential for compromising data integrity arises when researchers cannot consistently demonstrate stability, reproducibility, or accuracy of data analysis (Alan, Ashish, & David, 2005)

Extent of analysis

Upon coding text material for content analysis, raters must classify each code into an appropriate category of a cross-reference matrix. Relying on computer software to determine a frequency or word count can lead to inaccuracies. “One may obtain an accurate count of that word's occurrence and frequency, but not have an accurate accounting of the meaning inherent in each particular usage” (Louis August, 1995). Further analyses might be appropriate to discover the dimensionality of the data set or identity new meaningful underlying variables.

Q4. Describe key measures that are mandatory for data quality assurance at program level and explain the value of data quality assurance. (15 mrks).

Data quality assurance is the process of data profiling to discover inconsistencies and other irregularities in the data, as well as performing data cleansing activities (for example; removing outliers, missing data interpolation) to improve the data quality. (Statistics of Canada, 2009). The followings are the key measures that are mandatory for data quality assurance at program level

Rigorous data profiling and control of incoming data

In most cases, bad data comes from data receiving. In an organization, the data usually comes from other sources outside the control of the department or Organization. It could be the data sent from another organization, or, in many cases, collected by third-party software. Therefore, its data quality cannot be guaranteed, and a rigorous data quality control of incoming data is perhaps the most important aspect among all data quality control tasks. It is also essential to automate the data profiling and data quality alerts so that the quality of incoming data is consistently controlled and managed whenever it is received. (Lui, 1997)

Careful data pipeline design to avoid duplicate data

Duplicate data refers to when the whole or part of data is created from the same data source, using the same logic, but by different people or teams likely for different downstream purposes. When a duplicate data is created, it is very likely that it will lead to different results, with cascading effects throughout multiple systems or databases. At the end, when a data issue arises, it becomes difficult or time-consuming to trace the root cause, not to mention fixing it. Therefore, in order for an organization to prevent this from happening, a data pipeline needs to be clearly defined and carefully designed in areas including data assets, data modeling among others.

Accurate gathering of data requirements

An important aspect of having good data quality is to satisfy the requirements and deliver the data to clients and users for what the data is intended for. It is not easy to properly present the data. Truly understanding what a client is looking for requires thorough data discoveries, data analysis, and clear communications, often via data examples and visualizations. therefore, the requirement should capture all data conditions and scenarios. (Peersman, 2014)

Enforcement of data integrity

An important feature of the relational database is the ability to enforce data Integrity using techniques such as foreign keys, check constraints, and triggers. When the data volume grows, along with more and more data sources and deliverables, not all datasets can live in a single database system. The referential integrity of the data, therefore, needs to be enforced by applications and processes, which need to be defined by best practices of data governance and included in the design for implementation. In today's big data world, referential enforcement has become more and more difficult. Without the mindset of enforcing integrity in the first place, the referenced data could become out of date, incomplete or delayed, which then leads to serious data quality issues. (Brackstone, 1999)

Integration of data lineage traceability into the data pipelines

For a well-designed data pipeline, the time to troubleshoot a data issue should not increase with the complexity of the system or the volume of the data. Without the data lineage traceability built into the pipeline, when a data issue happens, it could take hours or days to track down the cause. Sometimes it could go through multiple teams and require data engineers to look into the code to investigate. data traceability lays the foundation for further improving data quality reports and

dashboards that enables one to find out data issues earlier before the data is delivered to clients or internal users.

Automated regression testing as part of change management

Obviously, data quality issues often occur when a new dataset is introduced or an existing dataset is modified. For effective change management, test plans should be built with two themes, that is to say, confirming the change meets the requirement and ensuring the change does not have an unintentional impact on the data in the pipelines that should not be changed. For mission critical datasets, when a change happens, regular regression testing should be implemented for every deliverable and comparisons should be done for every field and every row of a dataset. With the rapid progress of technologies in big data, system migration constantly happens in a few years. Automated regression test with thorough data comparisons is a must to make sure good data quality is maintained consistently.

Capable data quality control teams

Lastly but not the least, two types of teams play critical roles to ensure high data quality for an organization;

Quality Assurance team: This team checks the quality of software and programs whenever changes happen. Rigorous change management performed by this team is essential to ensure data quality in an organization that undergoes fast transformations and changes with data-intensive applications.

Production Quality Control team: Depending on an organization, this team does not have to be a separate team by itself. Sometime it can be a function of the Quality Assurance or Business Analyst team. The team needs to have a good understanding of the business rules and business requirements, and be equipped by the tools and dashboards to detect abnormalities, outliers, broken trends and any other unusual scenarios that happen on Production. The objective of this team is to identify any data quality issue and have it fixed before users and clients do.

These activities can be undertaken as part of data warehousing or as part of the database administration of an existing piece of application software.

Therefore, these steps or key measures helps to ensure data Validity, integrity, reliability, precision and timeliness which are key aspects of quality data (Brackstone, 1999)

Value of data quality Assurance

Data quality is important because without high-quality data, you cannot understand or stay in contact with your clients or beneficiaries. Data Quality Assurance provides the information you need to ascertain the quality of your data and whether it meets the requirements of your project. Data Quality Assurance ensures that your data will meet defined standards of quality with a stated level of confidence. Therefore, data quality assurance is of high value to an organization since it helps you produce data of known quality, enhance the credibility of your group in reporting monitoring results, and ultimately save time and money. However, a good Data Quality Assurance program is only successful if everyone consents to follow it and if all project components are available in writing. (Data Quality Experian, 2016)

.

Q5: In about 350 words, describe the main challenges to effective data interpretation and analysis. (10 mrsk)

Top of the list is the huge volume of data available for analysis. With today's data-driven organizations and the introduction of big data, data analysts and employees are often overwhelmed with the amount of data that is collected. An organization may receive information on every incident and interaction that takes place on a daily basis, leaving analysts with thousands of interlocking data sets to consider (The MITRE Corporation, 2008). Relatedly, with so much data available, it's difficult to dig down and access the insights that are needed most in limited time. When employees are overwhelmed, they may not fully analyze data or only focus on the measures that are easiest to collect instead of those that truly add value. The next issue is the challenge of trying to analyze data across multiple, disjointed sources. Different pieces of data are often stored in different systems. Employees may not always realize this, leading to incomplete or inaccurate analysis. Manually combining data is time-consuming and can limit insights to what is easily viewed. On a more practical note, to be understood and impactful, data often needs to be visually presented in graphs or charts (Brackstone, 1999). While this is incredibly useful, it's difficult to build them manually. Furthermore, nothing is more harmful to data analytics than inaccurate data. Without good input, output will be unreliable. A key cause of inaccurate data is manual errors made during data entry. This can lead to significant negative consequences if the analysis is used to influence decisions.

From a capacity perspective, some organizations struggle with analysis due to a lack of talent. This is especially true in those without formal risk departments. Finally, analytics can be hard to scale as an organization and the amount of data it collects grows. Collecting information and creating reports becomes increasingly complex. A system that can grow with the organization is crucial to manage this issue. While overcoming these challenges may take some time, the benefits of data analysis are well worth the effort.

References

Alan, F. K., Ashish, P. S., & David, L. B. (2005). *Data Quality: A Statistical Perspective.* Washington DC: National institute of Statistical Sciences.

Brackstone, G. (1999). *Managing Data Quality in a Statistical Agency.* Ottawa: Statistics Canada.

Data Quality Experian. (2016). *Global Data Management-Benchmark Report.* Nottingham: Experian.

David, R. B. (2000). *Apragmatic approach to the Demarcation problem.* London: Resnik B. David.

Louis August, G. (1995). *Content Analysis of Verbal Behavior.* Washington DC: American Psychological Association.

Lui, X. (1997). *Intelligent Data Analysis: Issues and Challenges.* london: International Institute for Advanced Studies and Cybernetics.

Munch, J. (2007). *Effective Data Interpretation.* Berlin: Jurgen Munch.

Peersman, G. (2014). *Data Collection and Analysis Methods in Impact Evaluation.* Florence: UNICEF OFFICE OF RESEARCH.

Shamoo, E. A., & Resnik, B. D. (2003). *Responsible Conduct of Research.* London: Oxford University Press.

Smeeton, N., & Goda, D. (2003). Conducting and Presenting Social work Research: some Basic Statistical Considerations. *British Journal of Social work*, 567-573.

Statistics Canada. (2003). *Survey Methods and practices.* Ottawa: Statistics Canada.

Statistics of Canada. (2009). *Quality Guidline.* ottawa: Statistics of Canada.

The MITRE Corporation. (2008). *Data Analysis Challenges.* Richmond: JASON, The MITRE Corporation.