**Module 5 Questions:**

**Registration number: Aipms 262/2019**

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

The **Data Collection**is a process by which the researcher collects the information from all the relevant sources to find answers to the research problem, test the hypothesis and evaluate the outcomes.

Broadly, the data collection methods can be classified into two categories:

**Primary Data Collection Methods**

The primary data are the first-hand data, collected by the researcher for the first time and is original in nature. The researcher collects the fresh data when the research problem is unique, and any other person does no related research work. The results of the research are more accurate when the data is collected directly by the researcher but it is costly and time-consuming.

**Secondary Data Collection Methods**

When the data is collected by someone else for his research work and has already passed through the statistical analysis is called the secondary data. Thus, the secondary data is the second-hand data, which is readily available from the other sources. One of the advantages of using the secondary data is that it is less expensive and at the same time easily available, but the authenticity of the findings can be questioned.

**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 (e.g. Synchrotron, sensor networks and computer simulation models) to low end paper instruments used in the field. Data with good metadata attached at the point of capture can expedite data sharing, publishing and citation.

**The method for capturing data are as follows;**

**Optical Character Recognition (OCR)**

OCR technology is used to capture data from structured documents, usually those that have been word-processed. The software works by converting documents into machine-readable files, once this has happened can search by keywords contained within it, great for files containing large amounts of data.

**Intelligent Character Recognition (ICR)**

ICR technology is an adapted version of OCR, the difference being that ICR is able to read handwritten text and convert it into computer readable information. Often used for unstructured documents such as letters, unstructured documents and other handwritten business correspondence.

**Automatic Data Capture**

Once documents have been OCR or ICR scanned, automatic data capture software can identify and extract key information from forms. This is becoming extremely popular for businesses looking to automate processing tasks, such as invoicing, purchasing and claims processing. The software is trained to look for specific types of information, such as reference numbers, names and addresses. Extracted information is output into a preferred format for import into existing systems, commonly used are csv, excel or html.

**Paperless Forms**

[Paperless forms](https://www.ukdataentry.com/data-entry-form-design-and-print/paperless-forms-design/) allow information to be captured whilst out in the field, transforming the way data comes in and out of a business. Data is captured using a mobile device and can be transferred straight into office systems, with no paper processing involved or delays.

**Barcode Technology**

Barcode technology is a data capture method that allows metadata such as customer name, address and contact numbers to be pre populated into barcode format prior to forms being sent. This form of data capture significantly reduces manual data entry requirement upon return.

**Double Blind Data Entry**

Double blind data entry is a form of manual data entry using two operators and validation software to increase accuracy. The software will flag up any discrepancies between both sets of data and errors must be corrected before they can move.

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

Data interpretation, in the end, helps improve processes and identify problems. It is difficult to grow and make dependable improvements without, at the very least, minimal data collection and interpretation. Vague ideas regarding performance enhancement exist within all institutions and industries. Yet, without proper research and analysis, an idea is likely to remain in a stagnant state forever (i.e., minimal growth). So there are a few of the benefit of data interpretation provided below;

**Informed decision-making**

A decision is only as good as the knowledge that formed it. Informed data decision making has the potential to set industry leaders apart from the rest of the market pack. [Studies have shown](https://hbr.org/2012/10/big-data-the-management-revolution/ar) that companies in the top third of their industries are, on average, 5% more productive and 6% more profitable when implementing informed data decision-making processes. 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.

If institutions only follow that simple order, then it will be easy to solve issues as they emerge in real time. Informed decision making has a tendency to be cyclical. This means there is really no end, and eventually, new questions and conditions arise within the process that need to be studied further. The monitoring of data results will inevitably return the process to the start with new data and sights.

**Anticipating needs with trends identification**

Data insights provide knowledge, and knowledge is power. The insights obtained from market and consumer data analyses have the ability to set trends for peers within similar market segments. A perfect 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 artists.

When industry trends are identified, they can then serve a greater industry purpose. For example, the insights from Shazam’s monitoring benefits not only Shazam in understanding how to meet consumer needs, but it grants music executives and record label companies an insight into the pop-culture scene of the day. 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**

Proper implementation of data analysis processes can provide businesses with profound cost advantages within their industries. A recent data study performed by [Deloitte](http://www2.deloitte.com/content/dam/Deloitte/global/Documents/Deloitte-Analytics/dttl-analytics-analytics-advantage-report-061913.pdf) vividly demonstrates this in finding that data analysis Return On Investment is driven by efficient cost reductions. Often, this benefit is overlooked because making money is typically viewed as “sexier” than saving money. Yet, sound data analyses have the ability to alert management to cost-reduction opportunities without any significant exertion of effort on the part of human capital.

**Clear foresight**

Companies 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)

The main concerns for data analyst while undertaking the task of data analysis are not limited to the below;

**Exploratory analysis**

Need to check for missing data and maybe stop analysis if proportion of missing data is too high. Check for high leverage points as pollution data tends to be skewed. Maybe log-transform and check for other outliers and note them for later (we may want to do a sensitivity analysis without those observations).

**Model fitting**

If the outcome is a count, then typically a Poisson regression model is used. We already know that maximum likelihood is an excellent approach and better than most others under reasonable circumstances. There is plenty of GLM software out there so we don’t even have to program the IRLS algorithm.

**Model building**

Since this is not a prediction model, the main concern data analyst have is that of proper adjusted for measured and unmeasured confounding. [Francesca Dominici](http://www.hsph.harvard.edu/francesca-dominici/) and some of her colleagues have done [some](http://www.ncbi.nlm.nih.gov/pubmed/18552590) [interesting](http://www.ncbi.nlm.nih.gov/pubmed/22364439) [work](http://www.tandfonline.com/doi/abs/10.1198/016214504000000656#.Ucye6BbHKZY) regarding how best to do this via Bayesian model averaging and other approaches. That said, I think simpler versions of the “ideal approach” can be easily implemented.

**Sensitivity analysis**

There are a number of key sensitivity analyses that need to be done in all time series analyses. If there were outliers during EDA, maybe re-run model fit and see if regression coefficient changes much. For time series models, unmeasured temporal confounding is a big issue so this is usually checked using spline smoothers on the time variable with different degrees of freedom. This can be automated by fitting the model many different times with different degrees of freedom in the spline.

**Reporting**

Typically, data analyst are concern about the reporting format before analysing its data. Some summary statistics for the data are reported along with the estimate + confidence interval. Estimates from the sensitivity analysis should be reported, and perhaps estimates from different lags of exposure, if that is a question of interest. It is slightly more complicated if one have a multi-city study.

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

The key mandatory measures for data quality assurance at program level are not limited to the following;

### The ratio of data to errors

This is the most obvious measure of data quality. It allows for tracking how the number of known errors such as missing, incomplete or redundant entries within a data set corresponds to the size of the data set. If it find fewer errors while the size of the data stays the same or grows, know that the data quality is improving.

### Number of empty values

Empty values, which usually indicate that information was missing or recorded in the wrong field within a data set, are an easy way to track this type of data quality problem. One can quantify how many empty fields are within a data set, then monitor how the number changes over time.

### Data transformation error rates

Problems with data transformation, the process of taking data that is stored in one format and converting it to a different format are often a sign of data quality problems. By measuring the number of data transformation operations that fail (or take unacceptably long to complete) one can gain insight into the overall quality of the data.

### Amounts of dark data

[Dark data](https://blog.syncsort.com/2015/01/big-data/dark-data-deal/?utm_source=Blog-Post&utm_medium=Blog-In-Text) is data that cannot be used effectively, often because of data quality problems. The more dark data one have, the more data quality problems one probably have.

### Email bounce rates

If running a marketing campaign, poor data quality is one of the most common causes of email bounces. They happen because errors, missing data or outdated data cause one to send emails to the wrong addresses.

### Data storage costs

[Data storage](https://blog.syncsort.com/2017/07/big-data/data-storage-best-practices-data-types/?utm_source=Blog-Post&utm_medium=Blog-In-Text) costs rises while the amount of data that is actually use stays the same. This is another possible sign of data quality issues. If one is storing data without using it, it could be because the data has quality problems. If, conversely, storage costs decline while data operations stay the same or grow, there is likely improving data quality front.

### Data time-to-value

Calculating how long it takes team to derive results from a given data set is another way to measure data quality. While a number of factors (such as how automated data transformation tools are) affect data time-to-value, data quality problems are one common hiccup that slows efforts to derive valuable information from data. The most important thing is to have some kind of data quality assessment plan in place, whatever its details may be.

There are several values of data quality; completeness, consistency, accuracy, validity, and timeliness. When each of these values are properly executed, it will result in high-quality data. It is also imperative that everyone who uses the data collected has a general understanding of what the data represents. The extent of a data initiative is not limited to the data produced by the company’s own research, it must include data obtained from external sources as well. High-quality data will ensure more efficiency in driving a company’s success because of the dependence on fact-based decisions, instead of habitual or human intuition.

**Completeness:**Ensuring there are no gaps in the data from what was supposed to be collected and what was actually collected.

This can be resolved by halting submission if the data is not complete. With Paper and Pencil Interviewing (PAPI), this can be exceptionally difficult as this method is prone to human error. On the other hand, the Computer Assisted Personal Interviewing (CAPI/electronic) method uses smartphones and tablets that allow the same data collection but the data is recorded on a device instead of paper. By using the mandatory fields function, data completeness is easily achievable. The respondent will not be able to complete and submit the data without the mandatory fields being filled. This will also ensure less time being wasted fixing mistakes resulting from incomplete data.

**Consistency:**The types of data must align with the expected versions of the data being collected. This can be ensured by using the drop down menus in a data collection application, which will result in data that is consistently collected in the expected format. Instead of free-form writing, there are predetermined numbers of options from which to choose. There will be consistency across the board and allow for complete search results.

**Accuracy:**Data collected is correct, relevant and accurately represents what it should. Accuracy is more challenging to remedy than data completeness and consistency. Accurate data is often the result of trained and competent employees. However, there is stillroom for human error. In order to reduce the likelihood of inaccuracies, it is vital to implement extra measures like adding picture capture, GPS location and time stamps to recorded events.

**Validity:**Validity is derived from the process instead of the result. When there is a need to fix invalid data, more often than not, there is an issue with the process rather than the results. This makes it a little trickier to resolve. Paper-based methods are more difficult to fix when it comes to issues of invalid data because changing forms can be expensive, wasteful and the more widespread the company is, the harder it is to change. With the electronic (CAPI) option, it will immediately take seconds to implement the change company-wide as all the data is collected on an electronic device. There are no old surveys to throw out or extra printing costs. And certainly, no one left using an old version of the survey.

**Timeliness:** The data should be received at the expected time in order for the information to be utilized efficiently. Anything slower becomes an inadequate source of information. With real time data and analytics, companies are better equipped to make more effective and informed decisions. There is a pressing need to eliminate the lag time between when a survey is completed in the field and when it is received.

Electronic methods allow field employees to collect the same data they would on paper but it would be safely recorded on a smartphone or tablet upon completion and then instantly submitted to the database. Another way to achieve timeliness is to employ Dattel Asia, ASEAN’s leading face-to-face [data collection service](http://www.dattel.asia/data-collection/) provider that utilizes tablets, digital tools and artificial machine learning systems to collect the true voice of the respondents across a vast urban and rural gamut.

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

**Handling Enormous Data in Less Time:**

Handling the data of any business or industry is itself a significant challenge, but when it comes to handling enormous data, the task gets much more difficult. Critical business decisions should be taken effectively, but we need to have strong IT infrastructure, which is capable of reading the data faster and delivering real-time insights.

**Visual Representation of Data:**

Another important task is the visual representation of data. One need to represent the data in an easy format that makes it readable and understandable to the audience. Handling an unstructured data and then representing in a visually attractive manner could be a difficult task.

**Application of Scalable data:**

The major factor to consider is the scalability factor of the applications. Several organizations are facing the same issue where the volume of data has been increasing each passing day. Due to the multiple layers between the databases and front-end, the data traversal takes time.

**Limited Sample Size**

Contrary to quantitative data where often have a great amount of data available, is sample size one of the challenges of qualitative data.

If browse on the internet, its find out there is no general agreement on the ideal sample size for qualitative research. It is very costly to perform extensive qualitative research with hundreds of participants.

**Sampling Bias**

“In statistics, sampling bias is a bias in which a sample is collected in such a way that some members of the intended population are less likely to be included than others.” In other words, qualitative sample will never include a representative overview of all the different people that come to the website.

**Self-Selection Bias**

Do you like to participate in surveys? A few of you might say “Yes” and others think “not”.

This is the exact problem here. It’s a free choice to participate in a research study or not.

On the other side, quantitative data is gathered from most people whether they like it or not.

Just sign up for hot jar, set up a heat map and the data will be collected.

**Hawthorne Effect**

“Participants in behavioral studies change their behavior or performance in response to being observed.” For example, one`s opinion about a particular website might be different when one know they are being observed if compared to when one don’t know they are being observed.

**Observer-Expectancy Effect**

Let’s say one is running a survey and function as an observer in the research room. And walking around and observe the participants. It is known that researcher’s beliefs or expectations causes him or her to uncon­sciously influence the participants of an experiment.

**Artificial Scenario**

Most experiments include pre-set goals in a specific environment. And can’t get feedback on things one don’t ask. For example, run an experiment for an ecommerce website, the goal is to find out whether the form (where people leave their personal information) functions well or if anything needs to be improved.