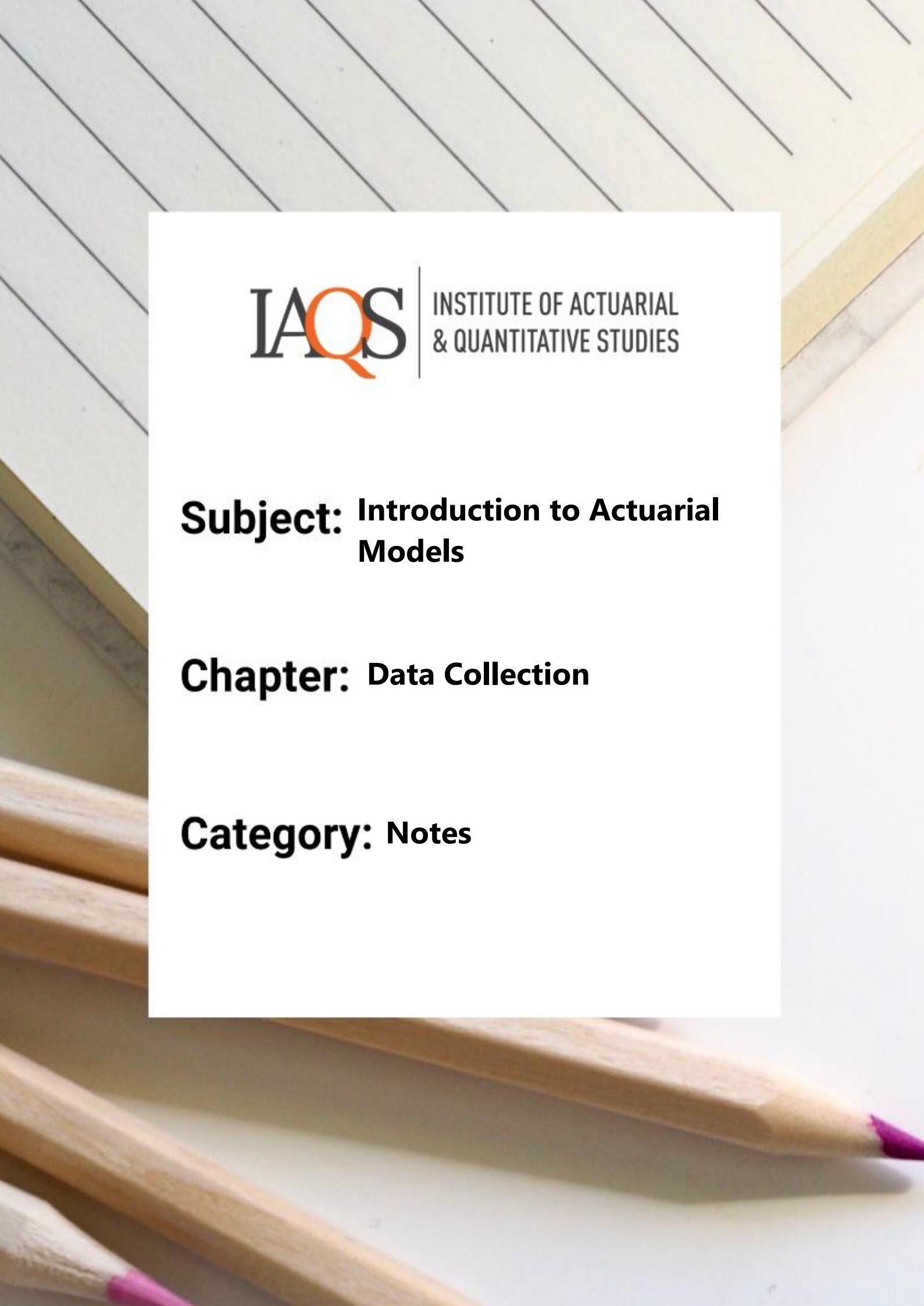


INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: Introduction to Actuarial Models

Chapter: Data Collection

Category: Notes

Introduction

Collecting Good Data

This section lays out some general principles for collecting data for construction of models. Using well-planned data collection procedures is often the difference between successful and unsuccessful experiments.

In addition, well designed experiments are often less expensive than those that are less well thought-out, regardless of overall success or failure.

Specifically, this section will answer the question:

What can the analyst do even prior to collecting the data (that is, at the experimental design stage) that would allow the analyst to do an optimal job of modeling the data?

After the data is collected, what steps need to be taken to prepare the data?

Design of Experiments (DOE)

Systematic Approach to Data Collection

Data for statistical studies are obtained by conducting either experiments or surveys. Experimental design or DOE is the branch of statistics that deals with the design and analysis of experiments.

Design of experiments (DOE) is a systematic, rigorous approach to problem-solving that applies principles and techniques at the data collection stage so as to ensure the generation of valid, defensible, and supportable conclusions. In addition, all of this is carried out under the constraint of a minimal expenditure of time, and money.

What is Experimental Design?

Experimental design is a way to carefully plan experiments in advance so that your results are both objective and valid.

Ideally, your experimental design should:

- Describe how participants are allocated to experimental groups. A common method is completely randomized design, where participants are assigned to groups at random. A second method is randomized block design, where participants are divided into homogeneous blocks (for example, age groups) before being randomly assigned to groups.
- Minimize or eliminate confounding variables (outside influences) which can offer alternative explanations for the experimental results.
- Allow you to make inferences about the relationship between independent variables and dependent variables.
- Reduce variability, to make it easier for you to find differences in treatment outcomes.

Design of experiments involves:

- The systematic collection of data

- A focus on the design itself, rather than the results
- Planning changes to independent (input) variables and the effect on dependent variables or response variables
- Ensuring results are valid, easily interpreted, and definitive.

The most important principles are:

- Randomization: the assignment of study components by a completely random method, like simple random sampling. Randomization eliminates bias from the results
- Replication: the experiment must be replicable by other researchers. This is usually achieved with the use of statistics like the standard error of the sample mean or confidence intervals.
- Blocking: controlling sources of variation in the experimental results

DOE Problem Areas

Main concerns in experimental design include the establishment of validity, reliability, and replicability. For example, these concerns can be partially addressed by carefully choosing the independent variable, reducing the risk of measurement error, and ensuring that the documentation of the method is sufficiently detailed.

Related concerns include achieving appropriate levels of statistical power and sensitivity.

Also, the analyst is interested in functionally modeling the data with the output being a good-fitting (= high predictive power) mathematical function, and to have good (= maximal accuracy) estimates of the coefficients in that function.

Basic steps in designing an experiment

1. Define the problem and the questions to be addressed.

Before data collection begins, specific questions that the researcher plans to examine must be clearly identified. In addition, a researcher should identify the sources of variability in the experimental conditions. One of the main goals of a designed experiment is to partition the effects of the sources of

ariability into distinct components in order to examine specific questions of interest.

2. Define the population of interest.
A population is a collective whole of people, animals, plants, or other items that researchers collect data from. Before collecting any data, it is important that researchers clearly define the population. The designed experiment should designate the population for which the problem will be examined. The entire population for which the researcher wants to draw conclusions will be the focus of the experiment.

3. Determine the need for sampling.
A sample is one of many possible sub-sets of units that are selected from the population of interest. In many data collection studies, the population of interest is assumed to be much larger in size than the sample. The results from a sample are then used to draw valid inferences about the population.

A random sample is a sub-set of units that are selected randomly from a population. A random sample represents the general population or the conditions that are selected for the experiment because the population of interest is too large to study in its entirety. Using techniques such as random selection after stratification or blocking is often preferred.

4. Define the experimental design.
A clear definition of the details of the experiment makes the desired statistical analyses possible, and almost always improves the usefulness of the results.

Defining the experiment involves identifying the variables and designing the structure.

Data Sources and Data Preparation

What is a data source?

A data source may be the initial location where data is born or where physical information is first digitized. It simply means what the words mean: where data is coming from.

Data sources can differ according to the application or the field in question.

We focus here on Primary and Secondary data. We discuss how the characteristics of the data are determined both by the primary source and the steps carried out to prepare it for analysis – which may include the steps on the journey from primary to secondary source.

The process of preparing the data for analysis.

Data preparation is the process in which data from one or more sources is cleaned and transformed to improve its quality prior to its use in data analysis. It's often used to merge different data sources with different structures and different levels of data quality into a clean, consistent format.

Data preparation is a key part of analytics.
Without data preparation, patterns and insights could be missing from the database and overlooked during analysis.
In all cases, knowledge of the details of the collection process is important for a complete understanding of the data, including possible sources of bias or inaccuracy.

Broadly the steps include:
1. Collecting primary data through the defined sources.
2. Summarize the data in a format suitable for further preparation and final analysis.
3. Study the details regarding the data collection process and the data collected itself.
4. Cleaning the data
5. Finally data is ready for analysis!

Let's look at each step.

Collecting the Primary data

Primary data is **data that is collected by a researcher from first-hand sources**, using methods like surveys, interviews, or experiments. It is collected with the research project in mind, directly from primary sources.

Primary data can be gathered as the outcome of a designed experiment. *i.e* through the DOE.

For example: The data collected by the insurance companies regarding their customers while writing a policy is a primary data for the company.
Here, the policyholders are the primary source. This data will be the refined by the company before using it for analysis.

Representing data in suitable format

There are many useful ways to present your data.

In general, large sets of numerical data should not be included in the text of a manuscript but rather summarized in tables or graphics. While tables are an excellent tool for giving structured numerical information, graphics are more appropriate for demonstrating trends and relationships or making comparisons.

Different formats that can be used to represent data.
1. Pie charts
2. Bar graphs
3. Line graphs
4. Scatter plots
5. Tables and Excel sheets etc.

Adding visual information to your paper can considerably enhance its impact and readability. So, take some time to make the right decision considering all the tools available and the story you want to tell.

Studying the data collection process and the collected data

This is a vital stage!
It is very important to understand different aspects of the data collection process in order to gain meaningful insight about the data quality, possibility of bias and inaccuracy.

Issues that the analyst should be aware of include:

1. whether the process was manual or automated;
2. limitations on the precision of the data recorded;
3. whether there was any validation at source; and
4. if data wasn't collected automatically, how was it converted to an electronic form?

For instance: scope of errors is more in manually collected data rather than an automated process.

When large sets of data is collected and sampling is done, it is important to know how the sampling is done.

The analyst should check regarding which sampling scheme is used:

- Simple random sampling

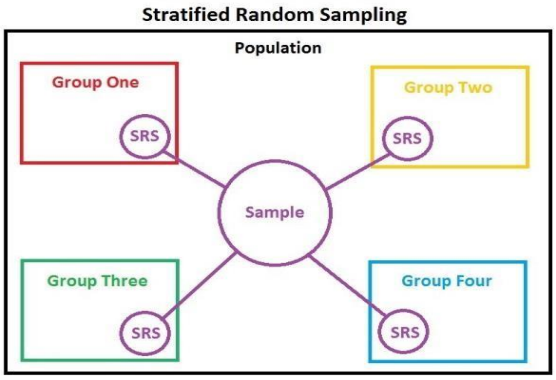
Random sampling is one of the simplest forms of collecting data from the total population. Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population.

- Stratified sampling

Stratified sampling refers to a type of sampling method.

With stratified sampling, the analyst divides the population into separate groups, called strata. Then, a probability sample (often a simple random sample) is drawn from each group.

Stratified sampling has several advantages over simple random sampling. For example, using stratified sampling, it may be possible to reduce the sample size required to achieve a given precision. Or it may be possible to increase the precision with the same sample size.



- Convenience sampling

A convenience sample is a type of non-probability sampling method where the sample is taken from a group of people easy to contact or to reach. For example, standing at a mall or a grocery store and asking people to answer questions would be an example of a convenience sample.

- Other sampling techniques:

1. Judgement sampling
2. Cluster sampling
3. Purposive sampling
4. Systematic sampling etc.

Other aspects of the data which are determined by the collection process, and which affect the way it is analysed include the following:

- ***Cross-sectional*** data involves recording values of the variables of interest for each case in the sample at a single moment in time.
- ***Longitudinal*** data involves recording values at intervals over time.
- ***Censored*** data occurs when the value of a variable is only *partially known*, for example, if a subject in a survival study withdraws, or survives beyond the end of the study: here a lower bound for the survival period is known but the exact value isn't.
- ***Truncated*** data occurs when measurements on some variables are not recorded so are completely unknown. For example: in case of claims above a benchmark which are passed to the reinsurer , the insurer only knows that the claim amount is above the benchmark however it does not know the exact amount. This is a truncated data.
- **Large or Big data.** The term *big data* is not well defined but has come to be used to describe data with characteristics that

- make it impossible to apply traditional methods of analysis A data can be classified as large data on different grounds like size- eg large observations, many attributes, speed –eg collected every second, variety – eg data from different sources in different structural variety.

When a data is extremely large, sampling is suitable.

All the above aspects need to be looked upon to understand the data well!

Cleaning the data

After understand the collection process, analyst also needs to study the data and clean it.

You really need to get to know your data before you can properly prepare it for downstream consumption.
Beyond simple visual examination, you need to profile, visualize, detect outliers, and find null values and other junk data in your data set, addressing unusual, missing or inconsistent values.

Zoom in to your data
Explore the columns you have in your data set and verify that the actual data types match the data that should be in each column. For example, a field titled "sales date" should have a value in a common data format like MM/DD/YYYY.
Similarly, you should understand the generic data type each field represents. If it's a numeric field, is it discreet or continuous? If it's a character field, is it categorical or a nominal free text field? Knowing these distinctions will help you better understand how to prep the data contained therein.

You need to read the pictures
Graphing key fields can be a great way to get to know your data. Use histograms to get a feel for the distributions of key fields, scatter plots for the all-important outlier detection etc

Iteratively cleanse and filter
Based on your knowledge of the end analytics goal, experiment with different data cleansing strategies that will get the relevant

data into a usable format. Again, start with a small, statistically-valid sample to iteratively experiment with different data preparation strategies and refine your record filters.

Data preparation is a messy but ultimately rewarding and valuable exercise. Taking the time to evaluate data sources and data sets up front will save considerable time later in the analytics project.

Data is ready for analysis.

We saw that in the data collection process, the primary source of the data is the population (or population sample) from which the 'raw' data is obtained. If, once the information is collected, cleaned and possibly summarised, it is made available for statisticians to model and analyse, and for others to use via a web interface.
This is then a secondary source of data.

Data Protection

Data security, privacy and regulation

In the design of any investigation, consideration of issues related to data security, privacy and complying with relevant regulations should be paramount.

Data security means protecting digital data, such as those in a database, from destructive forces and from the unwanted actions of unauthorized users, such as a cyberattack or a data breach.

Data privacy relates to how a piece of information—or data—should be handled based on its relative importance. Data privacy has always been important. For instance A single company may possess the personal information of millions of customers—data that it needs to keep private so that customers' identities stay as safe and protected as possible, and the company's reputation remains untarnished.

Another point to be aware of is that just because data has been made available on the internet, doesn't mean that that others are free to use it as they wish. This is a very complex area and laws vary between jurisdictions.

.