Chapter 3

Research Methodology

3.0 Summary

3.1 Data mining approach

3.2 Statistical hypothesis testing vs data mining

3.3 Ten questions questionnaire

3.4 Sampling and sample selection

3.5 Association mining using Apriori algorithm

3.0 Summary

This research work looks into the gap related to insurance need understanding and expectation of insurance service seekers from insurers. Like in any service, insurance service too has two sides – insurance service providers and insurance service seekers. The service is provided by insurers and taken by insured with various objectives. There can be financial and social objectives for insurers like earning profit, providing support to underprivileged in time of calamities and providing some cushion to reduce the impact of financial damages. Similarly, insured can have objectives like meeting regulatory requirement, risk mitigation, protection against future loss etc. However, risk understanding varies from person to person and also from time to time. Hence, such methodology has been selected that can help in investigating the role
of insurers in insurance need assessment and the likely impact on satisfaction level of the insured.

The approach used here is a combination of qualitative and quantitative techniques. However, for analysis purpose quantification has been done wherever possible. There are two types of insurance seekers – retail or individual and corporate. Their understanding about insurance and their expectations from insurers was to be determined. Also, insurers have their system of doing insurance operations. It is an important part of this research to understand the way insurers make use of information available at the insurance seekers level.

Statistical techniques for hypothesis testing are quite powerful in testing the hypothesis and helping in accepting or rejecting that. Every hypothesis testing evaluates the probability of the hypothesis to be true. If the probability is low, the hypothesis is rejected. The numerical value of high enough probability remains subjective. However, in such research approaches hypothesis has to be formulated and analysis of data speaks only about the hypothesis which is being tested. Data might have some strong patterns that go untapped if those patterns are not part of the hypothesis. Recent advancements in computational technology has made it possible to use more computation intensive techniques in analysing or mining data to reach to even the hidden patterns. Hence, in this research work, Apriori algorithm has been used to mine strong enough association rules from the data and test the hypothesis.

Insurance services have diverse interrelated aspects. It was quite complicated to structure that in the beginning of the research. Hence about 200 persons were interviewed to understand the dynamics of insurance needs, their ways of risk management, their perception about insurance options etc. Stratified random sampling technique was used to select the persons from available groups for interview representing various segments such as Individual Insurance service seekers, Software Industry, Manufacturing Organization, Chemical Industry, Transport &
Shipping Services, Agriculture and SSI etc. As there was huge variation in responses between individual insurance seekers and corporate insurance seekers, two different questionnaires were designed to collect the data. One was for individuals and the other one for corporate.

Apriori algorithm is the most common algorithm for mining association rules. It evaluates all possible associations against a given criteria of strength and outputs those associations that meet the strength criteria. Probability or p-value and the minimum support are the two criterions for the strength of association. Number of possible associations even with small number of data items runs in several thousands. Hence, TQQ (ten questions questionnaire) concept was used to design the questionnaires and restrict the number of questions in a questionnaire to 10.

To collect data from the other side of the insurance service, i.e., insurers, another TQQ was designed and used. Focus was on understanding the methods used for risk estimation, product/service design, assessment of loss, and classification of customers/services etc. and to understand the level of contribution made by Insurers as a risk management partner with the Insured.

These data were converted to a suitable format for use in data mining tool. Apriori algorithm of mining association rules requires p-value and minimum support values to find all such associations that satisfies these values. Various p-values and minimum support values were tried to get various strong and meaningful associations. This set of association rules were used to verify or test the hypothesis of this research work. The advantage of this technique is that it not only tested the hypothesis but also provided important leads for further research work.

Sampling for data collection: Random sampling technique was used for sample selection for data collection. The sample size was found by using the Slovin’s formula

\[ n = \frac{N}{1 + Nc^2} \]
Where \( n \) is the sample size, \( N \) is the population and \( c \) is the confidence level. The value of confidence level \( c \) has been taken as 0.05. This formula is quite useful in finding the sample size for large population with similar interests. Since, the respondents of the questionnaires were all insurance seekers or insurers, the sample size calculated using this formula is likely to provide a decent representation of the population.

For large population, the sample size becomes quite insensitive to changes in population and converges to 400. Hence, the sample size here has been taken as 400 from the population of insurance seekers and 400 from the population of insurers.

3.1 Data mining approach

Data mining approach, as the term indicates, is an approach to go deep into the data in search of valuable patterns. Data analysis techniques too work on data to find valuable patterns. However, the depth at which the data mining techniques work make them significantly different than the data analysis techniques. The entire approach and the kind of outcomes one can expect changes. Going deep into data is synonymous to going to the hidden level.

In data analysis techniques, various data analysis requirements are framed and the analysis is done accordingly. For example, if relationship between two variables are to be verified, this is considered as a requirement. Data analysis techniques can verify the relationship and can provide the output as whether such relationship exists or not. If more numbers of such relationships are to be verified, data analysis techniques can do that and provide the outputs accordingly based on the data. These techniques are quite rich and have been in use by researchers, analysts and managers for quite long. But, the responsibility of identifying the relationships to be verified rests on the researchers and managers. They can use data analysis techniques to verify any number of relationships that come to their mind based on their
experience or review of literature or by observation or by guess work. But in this process, the end result of the research or decision making depends heavily on their experience or intuition or guess. Mathematically, number of possible relationships explodes. Number of subsets of a set with n elements is equal to $2^n$. For a value of n as low as 10, this number exceeds 1000. The number of such relations that happening of a subset may be related to happening of another subset with some p-value, goes in several thousands. There is a high possibility of some important relationships that may be existing in the data but not identified because the researcher or the decision maker could not visualise that as a possibility. Some of such relationships may be valuable for the study but they get missed.

Data mining approach takes care of it. In a typical research, researcher is interested in the formulated hypothesis based on the objective of the research. This testing can anyway be done using the data analysis techniques. Data mining techniques open the scope of discovery of other valuable patterns too, if they exist in the system being researched. The algorithms used in these techniques explore all the possible combinations and provide as output all the strong enough relations among the variables. So, even if a relationship is not being evaluated but it is present in the system supported by the data, these techniques will surface them. This is somewhat similar to going deep into the data up to the hidden level and taking out the valuable patterns, something like mining.

There are mainly three techniques in data mining – clustering, classification and association. For each of these three techniques, there are several algorithms that make the computation efficient in different situations.

Clustering techniques look for clusters in data based on some similarity. The criterion for similarity has to be defined by the researcher. Clustering techniques takes the opposite of similarity as distance. It clusters the entire data in required number of clusters in such a way
that the distance between data elements within a cluster gets minimized. Data elements lying in the same cluster have strong similarity on the criterion used in clustering. Hence, a researcher can use this technique to test the hypothesis or the null hypothesis by checking the clusters in which the variables of interest are lying after clustering. This technique requires the definition of distance or similarity and the number of required clusters from the researcher. Results vary depending on the number of clusters formed. If the number of clusters given for clustering is low, then even the distant data items may get clustered in same cluster. If the number of clusters for clustering is high then the two data items need to be closely related to be in the same cluster. If number of clusters is kept as two and still two data items get clustered in different clusters, then they are very poorly related or practically not related. This matches with the concept of testing a null hypothesis.

Since, the clustering technique minimizes the distance among data items in clusters, the result depends on the data, definition of distance and the number of clusters. If some new data are included or some of the existing data are removed, then the clusters of some of the already clustered data may also change. Thus, the conclusion derived in research using this technique is based on the entire data and is quite scientific. However, the researcher needs to be aware about the impact of number of clusters on the strength of relationships among values in a cluster.

The other important data mining technique that is now being widely used is Classification. In this, the data entities are classified in different classes based on classification rules. The number of classes, the variables used for classification, the number of possible values of these variables and the rules framed on the basis of the combinations of these variable values decide the output of classification exercise. It can be concluded that the entities in the same class are similar and closely related. But, how much similar and how much close depends on the number of classes
and classification rules. It is difficult to quantify the strength of similarity here in such a way that it remains consistent irrespective of number of classes and the rules for classification.

Output of classification and clustering look similar. In both the cases, data entities are grouped in some pre-decided number of groups. Entities lying in same group are considered to be more closely related than the entities lying in different groups. However, the class of an entity is decided based on the classification rules. As long as the classification rules remain same and the value of the variables defining the entity doesn’t change, the entity remains in the same class. It doesn’t affect the class of already classified entities if more entities join the system or some leave the system. Contrary to this, in clustering, cluster of an already clustered entity may change with entry or exit of some entities even if the definition of distance and number of clusters are unchanged. This makes the selection of appropriate technique somewhat tricky.

The third main data mining technique is mining association rules. It identifies all the strong relationships from the data. It goes through the response data or the transaction data and evaluates the probability of something happening if something else has happened. If the probability is high enough and there are sufficient number of responses, then the two happenings are considered strongly associated. The algorithms used for mining association rules evaluate all possible combinations of happenings and provide as output all the strong enough associations meeting the criterion of required probability and required minimum number of incidences. The user of such technique has to specify the threshold probability and the threshold support value. If X has happened at least the threshold support value number of times and if at least the threshold probability fraction of times Y also happened whenever X happened, then X and Y are strongly associated with each other.

This technique is quite useful in research as it can provide all the relationships from the data that are strong enough and meets the minimum threshold support criteria. Researcher provides
the required p-value to consider a relationship strong. He also provides the minimum number of responses for a variable to consider that sufficient enough. Researcher can even experiment with these values and see the impact on the kind of associations coming as output. Thus, it checks the strength of the hypothesis and also takes care that rare combinations don’t influence the result. Since, it checks for all possible combinations, the number of relationships coming as output may be somewhat more than expected. One can check whether the hypothesis being tested comes as an output or not.

Thus, data mining techniques go much deeper into the data than the data analysis techniques and provide a holistic view of system being studied. They can not only be powerful tool in research but they are also capable of discovering some interesting and valuable patterns otherwise hidden. These patterns can be of interest in further research.

3.2 Statistical hypothesis testing vs data mining

Statistical hypothesis testing is about selecting the right statistical technique depending on the variables involved in the hypothesis and their dependencies, and applying the technique to test the strength of the hypothesis. Also, the conditions for accepting or rejecting the hypothesis is pre-decided based on the nature of research area. Statistical techniques are applied to compute the variables used in the acceptance criteria and based on their values the hypothesis is accepted or rejected depending on whether the acceptance criterion got met or not. These statistical techniques don’t do any investigation or exploration beyond the hypothesis. There are huge number of statistical techniques and picking the right technique for hypothesis testing can be a challenge sometimes.

Data mining techniques on the other hand focus on variable in the data rather than variables in the hypothesis. These techniques are much lesser in number compared to statistical techniques.
Selection of suitable data mining technique for research purpose depends on the nature of expected outcome. These techniques don’t focus only on hypothesis but explore all possibilities. Hypothesis is accepted or rejected depending on whether it comes as an output of exploration or not. These techniques are highly computation intensive. The computation load goes up exponentially with the number of variables. But they offer advantage in terms of discovering interesting, valuable and hidden patterns from the data. Since, data mining techniques are quite less in number, one can be very accurate in selecting right techniques for the required research purpose.

There are mainly three data mining techniques that belong purely to data mining. They are popularly known as Clustering, Association and Classification. Out of these, clustering and association techniques can be used in hypothesis testing.

Clustering technique forms clusters from data based on some definition of similarities. Variables coming in the same cluster after clustering of data are considered to be close to each other. This can be used in accepting or rejecting the hypothesis based on whether the variables of the hypothesis were placed in the same or different cluster after applying clustering technique. The strength of closeness or similarity in a cluster can be varied by changing the number of clusters. More the number of clusters, more the closeness required to be in the same cluster.

Mining association rules gives strong enough associations between possible combinations of situations by checking the threshold support and threshold probability criterion. This too can be used in hypothesis testing by selecting the support and probability criterion as per the nature of the research and checking whether the relations formulated in the hypothesis come as the strong enough association or not. Association mining has an advantage that the required strength of association to consider any association strong enough can be set by the researcher.
A researcher with not good enough statistical background too can safely choose association mining and experiment with values of threshold probability to explore his research area.

While selecting a statistical technique for hypothesis testing, a researcher has to look into the number of variables, whether distinction is made between dependent and independent variables, how the independent variables are to be treated etc. to decide the appropriate technique for his research analytics.

There are several statistical techniques that can be useful in doing research. A researcher can start with identifying the number of variables in the problem. In simplest of the situation where there is only one variable, one can choose the way that variable is to be treated with respect to the scale of measurement. If the scale of measurement is nominal, then finding mode is the appropriate technique to understand the central tendency. Looking at the relative frequency of the modal value can give the required idea about the dispersion of a single nominal variable. But, if the scale of measurement is ordinal, then median is the right measure to understand central tendency and the inter-quartile deviation is right for understanding the dispersion. If the scale of measurement is interval, then the researcher may be interested in knowing about the symmetry, dispersion, central tendency, peakedness, frequencies, normality etc. of the variable. In such case, skewness is a measure of symmetry. If it is symmetric, then mean tells about the central tendency but one will need mean and median both to understand the central tendency for skewed variable. To understand the peakedness in the variable, one has to compute the kurtosis. There are different tests for normality.

If there are two variables in the problem of research, then there can be six different possibilities related to their scale of measurement. It may be that both the variables are either nominal or ordinal or interval. Or, it may be that one is nominal and one is ordinal, or one may be ordinal and one interval, or one may be nominal and one interval. There are quite a few more factors
that need to be considered in choosing the right statistical technique for analysis. Is there
distinction being made between variables such as dependent and independent variables? Is the
relationship linear or something else? Are the nominal variables on two-point scale? What
needs to be studied – symmetry or covariance? There are several such questions that need to
be answered to finally choose the appropriate technique. It is not necessary that in all the cases
there are techniques to meet all the requirements accurately.

Sometimes, some approximation may be required. Similarly, if there are more than two
variables then again, the number of dependent variables, number of independent variables, their
scale of measurements, their relationships etc. will affect the selection of technique to be used.
It can be simple techniques like mean, median, mode, variance, relative frequencies, quartiles,
correlation, regression, chi square test, t-test, z-test etc. that can solve the purpose. Or, some of
the techniques like Anova (analysis of variance) or Levene’s W or some advanced statistical
techniques may be needed to meet the requirements.

The objective of this research and the formulated hypothesis to meet these objectives in the
research were such that the number of variables were quite high. Research was conducted
related to risk management in insurance services with focus on insurance needs and
expectations. Also, it was preferred to be in exploratory mode and remain open to some
interesting discoveries that can excite future researchers to further investigate in this area.
Hence, mining association rule, a technique of data mining was used to test the hypothesis.

3.3 Ten questions questionnaire

Questionnaire technique is perhaps the most common technique used for data collection by
researchers when the number of respondents are high and are demographically spread over.
Ten questions questionnaire (TQQ) is also a kind of questionnaire with an additional speciality
that such questionnaires have exactly 10 questions. Ten questions questionnaire offers quite a few advantages in terms of ease in getting responses, better accuracy in responses, suitability in applying data mining techniques etc.

Collecting data through pure questionnaire technique where the researcher is not available with the respondent at the time of data collection, offers certain challenges. Ten questions questionnaire has been found to be quite effective in overcoming those challenges and improving the quality of the data.

It is difficult to motivate respondents to respond to the questionnaire by finding some time from their daily routine. This becomes more difficult if the perceived time required to fill the questionnaire is high. The length of the questionnaire, number of questions in the questionnaire etc. influence the respondent’s perception about the time required to fill them. If it is high, it may be difficult for the respondents to agree and get ready for answering the questions. Even if he does this under some pressure or interest, the data thus captured may not be of good quality. Ten questions questionnaire creates an initial impression that there are only ten questions. This provides the required entry for the researcher to convince the respondents to respond to the questionnaire.

Another reason of reduced quality of data collected through questionnaire is that the questionnaire may be filled in multiple sittings. This happens more if the questionnaire has relatively large number of questions. The respondent may take one or multiple breaks during his response. This can result in inconsistency in his response if questions are related to opinion, emotion, sentiments etc. Sometimes, respondent may fill the questionnaire in casual way resulting in distorted data if the length of the questionnaire is big. A reliable questionnaire is one which gets consistent response from the same respondent if used at different times within
a reasonably uniform time duration. It is easier to design a TQQ with high reliability than a questionnaire with larger number of questions.

Sometimes researchers misinform the respondents giving the estimated time required to fill the questionnaire as about 10 minutes or so in desperate attempt to get them ready to respond. But, actually it starts taking much more time than what is communicated. The respondents lose the interest and finish filling the questionnaire in quick way just to complete the formality. This too results in capturing of unrepresentative data and thus, wrong research results. Ten questions questionnaires (TQQ) are quite pleasant in responding. In worst case, even if a respondent loose interest in the middle, he may find himself close to the end of questionnaire. Hence, he may prefer completing his response with due seriousness.

It also provides significant advantage while applying some of the computation intensive techniques during analysis such as data mining techniques. Mining association rules provide all the strong enough relations among various combinations of variables in the data. If a combination is represented by a subset of the variables, then the number of subsets of a set with n elements is equal to $2^n$. Though, present days’ computers are capable of handling large volume of data and do high speed computations at affordable cost, it is always a better strategy to keep the number of variables within limits. With this view, keeping only ten number of questions in a questionnaire makes it possible to have comprehensive view of intervariable relations. Applying data mining techniques on data collected for research purpose requires use of suitable computer based tools. The algorithms used in the software to build the tool is usually not available to the users. Ten questions questionnaire keeps the number of possible intervariable relations within such a limit that allows the researcher to improvise with the algorithm and reach to meaningful results.
Though TQQ offers so many advantages in terms of quality of data, ease in data collection and application of advanced data analytic techniques, researchers sometimes find it difficult to accommodate all their data requirements in just 10 questions. This, in fact, is not required. Putting complex questions or multiple sub-questions to forcefully keep the number of questions in a questionnaire to ten defeats all the advantages of a ten-questions questionnaire. It may make the questionnaire too complex for the respondents to accept it with due sincerity. Thus, getting back to all the reasons of poor quality data collection through questionnaires. Also, the number of combinations of possible answers to the questions goes exponentially up because the base of the exponent itself goes up due to multiple parameters in a question.

Hence, it is desired to keep the questionnaires simple. If required, multiple questionnaires can be designed and used to collect the required data. Each of the questionnaires can be based on the TQQ concept to get all the advantages it offers.

In this research, responses were needed from the insurers and also the insured. Also, the insured may be individual or may be corporate. Hence, multiple TQQ were used to collect the data. Those individuals or corporates who were already insurance customers or were potential insurance seekers were part of the population of insured. Two different ten questions questionnaires were used to collect required information from them – one questionnaire was for retail insurance seekers and the other was for corporate insurance seekers. The questions were such that it was not required to refer any data or report for answering them. These were mentioned on the questionnaires so that respondent could feel comfortable about the time and effort required to provide their response to the questionnaires.

For the corporate insured or corporate insurance seekers, the questionnaire included questions related to the types of insurance coverages, the reasons for opting for insurance, factors considered in choosing the options, insurer role in identifying the service, risk management
dependencies, insurance process, and some questions on quality etc. The idea was to collect relevant information from corporate insurance seekers about their way of choosing insurance options and their understanding about the insurers way of finalizing the important insurance parameters.

Similar were the requirements of information from retail insurance customers or seekers as well. In case of corporate insurance seekers, the respondents to the questionnaire were representing their organizations. But, in case of retail insurance customers or seekers, the respondents were representing themselves. Hence, a different ten questions questionnaire was developed and used for collecting information from them. Here too the questions were related to insurance coverage, mechanism of insurance need determination, awareness about claim denial reasons and preferences related to insurance services etc. Questions were designed in such a way that they can be answered easily and precisely. These questions were not directly on hypothesis but were aiming to capture information about various aspects of insurance needs and insurance service selection.

Another ten questions questionnaire was used to collect information from the insurers or the insurance professionals. Respondents of these questionnaires were representing the insurance organization. The questions were related to information used in underwriting risk, use of information technology, impact of information technology on insurance practices, claim denial process and the steps taken to reduce the claim denial etc.

Any questionnaire should be reliable, valid and verifiable. A reliable questionnaire is one that gets similar answers to the questions from the same respondent if administered on different times within a time period not including any major environmental change. Since, these questionnaires were quite precise and required less data or effort and more of clarity in understanding, they were found reliable. Valid questionnaires are such questionnaires that
collect such data for which they have been designed. Such questionnaires should not have ambiguous questions and should be understood in similar ways by different people. The questionnaires designed here took care of this. A good questionnaire also includes some questions that can be considered as verification questions. Their answers can be verified from the answers of some questions in the same questionnaire. This helps in identifying if the responses were captured with due sincerity and the data could be considered of good quality. TQQ, normally don’t have such problems as they take less time and effort of the respondents. However, these questionnaires included questions for verification purpose as well.

3.4 Sampling and sample selection

It was a requirement to collect information from insurance customers or seekers of retail and corporate insurance and also from the insurance professionals. With hundreds of insurance products and close to fifty insurance companies in India, the insurance service is quite vast. Hence, it was necessary to use sampling technique to collect data.

Sampling technique is used when large number of individuals or objects (Population) are to be observed or investigated in the research. A sample of the individuals or objects are selected and the required experiment or data collection is done through them to draw conclusion about the population. Sample design involves the method to be used for sample selection and the kind of estimates to be done. In this research, TQQ was to be administered on the individuals representing themselves as insurance customer or representing their organization as insurance corporate customer or representing themselves and their organization as insurance professionals. Ten questions questionnaire was preferred because of its ease in using data mining techniques and ability to capture high quality data from the respondents.
Sample should represent the population and sample selection should be unbiased. There can be a sample selected randomly from the population as per the required sample size. Or, if there is some clear characteristics visible on which the population can be divided into various strata then stratified sampling can be done. In stratified sampling, sample is selected from each strata or randomly sampled strata. Sample can be clustered as well in which within the sample various samples are selected based on clusters.

Selection of the right sampling technique is very important to have a sample that represents the population and also has no bias that can adversely affect the quality of research result. Insurance services have diverse interrelated aspects. It was quite complicated to structure that in the beginning of the research. Hence about 200 persons were interviewed to understand the dynamics of insurance needs, their ways of risk management, their perception about insurance options etc.

Individual insurance seekers may be related to some of the common insurance services like health insurance, life insurance, motor insurance, travel insurance etc. or may be for some of the not so common services like property insurance, liability insurance etc. Corporate insurance seekers can be for project insurance, some specialized insurance, group insurance etc. They can be from various industries. Stratified random sampling technique was used to select the persons from available groups for interview representing various segments such as Individual Insurance service seekers, Software Industry, Manufacturing Organization, Chemical Industry, Transport & Shipping Services, Agriculture and SSI etc.

As there was huge variation in responses between individual insurance seekers and corporate insurance seekers, two different questionnaires were designed to collect the data. One was for individuals and the other one for corporate. Sample selection for collecting responses through the questionnaires from individual insurance seekers were done through simple random
sampling. In case of corporate insurance service seekers, there were clear characteristics that could differentiate them in different types of industries. Hence, a sample was selected for corporate and within each of the sample items random sampling was done to prepare the sample.

It is quite difficult to eliminate any bias if the sample items are selected by human beings manually. There is some kind of preferences or prejudices in the mind of any individual that can influence the selection of items from the lot for sample purpose. This creates biasedness. If individuals are being sampled for collecting responses through questionnaires, there may be tendency of selecting sample individuals based on look, height, color, body language etc. or ignoring some based on some arbitrary criterion. These may happen unintentionally and the researcher may feel that he has not created any bias. To avoid such situations, random numbers were used to create the randomness.

Random numbers are such numbers that the probability of any one number getting selected remains same irrespective of the number. Every programming language or computer software used in analytics etc. have random number functions. There is random number table as well that contains numbers using which one can generate required number of random numbers. These are based on some algorithms to keep the probability of any number being selected same as other numbers. In random sampling, these random numbers are of great help in selecting sample without any bias. For this, random numbers are generated and then these numbers are associated with the real system variables. Some planning may be required to associate the random numbers with real system variables.

For this purpose, the lists of all the individuals from which the sample was to be selected were prepared in excel sheet. Then, random numbers between 0 and 1 were generated using the rand function of excel. These random numbers were generated in every row of the excel sheet
containing detail of any prospective respondent. These data were then sorted in ascending order of random numbers. The order of sorting is not important, it can be anything – ascending or descending. This is equivalent to shuffling the list vigorously. Then, the required number of individuals were selected from the top of the list. The rule for selection should be made in advance before generating the random numbers. Sorting order, the position in the list from where sample is to be selected, number of rows to be taken as sample etc. need to be pre-decided to avoid any bias. In this case, the list was sorted in ascending order and the records were selected from the top.

The Sample size depends on the population size, variation in the data, the level of accuracy required and the confidence level. One can get idea about whether rice has been cooked properly or not by testing two-three grains only. But this sampling concept gives correct result when heat is uniformly distributed in the entire volume of rice being cooked. If there are multiple types of rice being cooked together and the heat distribution is not uniform, then one has to select samples from each type of rice and each segment of heat distribution. If the variation in data is more, sample size will be more. Similarly, accuracy required is a measure of how accurately the results obtained from sample investigation represents the population. Confidence level talks about the probability which the result is likely to be that accurate. So, higher confidence level and accuracy required, higher the sample size would be. But, variation in the data is quite important in deciding the sample size. For a completely uniform population even a small sample can give accurate result with high confidence level.

In this research, the respondents of the questionnaires were all insurance seekers or insurers. Also, different TQQs were used to collect information based on the type of respondent. Though the population size is quite big but there is uniformity in interests. The sample size was found by using the Slovin’s formula
\[ n = \frac{N}{1 + Ne^2} \]

Where \( n \) is the sample size, \( N \) is the population and \( e \) is the confidence level. The value of confidence level \( e \) has been taken as .05. This formula is suitable for such situations where the population size is big, respondents are of similar interests and questionnaires are based on types of respondents.

Inverse of sample size \( n \), i.e., \( 1/n \) in this formula is equal to \( (1/N) + e^2 \). For high population size \( N \), \( 1/N \) becomes very small and can be neglected. The other term \( e^2 \) is .0025 for \( e = .05 \). Thus, the inverse of the sample size can be approximated to .0025 for high value of \( N \) (Population). This gives sample size as \( 1/ .0025 = 400 \). So, the sample size for insurers as well as insurance seekers were taken as 400.

### 3.5 Association mining using Apriori algorithm

Association mining is similar to identifying strong relationships with a difference that it works as a discovery technique rather than a testing technique. It works on the data. If something say A happens sufficient number of times and when A happens then something else say B also happens with high enough frequency, then it suggests that A and B are strongly associated. Sufficient number of times for A and high enough frequency for B are the two parameters that the researcher has to define. If A is a rare occurrence, then it may not be of interest of the stakeholders. However, how much rare is considered to be rare depends on the volume of data and the presence of A in that. Similarly, the strength of the relationship too depends on the area of research and the objective behind conducting the research. The algorithms used for mining association rules discovers all the strong enough associations that are supported by the data.
Suppose the threshold support value is $s$. That means, something should happen minimum $s$ number of times in the data to qualify for further investigation. Suppose A happens at least $s$ number of times then further investigation is done on that. While applying association mining, a researcher is interested in knowing the possibility of something else say B happening if A happens.

So, a superset of A gets formed that contains elements of A and also some extra elements to make it B. When it is said that whenever A happens, there is high possibility of B happening (or B also happens) then it is understood that B contains elements of A as well. There will always be some possibility of B happening whenever A happens. This possibility, if defined as probability can be zero or nonzero. Such a nonzero probability that can be considered as high enough is taken as threshold probability. Let this be represented by $p$. Output of mining association rules are in the form of statements like “If A happens then B also happens”. In this the threshold support criteria and threshold probability criteria must be met by A and B.

There are several algorithms to mine association rules. Since, the number of subsets and possible associations are quite high, the association algorithms focus on achieving computational efficiency to provide the results in acceptable time. Apriori algorithm for mining association rules is one of the most popular association mining algorithm. It uses some basic principles of set theory and the probability concepts. Some of the theories that are used in this algorithm are:

a) There are $2^n$ subsets of a set containing $n$ elements.

b) If a subset has frequency $f$ then its supersets will have frequency less than or equal to $f$. In no case the superset can have frequency more than that of its subsets.

c) If a subset doesn’t meet the threshold support criteria then its supersets too will not meet this criterion.
d) If the threshold support value is $s$ and the threshold probability is $p$, then if frequency of $A$ is more than $s$ then frequency of its superset say $B$ should be at least $p$ times the frequency of $A$ to qualify for a strong enough association. If the frequency of $B$ is not meeting this criterion, then its supersets too will not meet this criterion.

Apriori algorithm works by checking the frequency $f$ against threshold support value $s$ for subsets and eliminating supersets of that for further consideration when this criterion is not met. It first checks the subsets of single element and counts the frequencies $f$ of their occurrences. While counting the frequency, it includes all such incidences in which that element happened. If something more also happened along with that element, then that too is counted in that. The frequency $f$ is then compared with the threshold support value $s$. If the frequency is less than the threshold support value, all the supersets of the single item subset are removed from the list of subsets to be examined. But, if the frequency is more than the threshold support value, then the frequency is multiplied by the threshold probability $p$ to get the value of $fp$.

Now, immediate supersets of the subset which met the threshold support criteria are checked. Immediate superset includes one more element in the subset. The frequency is counted and compared against $fp$. If the frequency is less than $fp$, then it doesn’t give a strong enough association. But, if the frequency is higher than $fp$ then the association is strong enough and valid. It indicates that if the elements of the subset happen then the additional elements of the superset too will happen. The algorithm continues building the supersets, counts the frequency, compares that with $fp$ and accordingly include or not include as strong enough association. Since, the frequency of various subsets is counted during this process and they may be required for checking the strength of association while checking the threshold probability criterion for other subsets, the subsets and their frequencies are stored in proper format to avoid repeat of counting.
After exploring all the subsets of single element, the algorithm, then checks for the 2 elements subsets that have not been eliminated on threshold support criterion as superset of their subsets. For these two elements subsets, frequencies are counted and compared with the threshold support value s. If the frequency is less than s then the subset and all the supersets are eliminated from the list. If the frequency is more than s then further computation is done for finding strong enough associations by checking the threshold probability criterion of supersets. This process continues till all such possible associations meeting the threshold support and threshold probability criterion are identified.

Since, this algorithm starts checking for threshold support criterion with smallest subsets, a large number of supersets gets eliminated for checking on this criterion if a subset fails to meet this. For example, if there are 20 elements in the complete set, it has $2^{20}$ subsets out of which one is null subset and 20 are with single element. If one of the subset of single element fails on the threshold support criterion, all its supersets too will fail because their frequency cannot be more than the frequency of this subset. Number of supersets of this single element subset will be $2^{19} = 524288$, including the subset itself. Thus, 524288 subsets get eliminated from the list of subsets to be checked for threshold support criterion. This approach reduces the computational workload drastically during mining association rules.

Because of the computational advantages the Apriori algorithm provides in association mining, various computational tools have built association mining features using this algorithm. Most of the popular open source software or programming platforms that are especially suitable for data mining applications have in-build features based on Apriori algorithm to mine association rules. Data can be created in these tools directly or may be imported through other common data analytic tools like excel or calc. R is one such open source programming language that has such features that make it highly powerful tool for statistical analysis and data mining.
Similarly, Weka is another such tool that is specialized for data mining application. Data stored in csv format can be opened in Weka to apply Apriori algorithm for association mining.

Data collected in this research through the questionnaires were stored in excel sheet for ease in data entry. Preliminary frequency checking was done through excel operations. The data were converted to csv format to conduct experiments using Weka. It allowed experimenting by changing threshold support and threshold probability values. Though, it discovered quite a few interesting and powerful associations, but the main interest in this research was on the objectives of the research and the formulated hypothesis.