International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 4 Issue 11, November 2017

ISSN: 2394-4404

From Methodology To The Applied Statistics - A Research Case In Telecommunication Industry

Eralda Caushi

Klejda Caushi

Universita' Cattolica "Nostra Signora Del Buon Consiglio", Albania

***Abstract:** One of the most important steps in Telecommunications industry is to understand the behavior of the customers, encourage them in spending more and then predicting their future by preventing their attrition. The churn might be voluntary in cases they want to leave the network they actually are using, or involuntary churn in case of unpaid bills. The methodology used to do the right evaluations in order to achieve strong results in this field is very large and varied. The paper intends to give insights about the methodology used to evaluate the best models in statistical fields. The best approach to start is by studying the existing literature and then developing new structures. That's why the study is focused in standard disciplines combined with the applied statistics.*

***Keywords:** Bayes, Sample, Inference, Telecommunication*

***Jel Codes:** A20, C00, L96*

I. INTRODUCTION

The paper considers the main pillars of the PhD degree program in order to deal with the research of the identification and analysis of the methodologies used to estimate the customer attrition in the industry of the telecommunication. Descriptive Statistics is the very first discipline that will help us evaluate the data we have in our possession. Graphical analysis is the main representation of the analytics. Inferential statistics will help examining the entire population instead of a sample. The sampling techniques allow us optimizing the sample extraction criteria so that we can obtain the same information from the sample, which would have been obtained by having the entire collective. So, that's the reason that sampling techniques is one of the most important discipline that we will refer in this paper. Methods of collecting, summarizing, analyzing, and interpreting variable numerical data are summarized in the chapter Statistical Methods. They can be contrasted with deterministic methods, which are appropriate where observations are exactly reproducible or are assumed to be so. Indicators are the tools that can measure the behavior of a phenomenon that is considered representative for the analysis and are used to monitor or evaluate the degree of success or adequacy of the activities implemented. A strong point to be focused in is the information system, without which every single calculation would be too complex to calculate. Considerable should be the knowledges in the field of Mathematical Analysis and Marketing. The second one helps us formulate the needs of any business.

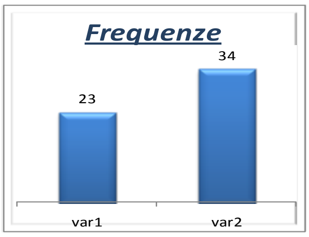
II. LITERATURE REVIEW

In this paper I will briefly mention the main knowledge that is needed in order to do an accurate analysis of the customer attrition in the telecommunication. The programs are developed within the attention of the professors of the university Our Lady of Good Council that meantime are teaching at the University of Bari "Aldo Moro". The main objective of the Research Site is to acquire a thorough preparation in the field of statistical methodology for applications in particular contexts and research issues. In particular, the areas of interest are aimed at corporate, social, economic, informatics, biological, environmental etc. Our area of interest is telecommunication and the entire program is adapted by us in order to be applied to this field. The main source we have been based in is the book Research Methods for Business- A skill building approach by Uma Sekaran and Roger Bougie. The scope is to be useful when we use the methodology to the applied statistics.

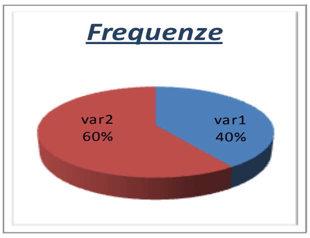
III. COURSES AND CONTENTS

DESCRIPTIVE STATISTICS

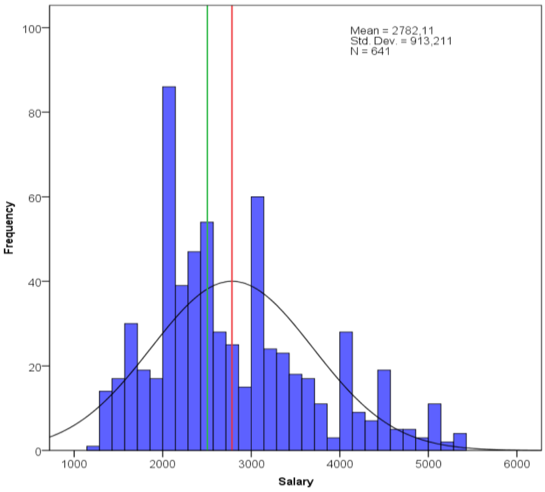
Descriptive statistic is the discipline in which the methodologies used by a tester are studied to collect, represent and process the observed data for the purpose of analyzing a certain phenomenon. Data is collected through a total survey, census detected on all population units or by sampling. Data may be quantitative or qualitative. The data is sorted in statistical tables. Graphic representations: Bar graph, pie chart, histogram. The analytical representation of the theoretical distributions is to find an interpolating mathematical function that adequately represents an observed statistical phenomenon. Its purpose is to get a general view of the data and the distributions of the variables by diagrams, tables, and basic statistics, such as mean and standard deviation. The descriptive analysis is a necessary part of the research and is always conducted before doing any statistical tests or more complicated modeling. This part presents some common techniques for descriptive data analysis, while the next section inferential statistics focuses on statistical testing and modeling.



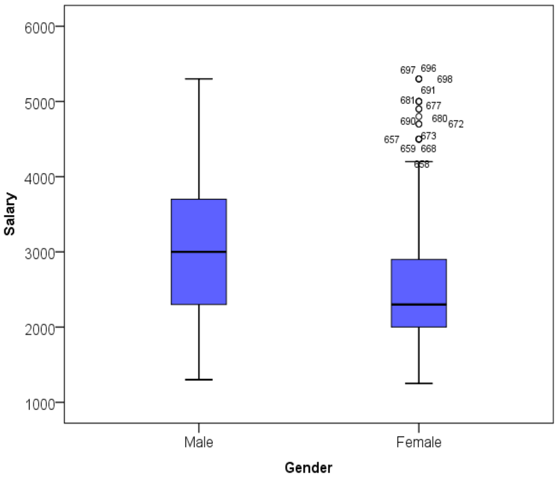
Graph 1: Chart of frequencies



Graph 2: Pie-Chart of two variables



Graph 3: Histogram



Graph 4: Box Plot

INFERENTIAL STATISTICS

Regression: examines the linear relation between one or more explanatory variables (or independent ones) and a criterion (or dependent) variable: Y = α + βX. By correlation is meant a relationship between two statistical variables such that each value of the first variable matches with “certain regularity" the value of the second one. The degree of correlation between two variables is expressed by the so-called correlation indices which assume values between - 1 and + 1. We refer to descriptive statistics when addressing a direct population survey, instead of statistical inference when starting from examining a sample to have information about the entire population. Addressing a statistical inference problem should refer to a Model. The three main techniques used in the inference are theory of estimation, use of confidence intervals and test theory. In probability, it is considered a phenomenon that can be observed solely from the point of view of the possibility or not of its occurrence, regardless of its nature. Between two extremes, known as event and event impossible, there are more or less likely events.

Figure 1: Probabilities of a coin

A random variable X is a numeric variable whose measured value may change by repeating the same measurement experiment. X can be a continuous or discrete variable. The maximum information that can be given to questions of the type: what is the probability that in a future measurement the value of X is between a and b, where a <b are two numbers a real assignment? These probabilities identify the distribution of the random variable. In the field of statistical inflection, two schools of thought are distinguished, linked to different concepts, or interpretations, of the meaning of probability: Classical / frequentist inferences and Bayesian inferences. A probability distribution is a mathematical model

that links the values of a variable to the probability that these values can be observed. Formally, probability distributions are expressed by a mathematical law called probability density function (indicated by f (x)) or probability function (indicated by p (x)) respectively for continuous or discrete destruction. Discrete distribution examples are Binomial, Poisson etc. Instead of continuous distribution are the Normal, Exponential, and Weibull.

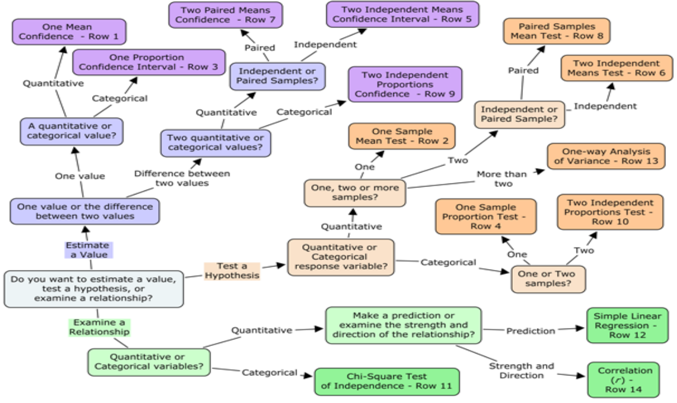

Two Paired Means Confidence - Row 7
Two Independent Means Confidence Interval - Row 5
One Mean Confidence - Row 1
Paired
Independent
Paired Samples Mean Test - Row 8
One Proportion Confidence Interval - Row 3
Independent or Paired Samples?
Two Independent Proportions Confidence - Row 9
Paired
Two Independent Means Test - Row 6
Quantitative
Categorical
Quantitative
Categorical
Independent or Paired Sample?
Independent
A quantitative or categorical value?
Two quantitative or categorical values?
One Sample Mean Test - Row 2
One-way Analysis of Variance - Row 13
One
Two
More than two
One value
Difference between two values
One, two or more samples?
One value or the difference between two values
One Sample Proportion Test - Row 4
Two Independent Proportions Test - Row 10
Quantitative
Estimate a Value
Quantitative or Categorical response variable?
One
Two
Test a Hypothesis
Categorical
Do you want to estimate a value, test a hypothesis, or examine a relationship?
One or Two samples?
Examine a Relationship
Quantitative
Make a prediction or examine the strength and direction of the relationship?
Prediction
Simple Linear Regression - Row 12
Quantitative or Categorical variables?
Categorical
Chi-Square Test of Independence - Row 11
Strength and Direction
Correlation (r) - Row 14

Figure 2: Variables and statistical tests

SAMPLING TECHNIQUES

Statistical data may come from the following sources: censuses, sample surveys, and data processing collected within administrative procedures, that is, administrative source data. Statistical surveys - performed on a sample or total population - are developed according to a process that, starting with the definition of detection targets, collects and processes the data and concludes with the analysis and dissemination of the results. The questionnaire is the tool with which you can find the data of interest. Questionnaires can be formulated in such a way as to provide different ways to answer open questions or structured questions.

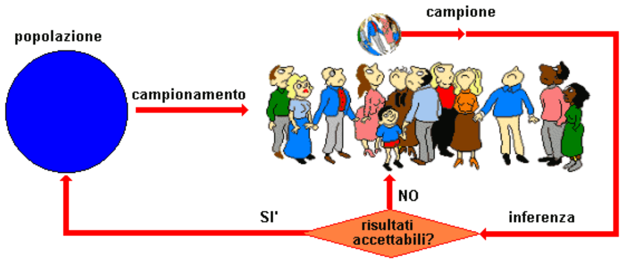


Figure 3: Sampling and Population

The sampling techniques allow optimizing the sample extraction criteria so that we can obtain the same information from the sample, which would have been obtained by having the entire collective. In this way you can get the same information, with costs, however, significantly lower and, often very important, with extremely quick times (eg electoral projections). By sample is meant that group of elementary units, a particular subset of the population, identified in it so as to allow, with a definite risk of error, the generalization of analysis results to the whole population. A sampling plan defines a method by which you select items that are part of the sample. A first major distinction to note is that of probabilistic samples and non-probabilistic samples. The probabilistic sample selection methods can be different. They are distinguished by: simple random sample, stratified sample, cluster sample, systematic sample, multi-stage sample, etc. Depending on how the sample units are selected. The parameters of interest in the population are usually the average and the total. The estimators that are often used are those of Hansen-Horvitz and Horvitz-Thompson. Correctness and efficiency evaluate the bounty of the estimators used.

STATISTICAL METHODS

The analysis of historical series includes a series of statistical methods to investigate a historical series, determine the process underlying it, and make predictions. According to the traditional approach, it is assumed that the process has a deterministic part, which allows it to be dissociated In trendy, cyclical and / or seasonal components, and that the difference between the theoretical data of the deterministic model and the observed data is attributable to a residual random component. According to the modern approach, however, it is assumed that the process described has been generated from a stochastic process described by a parametric probabilistic model. The distances-the groups should be unit assemblies on the one hand as homogeneous as possible and on the other as separate as possible. This suggests introducing distance indices so as to clarify the notion of proximity and homogeneity. Cluster Analysis (CA) consists of a set of statistical techniques to identify groups of units similar to a set of characters taken into account, and according to a specific criterion. The objective that we set ourselves it is basically that of bringing together heterogeneous units into more subordinate and mutually exhaustive subsets. The statistical units are, in other words, subdivided into a number of groups depending on their level of "resemblance" from the values that one Series of preset variables assumes in each unit.

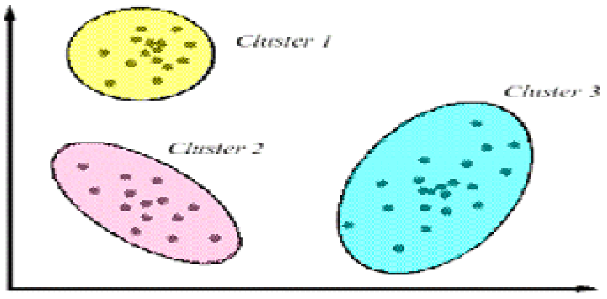


Figure 4: Clustering

The joint measurement analysis and a multivariate analysis technique that takes into account consumer preferences in the choice of goods and services. Through Conjoin Analysis you can check: the degree of relevance to each level or mode of each feature, and the importance each individual attributes to a feature of a product or service.

STATISTICAL METHODS FOR BUSINESS

Performance Plan - The Plan's definition process has followed some logical phases: defining the history, the current and the identity of the organization; Analysis of the external and internal context; Definition of strategic objectives and strategies; Definition of operational objectives and operational plans; the development of programming and control systems and the improvement actions to be promoted.

Figure 5: Performance Plan

BUILDING INDICATORS

The index is a ratio between two numbers and is intended to compare two entities. The indicator is the sum of one or more factors such as GDP, which will affect the final result from the sum of the individual factors. Indicators are tools that can show (measure) the behavior of a phenomenon that is considered representative for the analysis and are used to monitor or evaluate the degree of success or adequacy of the activities implemented.

Some important indexes in the telco are:

ARPU (Average Revenue per Unit) = Total Revenue / Nr of MSISDN

ARPA (Average Revenue per Account) = Total Revenue / Nr of Accounts

Churn Rate = No. of Customers Churning / (Closing Base-Opening Base)

Margin of Postpaid Customers = Total Postpaid Customers / Total Customers

INFORMATION PROCESSING SYSTEMS

The information system consists of all the information used, produced and transformed by a company during the execution of business processes, the manner in which they are managed and the human and technological resources involved. This is from data describing business or environmental phenomena. The IT infrastructure consists of a set of shared technology resources that integrate each other to provide the operating environment for enterprise applications and business processes. These resources are of hardware type such as server and storage, software such as operating systems or services such as configuration installations and customizations. Microsoft Access, also known as Microsoft Office Access, and a database management system from Microsoft that combines the relationships between the Microsoft Jet Database Engine relational model and a graphical user interface and software development tools.

BAYESIAN INFERENCE

The Bayesian approach might have a very important role in Telco industry due to computational reasons. Epistemological reasons and pragmatic reasons are also considerable reasons that guide us in our study. As explained by Brunero Liseo in “Introduzione alla statistica Bayesiana”, from an epistemological point of view the reasons for using this method are based on a simple and direct inductive reasoning method, according to the information available on a certain set of phenomena, in a certain moment of life that wants to calculate the probability of future events or, more generally of events for which it is not known whether they are verified or not. Bayesian logic is consistent with very logical basis and free of risk counterexamples, always waiting for innovation when it is used the method of induction, and it is necessary to produce statements of probabilistic nature of events that we do not know if it will happen or less. Pragmatic reasons are related to the need of taken in consideration the extra-experimental information of the problem that need to be solve. That refers to the Bayesian setting. In telecommunication, for example, when assessing the probability that a customer might leave the network due to the reduction of some particular offer those that are the a-priori probabilities (extra-experimental info) and are nothing else but the information on the specific offer we need to include in our problem solving. Also very useful in this sector is to have the information at a level of disaggregation sufficiently high. This need goes under the name of “small area estimation” that refers to the difficulty of producing information for areas of which we do not have access to the sample. So, estimating the possibility to churn for a single customer that belongs to a sample of a company for which we do not have data, might be possible using the Bayesian method. So, an intrinsic characteristic of the Bayesian method is precisely that of being able to assume, in a simple and natural, different levels of association between the units of the sample, allowing the phenomenon of “borrowing strength” which allows the production of estimates sufficiently stable for those areas with no sample data. The Bayesian Method gives the possibility to integrate, using Bayes theorem, all the information generated by the statistical experiment with the “a priori” data. Monte Carlo methods, based or not on the properties of Markov chains, gives the possibility to generate a sample of whatever dimensions, independent identically distributed by the distribution a posteriori of the parameters we are interested in. That’s why in a very large contest the Bayesian approach permits the flexibility of a model which is very difficult to be achieved through classical methods. The following figure shows clearly the concept of borrowing concept. In high dimensions, as we are considering telco data, the potential gain is large. A-priori knowledge should make this gain even bigger.

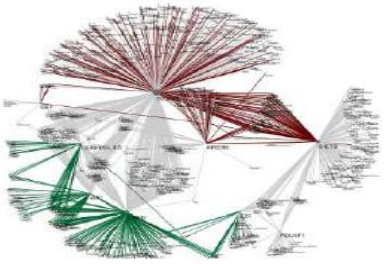
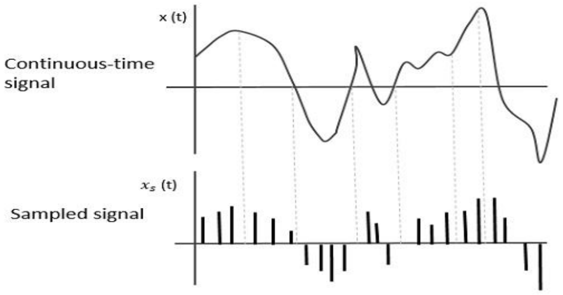


Figure 6: Borrowing strength

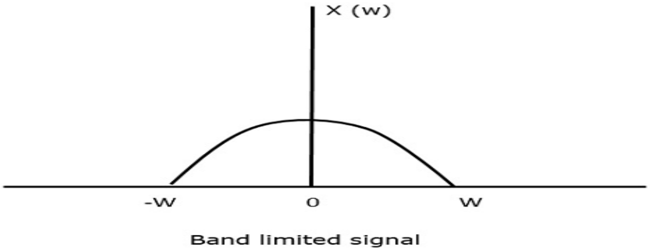
DIGITAL COMMUNICATION SAMPLING

Sampling is defined as, “The process of measuring the instantaneous values of continuous-time signal in a discrete form.” Sample is a piece of data taken from the whole data which is continuous in the time domain. When a source generates an analog signal and if that has to be digitized, having 1s and 0s i.e., High or Low, the signal has to be discretized in time. This discretization of analog signal is called as Sampling. The following figure indicates a continuous-time signal x (t) and a sampled signal xs (t). When x (t) is multiplied by a periodic impulse train, the sampled signal xs (t) is obtained.



Graph 5: Continuous-Time and sampled Signal

To discretize the signals, the gap between the samples should be fixed. That gap can be termed as a sampling period Ts. Sampling Frequency=1Ts=fs Sampling Frequency=1Ts=fs Where, Ts is the sampling time fs is the sampling frequency or the sampling rate. Sampling frequency is the reciprocal of the sampling period. This sampling frequency can be simply called as sampling rate. The sampling rate denotes the number of samples taken per second, or for a finite set of values. For an analog signal to be reconstructed from the digitized signal, the sampling rate should be highly considered. The rate of sampling should be such that the data in the message signal should neither be lost nor it should get over-lapped. Hence, a rate was fixed for this, called as Nyquist rate. Suppose that a signal is band-limited with no frequency components higher than W Hertz. That means, W is the highest frequency. For such a signal, for effective reproduction of the original signal, the sampling rate should be twice the highest frequency. The sampling theorem, which is also called as Nyquist theorem, delivers the theory of sufficient sample rate in terms of bandwidth for the class of functions that are bandlimited. The sampling theorem states that, “a signal can be exactly reproduced if it is sampled at the rate fs which is greater than twice the maximum frequency W.” To understand this sampling theorem, let us consider a band-limited signal, i.e., a signal whose value is non-zero between some –W and W Hertz. Such a signal is represented as x(f)=0for|f|>Wx(f)=0for|f|>W . For the continuous-time signal x (t), the band-limited signal in frequency domain, can be represented as shown in the following figure.



Graph 6

IV. CONCLUSIONS

So far we have done an evaluation of all the knowledge we have gained in order to advance in our research. That’s why the study is focused in standard disciplines combined with the applied statistics. As we have reaffirmed above in Telecommunications industry is to understand the behavior of the customers, encourage them in spending more and then predicting their future by preventing their attrition. The methodology used to do the right evaluations in order to achieve strong results in this field is very large and varied. Descriptive Statistics is the very first discipline that will help us evaluate the data we have in our possession. Graphical analysis is the main representation of the analytics. Inferential statistics will help examining the entire population instead of a sample. The sampling techniques allow us optimizing the sample extraction criteria so that we can obtain the same information from the sample, which would have been obtained by having the entire collective. So, that’s the reason that sampling techniques is one of the most important discipline that we will refer in this paper. Methods of collecting, summarizing, analyzing, and interpreting variable numerical data are summarized in the chapter Statistical Methods. They can be contrasted with deterministic methods, which are appropriate where observations are exactly reproducible or are assumed to be so. KPIs or Indicators are the tools that can measure the behavior of a phenomenon that is considered representative for the analysis and are used to monitor or evaluate the degree of success or adequacy of the activities implemented. A strong point to be focused in is the information system, without which every single calculation would be too complex to calculate. Considerable should be the knowledges in the field of Mathematical Analysis and Marketing. The second one helps us formulate the needs of any business. The Bayesian approach might have a very important role in Telco industry due to computational reasons. Its logic is consistent with very logical basis and free of risk counterexamples, always waiting for innovation when it is used the method of induction, and it is necessary to produce statements of probabilistic nature of events that we do not know if it will happen or less. Sampling is defined as “The process of measuring the instantaneous values of continuous-time signal in a discrete form.” Sample is a piece of data taken from the whole data which is continuous in the time domain.

REFERENCES

[1] Liseo B. (2008) Introduzione alla statistica Bayesiana” – Dispensa
[2] Sekaran U. Bougie R., (2016) Research Methods for Business, A skill-building approach. John Wiley & Sons LTD
[3] Ross, S.M. (2013) Calcolo delle Probabilità. Apogeo
[4] Resnick, S.I. (1999) A Probability Path. Birkhauser
[5] Pace L., Salvan A., (2001) Introduzione alla statistica II Inferenza, verosimiglianza, modelli. Padova: Cedam
[6] Azzalini A., (2001) Inferenza statistica: una presentazione basata sul concetto di verosimiglianza Milano: Springer-Verlag Italia

[7] Verbraken T. Verbeke W. Baesens B. (2013) Profit Optimizing Customer Churn Prediction with Bayesian Network Classifiers, Amsterdam