**SAMPLING PLAN**

#### 

# **1.Sampling Strategy**

## **Objectives and Reliability Requirements**

Patients receive appointments to the doctor according to the disease they are suffering from This means that patients suffering from different diseases i.e(hypertension,diabetes,Alcoholism,handicap) are all grouped together according to the listed diseases and each disease has its day to see the doctor. The patient receives an sms before the date of appointment to remind them and also inform them the time set for the appointment.

However, there are cases where a patient receives an sms and is not able to see the doctor because of unavoidable circumstances or the patient does not receive an sms and he/she sees the doctor.

The main objective of the project is to help the medical company to follow up on patients who received sms and never showed up for an appointment.

The data i used had 110527 rows \* 14 columns which is huge and time consuming to analyze. Therefore, as a data analyst I came up with a sampling plan for ease of analyses.This plan is a good initiative because it will ensure that the experimental sample(estimated value) is truly representative of the entire population.

## **Target Population**

The data used was extracted from an open source website.This medical data was collected for the purpose of understanding their clients more.

The total data collected was that of 110527 patients which included:

1. The patients who received an sms and showed up for appointments.
2. Those who never received an sms and showed up for appointment.
3. Those who never received an sms and never showed up for appointment
4. Those who received an sms and never showed up for appointment..

The target population in this case were the number of people who never showed up for the doctors appointment and received an sms.

The variables associated with the target population in this dataset were PatientId, AppointmentID, ScheduledDay,Gender , Appointment Day, Age, Neighbourhood, Scholarship, Hypertension, Diabetes, Alcoholism, Handicap, SMS\_received, No\_show

## **Sampling Method**

For the accuracy method the sample to be used was to be randomly selected using the simple random sample.

This is because the dataset was large and it would be challenging to group them into clusters or even stratify them. While these sampling methods would have also done the same job, I decided to use simple random sampling since it cuts across the dataset, It is more accurate whereas it is not prone to sampling errors and was easier to use for such a large dataset of 110527 entries.

## **Sample Size**

From the large sample of 110,527\*15 entries i first located the patients who received an sms and never showed up for appointment which i found out to be 9784 patients there is where i got a sample size of 978 which was 10% of the total population of patients.

These patients were a combination of randomly selected males and females with random selection of their PatientId, AppointmentID, ScheduledDay, Appointment Day, Age, Neighbourhood, Scholarship, Hypertension, Diabetes, Alcoholism, Handicap, SMS\_received, No\_show.

There were no specifications when running the simple random sampling algorithm, which made it easier and more precise.

From the sample, the sample mean should be an equivalent of the population mean, an ideal that cuts through the different measures of dispersion.

## **Sampling Frame**

In this case, the sampling frame is the Mexican states hospitals. It is from this frame that we get the population from which we also get the samples.

It is important to note that there has to be a sample frame, no matter where the sample is obtained. The Mexican cities gives us the neighbourhoods data which would later be used in the analysis.

# **2. Data**

## **Field Measurements**

For the success of my project a proper study and consideration of the dataset has to be made as this will help the researcher in understanding and analysis of the data provided and sample.

For this stage, the variables under consideration are: Gender, Age, SMS\_received, No\_show, AppointmentID, ScheduledDay, PatientId.

The frequency of the data is 9784 after we located those who received an sms but never showed up for appointment however we will be using the population sample of n=978 patients for analysis and possible arrival to the necessary conclusions

## **Quality Assurance / Quality Control**

In Data collection, Data integrity is one of the most crucial steps in data analysis, responses and conclusions bound to be derived from the data. If data quality is low,the results will be highly misleading starting from the sampling error, a result of data without integrity.

Since the data was obtained from an open source, we assume the researchers did not go through the data quality process so as to avoid any biases.

Good quality data must be accurate, complete,relevant,available,detailed and timely.

Some of the ways in which data can be assured are:

### **Performing data cleaning**

#### **i) Validity**

I had to make sure there were no irrelevant variables that would not be needed / helpful in my analysis. Some of this relevant variables are: Neighbourhood, Scholarship, Hypertension, Diabetes, Alcoholism, Handicap

#### **ii.) Accuracy**

I had to make sure that my data was between my timeframe which was 2016.

#### **iii)Completeness**

I also have to make sure that my data was complete hence there were no missing entries.

#### **iv)consistency**

I have to make sure that there are no double entries in my data.

### **Availability**

The data to be used during analysis must be freely accessible and available.

### **Relevance**

The data I have must consist of relevant information like the Patients Id, the patient's Appointment ID, Scheduled Day for appointment, Appointment Day, Age, patients who received an SMS and Patients who did not show up for the appointment, which is important for the success of my research .

### **Timeliness**

My data must be timely as it should be in the range of my time frame

### **Using data obtained and not cherry pick the data**

There have to be no modifications made to the data since any modification would lead to low quality so i should opt to use the data i would have collected.

### **Data profiling**

To ensure that data did not fail to represent the intended required goal. This can be done by reviewing the data collected, division of the data into sets while ensuring that the data is thorough and finally ensuring that the summation of the data is a representation of the whole data collected.

### **Building a quality assurance team**

To ensure the whole process of handling data is properly reviewed and data collected is accurate.

### **Data matching, Data parsing, Data enrichment and Data monitoring**

Identifying, matching and merging records that correspond to the same entities from several databases or even within one database.

To ensure data is transformed into a more readable format.

To ensure data is tracked using relevant dashboards, alerts and reports

## **Analysis**

The sample obtained from the data will be used to determine the distribution of the total population, the measures of central tendencies and dispersion for easy analysis of the larger dataset.

The sample will also be used to answer the research question along with the others, For example:

* How many females and how many males received an sms and did not go for appointments.
* Which ages are more prone to miss out on the appointments despite being sent an sms.
* Which day and month had the most number of patients missing out on the appointments yet they received an sms despite scheduling for the appointment

# 

# **3. Implementation**

In the implementation stage of this sampling plan, a qualified research team was sent out on the field to collect this data from several databases or individual responses.That is how it needed to be done for the data to reach the analysis stage.

However, considering I used open source data, the implementation of this sampling plan remains an assumption that I made in reaction to the collected data.

The implementation stage of the whole process when the data is collected is a section I didn't tackle as the report did not require that.

.