# ELIZABETH JOSEPHINE SAMPLING PLAN

**Python notebook:**

- Link

## THE SAMPLING PLAN

### 1. Sampling Strategy

#### a. Objectives and Reliability Requirements

Like any huge data analysis process, it was necessary to come up with a sampling plan for easy and thorough analysis of the provisional dataset. This is because it would be time consuming compared to navigating through the whole dataset of (150, 000 * 6) entries.

This plan would ensure that the estimated value obtained from this sample would be a good representation of the city population's true value. Supposing the researcher opted for the longer way out.

The main objective for this sampling plan is to help a US insurance company in the determination of the cities within which its clients have the highest annual income for the people who are aged 40 and below.

#### b. Target Population

The dataset used was extracted from an open source website. This medical data was collected from all the cities across the  united states for the purposes of understanding their clientele base more. The total data collected was that of 150, 000 US citizens.

The target population in this case were the United states citizens aged between 40 and below who had insurance cover with this insurance company.

The variables that were associated with the target population in this dataset were the residential city, the gender of the respondents, the age of the respondents, their income ranges and whether or not they were having any illnesses at the time within which the data was collected.

### c. Sampling Method

For accuracy purposes, the sample to be used was randomly selected using the simple random sampling technique.

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 opted to use simple random sampling because it cut across the dataset and was easier to use for such a large dataset of 150, 000 row entries.

### d. Sample Size

From the large sample of 150, 000 * 6 entries, I got a sample size of 30 individuals who were randomly selected to ensure there were no biases in the selection methods.

These individuals were a combination of randomly selected males and females with random income ranges, ages, residential cities and even their illnesses. There were no specifications while running the simple random sampling algorithm, which made it even more precise.

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

### e. Sampling Frame

By definition, a sampling frame is a list of all the units of the population of interest. In this insurance case, the sampling frame is the United states hospitals. It is from this frame that we got the population from which we also got the samples.

It is imperative to note that there has to be a sample frame, no matter where the sample is obtained from. The American states give us the cities which would later on be  used in this analysis.

## 2. Data

### a. Field Measurements

To solve the research question, a proper study and consideration of the dataset has to be made as this will help the researcher in the understanding and analysis of the provisional data and sample. For this stage, the variables under consideration are the ages and income along with the gender and cities of residence of the clientele of this insurance company.

The frequency of all this data is 150, 000, however, we will be using the population sample n=30 for the analysis and possible arrival to the necessary conclusions.

### b. Quality Assurance / Quality Control

When collecting any type of data, the integrity of the data being collected is very crucial to the analysis, responses and conclusions bound to be derived from the data. If the quality of the data is low, then the results of the analysis will probably be misleading starting from the sampling error, a result of data without integrity.

While the data used for illustration of this plan was obtained from an open source, we assume that the researchers did go through the data quality control process so as to avoid any biases in the conclusive information.

It must be noted that good quality data is accurate, complete, relevant, available, detailed, and timely.

Some of the ways in which the quality of data can be assured include:

- **Use the data obtained and not cherry pick the data:** this is to avoid any biases in the dataset.
- **Data profiling:** this is to ensure that the data doesn't fail to represent its intended goal. This can be done by reviewing the data collected, division of the data into sets while ensuring the data is thorough and finally ensuring that the summation of the data is a representation of the whole data collected.
- **Data standardization:** helps ensure the consistency in the data collected.
- **Data cleaning:** This is the process of checking for missing data and removal of duplicates.
- 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 are also ways in which the quality of data can be ensured.

### c. Analysis

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

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

- Which city had the highest number of individuals who were aged above 40 and were ill?
- Which city had the highest count of ill people in the united states
- What is the average age of people who are insured
- In the dependent age bracket, which city had the highest number of individuals.

A normality test on the worked data is done to determine whether or not the modelling of the data was done by a normal distribution. It is also done to determine the likelihood of an underlying random variable in the data to have a normal distribution. The goodness of fit scale for example, is a standard measure for normality distributions. This test helps a researcher understand whether their data was drawn from a population that was normally distributed or not, a concept which is also applicable in the dataset from which the n=30 sample was derived.

## 3. Implementation

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

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

Attached is a link to my python notebook which clarifies the method or sampling I opted for in this plan including the dataset I used. The implementation stage of the whole process when the data is collected is a section I did not tackle as the report did not require that.

#Kindly do note that you will **not** be required to perform any implementation within the scope of this IP.