New Light on Dark Universe

Topic: Use of Sampling Techniques in Astronomy.

ISHITA DEODHAR | KARNALI NAIK | MAHIMA CHADHA | RAUNAK GUHA |SANKET PATIL | SAYALI BORA | VISHAKHA PATIL

Email : mahima.chadha@ssi.edu.in

Abstract:

Milky Way Galaxy, a large spiral system consisting of several hundred billion stars, the irregular luminous band of stars and gas clouds that stretches across the sky as seen from Earth. A thick layer of interstellar dust obscures much of the galaxy.

With digital transformation and growing importance of decisions and functions statistics holds a critical position in the process of analysing, controlling, and presenting data. In order to detect structure in astronomical data and to make predictions, statistics provides measures and methods to evaluate insights out of data by getting the right mathematical approach for data.

In this report the results of Descriptive and Inferential statistics performed on 'Star datasets to predict Star types' is reported. Confidence interval for population mean of temperature is evaluated using One Sample Z – test at 5% level of significance (α). Descriptive and Inferential analysis being a significant aspect of statistics was performed using various statistical software including R and Excel.

Introduction:

This report presents a framework for envisioning use of sampling techniques in astronomy. Astronomy is the science that encompasses the study of all extra-terrestrial objects and phenomena. Our universe is both ancient and vast and expanding out farther and faster. The existence of an expanding universe implies that the cosmos has evolved from a dense concentration of matter into the present broadly spread distribution of the galaxy.

Sampling is considered with the selection of a subset of individuals from within a statistical population to estimate the characteristics of the whole population. In a simple sense sampling refers to the method used to select a given number of people or items from a population. Sampling being an integral part of research methodology, refers to the selection of some items from the population which represents the population. This report gives insights on use of sampling techniques in Astronomy. Mainly three sampling techniques are discussed: simple random sampling, stratified sampling and systematic sampling.

Descriptive statistics provide summarizing information of the characteristics and distribution of values in one or more datasets. The classical descriptive statistics allow analysts to have a quick glance of the central tendency and the degree of dispersion of values in datasets. They are useful in understanding a data distribution and in comparing data distributions. While descriptive statistics are simple concepts in statistical analysis, they are important and useful in today's era of big data. With increasing large volumes of data being produced constantly and distributed via Internet, the effectiveness and usefulness of descriptive statistics should not be overlooked. Components of descriptive statistics have briefly been discussed in the report and its graphical representation is shown.

Inferential statistics helps to suggest explanations for a situation or phenomenon. It allows us to draw conclusions based on extrapolations, and is in that way fundamentally different from descriptive statistics that merely summarize the data that has actually been measured. There are many types of inferential statistics and each is appropriate for a specific research design and sample characteristics. Researchers should consult the numerous texts on experimental design and statistics to find the right statistical test for their experiment. However, most inferential statistics are based on the principle that a test-statistic value is calculated on the basis of a particular formula. That value along with the degrees of freedom, a measure related to the sample size, and the rejection criteria are used to determine whether differences exist between the treatment groups. The larger the sample size, the more likely a statistic is to indicate that differences exist between the treatment groups. Thus, the larger the sample of subjects, the more powerful the statistic is said to be.

A Z-test is a type of hypothesis test—a way to figure out if results from a test are valid or repeatable. A hypothesis test tells if it’s probably true, or probably not true. A Z - test, is used when your data is approximately normally distributed.

Purpose:

Astronomers want to research qualitative features of stars such as Absolute temperature, Relative luminosity, Relative Radius, Absolute Magnitude to make confidence intervals for population means, which further can be used in research to study star colour and star types. Inferential statistics is done to find the confidence interval which in turn will provide researchers a range which provides them data to predict star types.

Objectives:

- To broaden the aspects of statistics specifically sampling theory in Astronomy.
- To enhance analytical thinking by applying statistical tools on astronomical data.
- To build conceptual knowledge on how scientists utilize statistical techniques in astronomy to derive meaningful results.
- To collect and analyse data and to apply various sampling techniques.
- To perform descriptive and inferential statistical analysis for comparing the sample and population data.
- To work collectively as a team for improving the efficiency of analysis and performance of the group.

Important Terminologies:

- **Absolute temperature**: The temperature measured using the Kelvin scale where zero is absolute zero.
- **Luminosity:** It is an absolute measure of radiated electromagnetic power (light), the radiant power emitted by a light-emitting object over time.
- **Absolute Magnitude**: The absolute magnitude of a star, *M* is the magnitude the star would have if it was placed at a distance of 10 parsecs from Earth.
- **Spectral class**: Various groups into which stars are classified according to characteristic spectral lines and bands.
- **Star type**: Stars are also classified by their spectra (the elements that they absorb) along with their brightness.
- **Relative Radius**: A measure of the size of the star.

Methodology:

- The data was analysed meticulously and various sampling techniques were studied.
- The population of 240 observations were taken under consideration.
- Three sampling techniques i.e. simple random sampling without replacement, stratified random sampling and systematic sampling were performed on the population.
- Samples of 120 observations each were drawn using excel.
- Evaluation of mean, standard error, standard deviation, sample variance and confidence level of 95% was performed.
- Hypothesis Testing was computed on the samples.
- Data Visualization was done on descriptive as well as inferential statistics using excel and R.

Statistical Analysis:

The dataset consists of several features of the stars such as Absolute temperature, Relative luminosity, Relative Radius, Absolute Magnitude, Star Colour, Spectral Class and Star Type. After this data was collected, it was compiled into usable form; both quantitative and qualitative tools and techniques were used to analyse the data. The performance effectiveness and efficiency of the selected data was assessed.

Sampling Techniques:

Samples of the same size (n = 120) were drawn using three methods.

- **Simple Random Sampling without replacement.**
  In this method of selecting a sample of the population units, every sample of a fixed size is given an equal chance to be selected.
- **Systematic Random Sampling.**
  In this method the population is divided into k groups of size n=N/k in each. One unit is chosen randomly from the first k units and every kth unit following it is included in the sample
- **Stratified Random Sampling**.
  In this method the entire population is divided into homogeneous groups known as strata, and then random samples are selected from each stratum.

Mean: A mean is the simple mathematical average of a set of two or more numbers.

Standard Error: Standard Error is a statistical component that measures the accuracy with which a sample represents a population by using standard deviation.

$$SE_X = \frac{s}{\sqrt{n}}$$

The positive square root of the variance is termed as the standard error of the estimator. The variance of the sample estimator is used to decide the precision of the sample estimator.

Standard Deviation: The standard deviation is a measure of the spread of scores within a set of data.

Variance: The sample variance is used to calculate how varied a sample i.e. sample is a select number of items taken from a population.

Confidence Level: The 95% confidence interval is a range of values that you can be 95% confident contains the true mean of the population.

Statistical tools used: (1) Excel

(2) R Software

Descriptive Statistics:

	Descriptive Statistics	Temperature (K)	Luminosity(L/Lo)	Radius(R/Ro)	Absolute Magnitude(Mv)	Star Type
Population	**Mean**	10497.4625	107188.3616	237.1577814	4.382395833	2.5
Simple Random Sampling	**Mean**	11198.76667	108069.2651	212.5321505	5.111791667	2.391666667
Stratified Sampling	**Mean**	10292.56667	100686.496	246.3414662	4.402291667	2.5
Systematic Sampling	**Mean**	10850.56667	109053.3058	241.8463529	4.293633333	2.5

TABLE 1

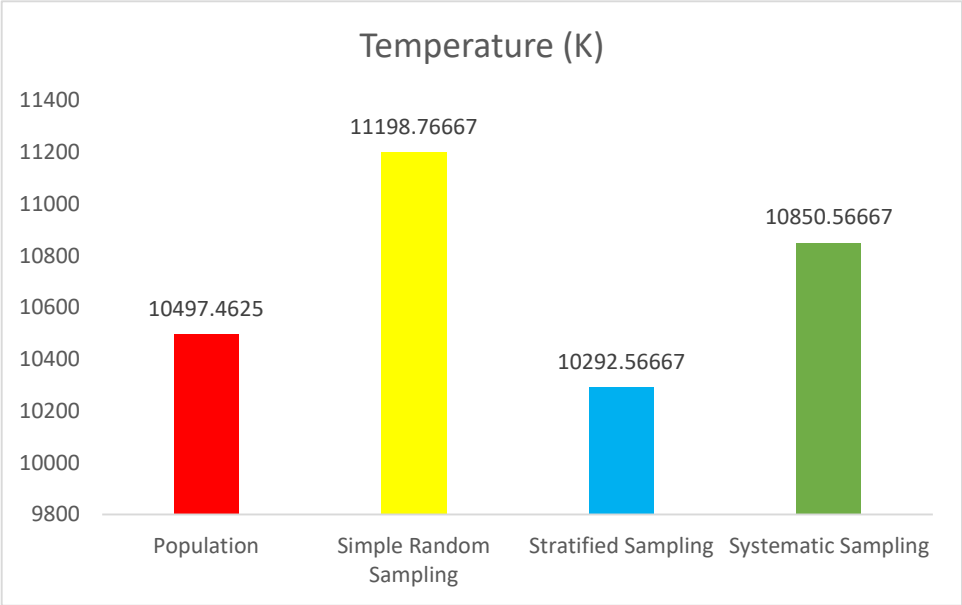


Fig 1.1

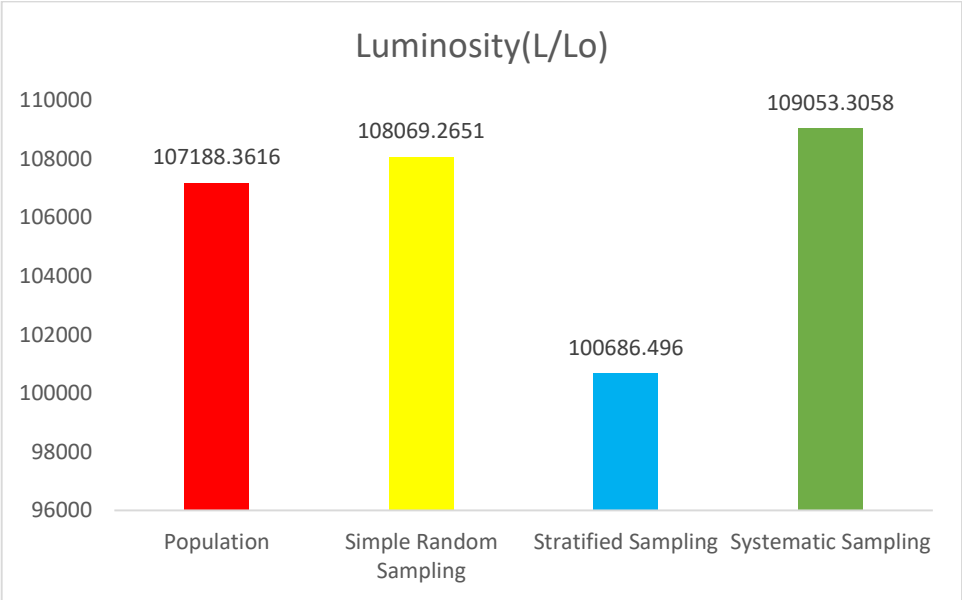

Luminosity(L/Lo)
110000
108000
106000
104000
102000
100000
98000
96000
107188.3616
108069.2651
100686.496
109053.3058
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 1.2

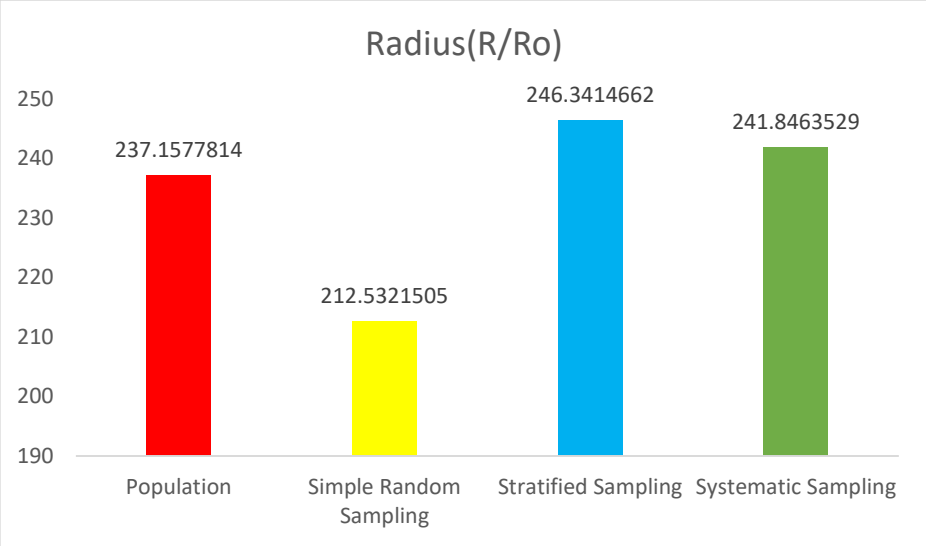

Radius(R/Ro)
250
240
230
220
210
200
190
237.1577814
212.5321505
246.3414662
241.8463529
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 1.3

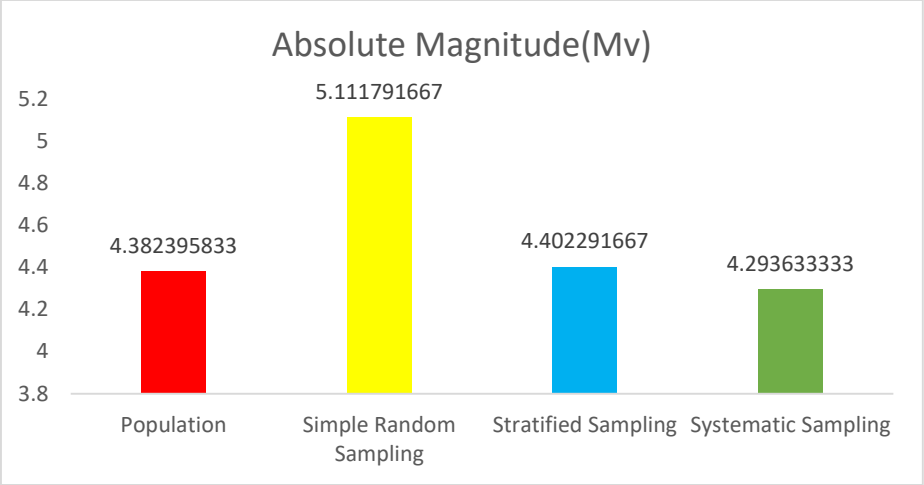

Absolute Magnitude(Mv)
5.2
5
4.8
4.6
4.4
4.2
4
3.8
4.382395833
5.111791667
4.402291667
4.293633333
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 1.4

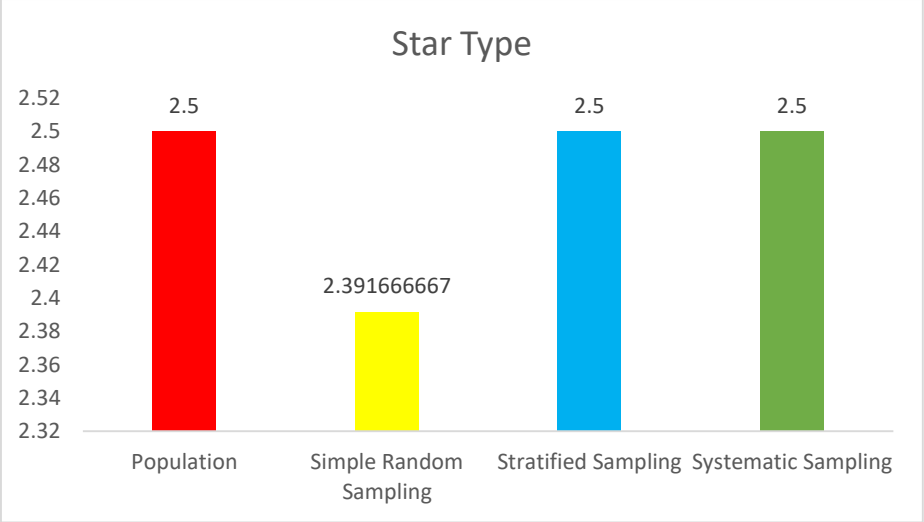

Star Type
2.52
2.5
2.48
2.46
2.44
2.42
2.4
2.38
2.36
2.34
2.32
2.5
2.391666667
2.5
2.5
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 1.5

	Descriptive Statistics	Temperature (K)	Luminosity(L/Lo)	Radius(R/Ro)	Absolute Magnitude(Mv)	Star Type
Population	**Standard Error**	616.6063847	11582.30161	33.38226098	0.679870749	0.110470024
Simple Random Sampling	**Standard Error**	884.0048547	18199.9511	45.08970575	0.967289891	0.155791929
Stratified Sampling	**Standard Error**	835.9157354	15755.93487	49.15065312	0.959213673	0.156556073
Systematic Sampling	**Standard Error**	930.1043975	16945.36513	48.14394476	0.961269562	0.156556073

TABLE 2

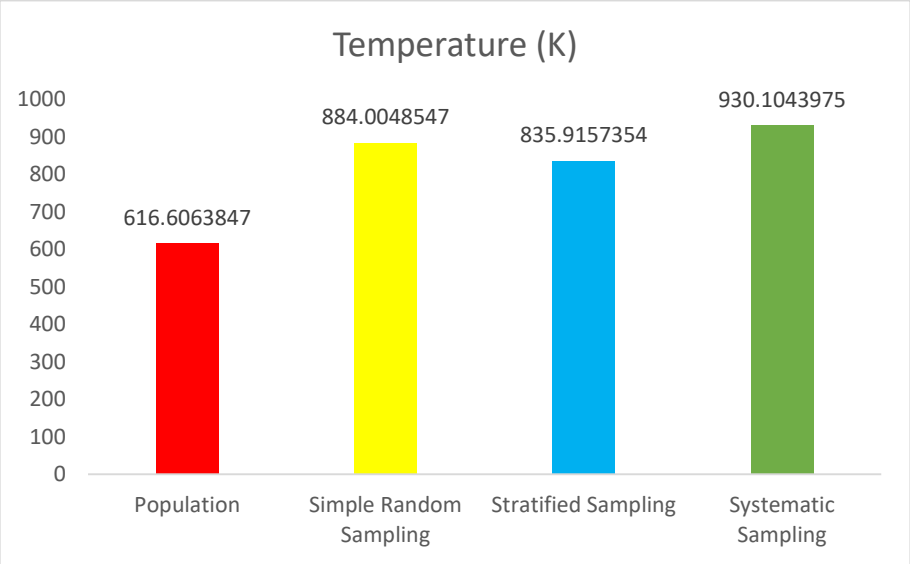


Fig 2.1

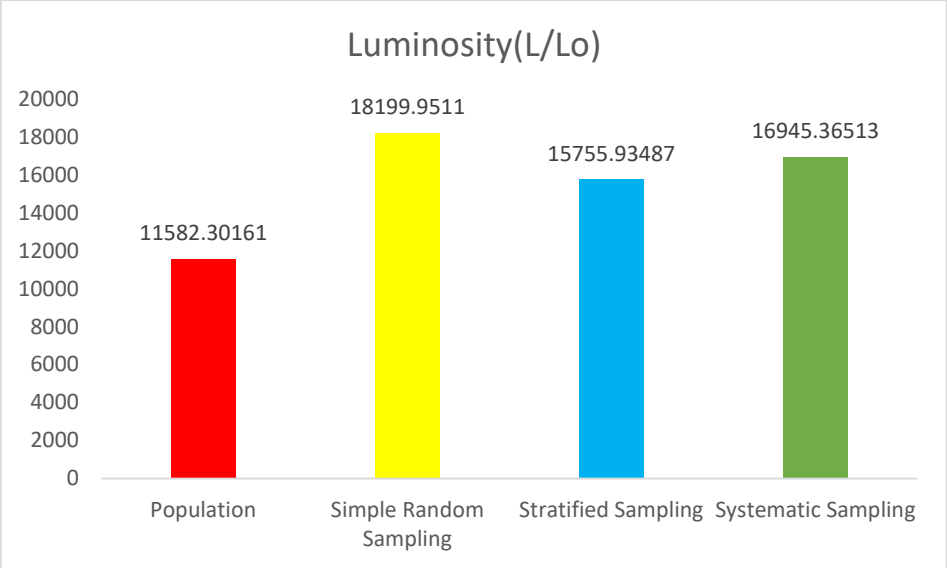

Luminosity(L/Lo)
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
11582.30161
18199.9511
15755.93487
16945.36513
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 2.2

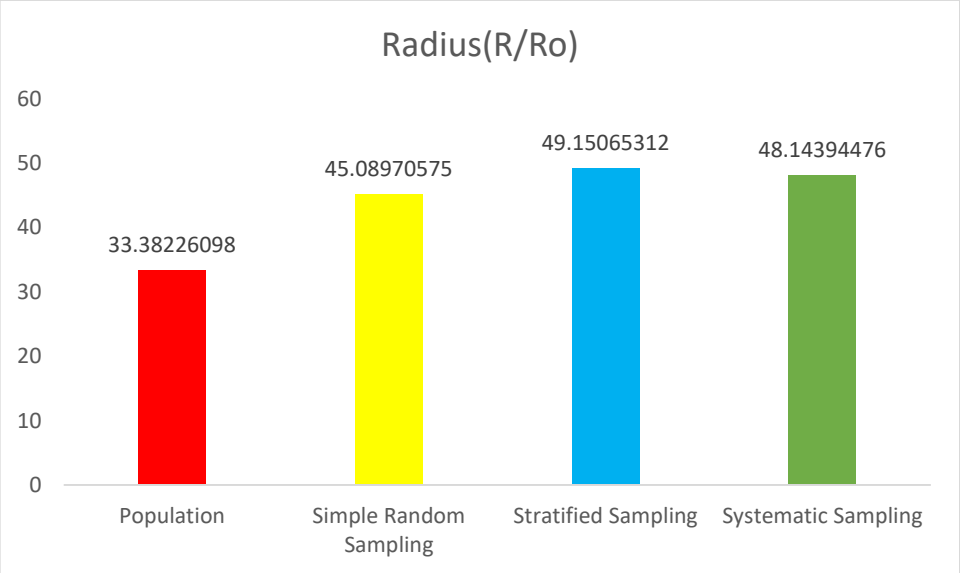

Radius(R/Ro)
60
50
40
30
20
10
0
33.38226098
45.08970575
49.15065312
48.14394476
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 2.3

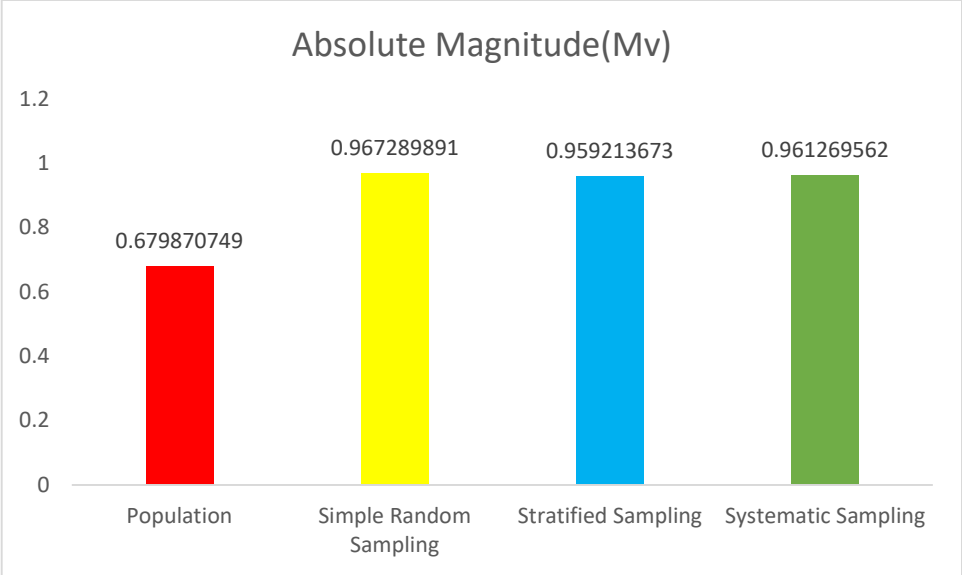

Absolute Magnitude(Mv)
1.2
1
0.8
0.6
0.4
0.2
0
0.679870749
0.967289891
0.959213673
0.961269562
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 2.4

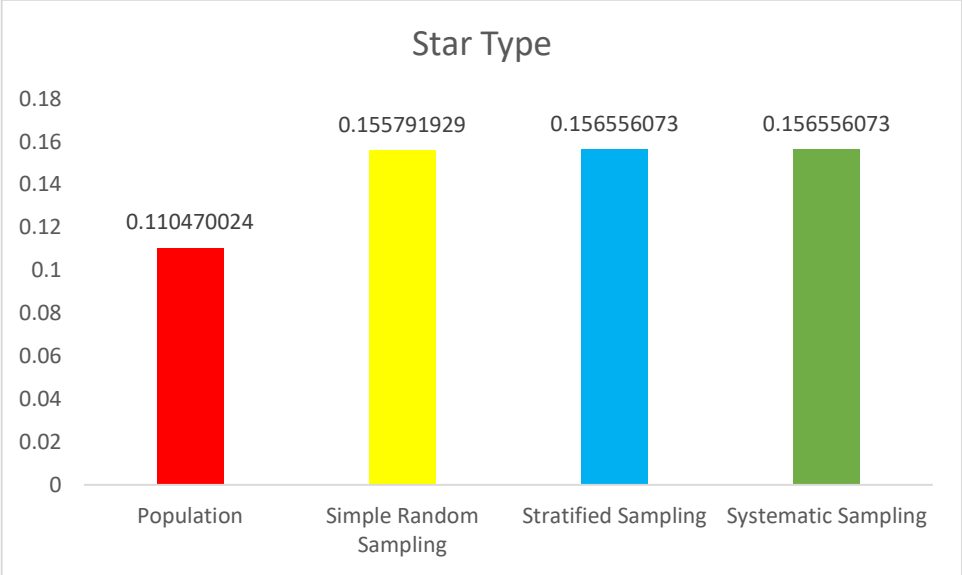

Star Type
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0.110470024
0.155791929
0.156556073
0.156556073
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 2.5

	Descriptive Statistics	Temperature (K)	Luminosity(L/Lo)	Radius(R/Ro)	Absolute Magnitude(Mv)	Star Type
Population	**Standard Deviation**	9552.425037	179432.2449	517.1557634	10.53251235	1.711394254
Simple Random Sampling	**Standard Deviation**	9683.787997	199370.4753	493.932979	10.59612986	1.706615075
Stratified Sampling	**Standard Deviation**	9156.998089	172597.6189	538.4184286	10.50765933	1.714985851
Systematic Sampling	**Standard Deviation**	10188.78319	185627.1746	527.3904911	10.53018046	1.714985851

TABLE 3

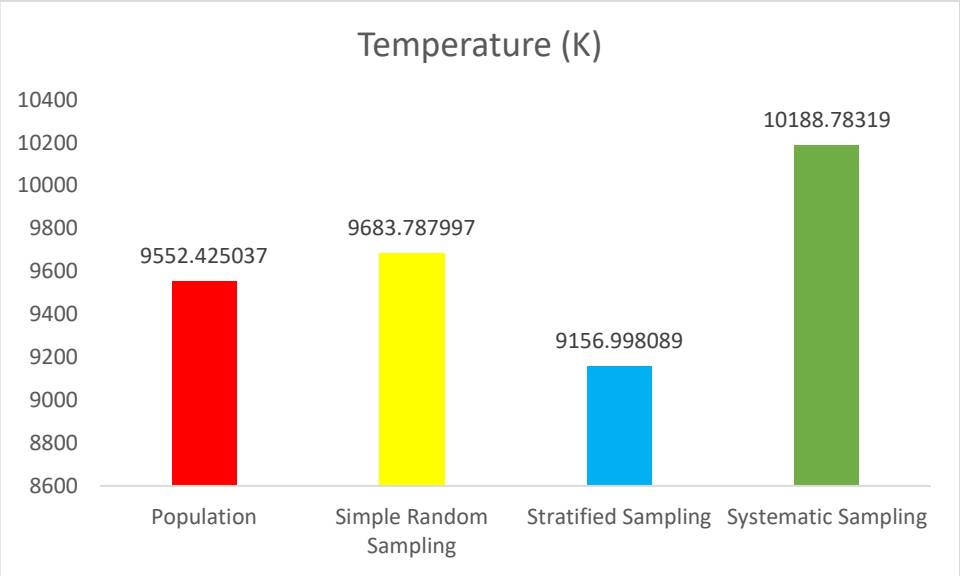


Fig 3.1

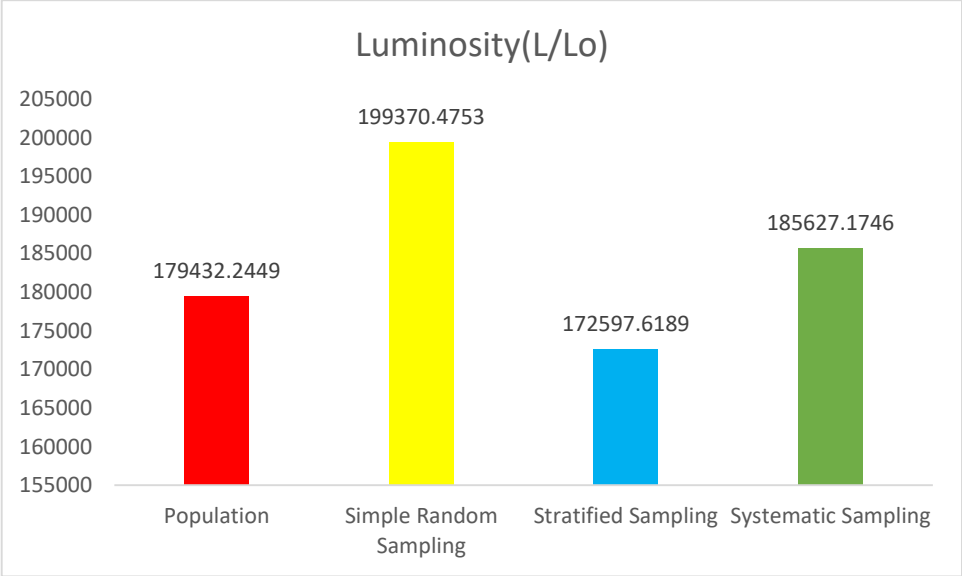

Luminosity(L/Lo)
205000
200000
195000
190000
185000
180000
175000
170000
165000
160000
155000
179432.2449
199370.4753
172597.6189
185627.1746
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 3.2

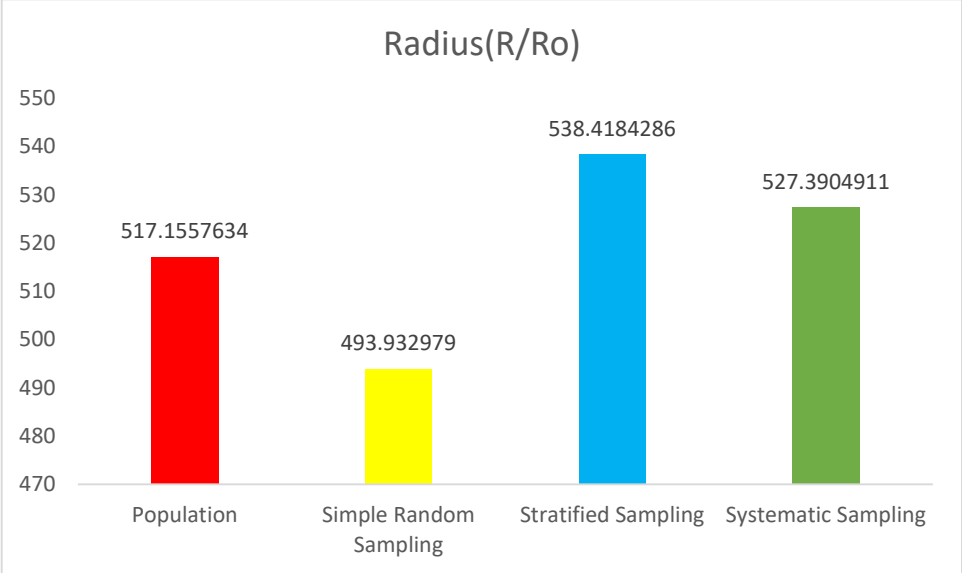

Radius(R/Ro)
550
540
530
520
510
500
490
480
470
517.1557634
493.932979
538.4184286
527.3904911
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 3.3

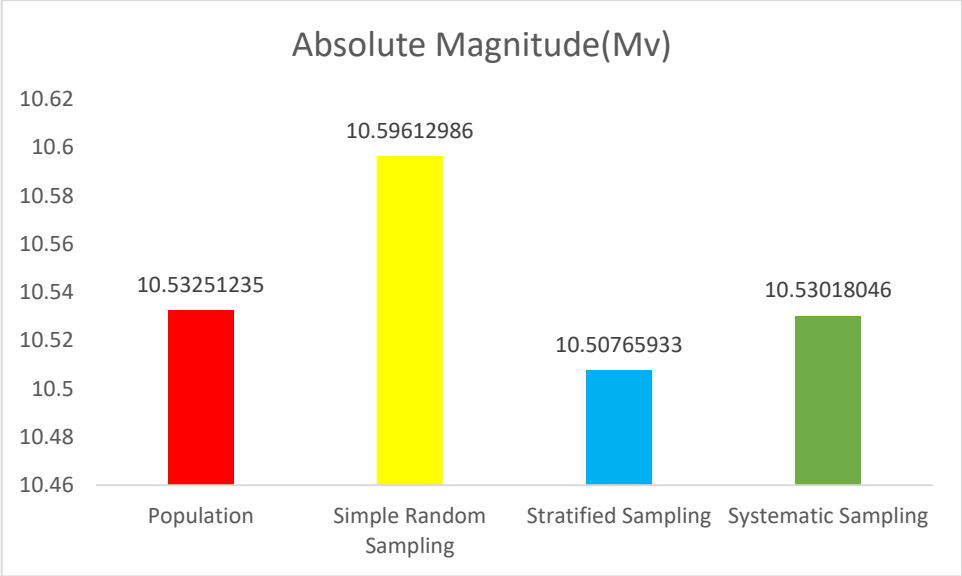

Absolute Magnitude(Mv)
10.62
10.6
10.58
10.56
10.54
10.52
10.5
10.48
10.46
10.53251235
10.59612986
10.50765933
10.53018046
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 3.4

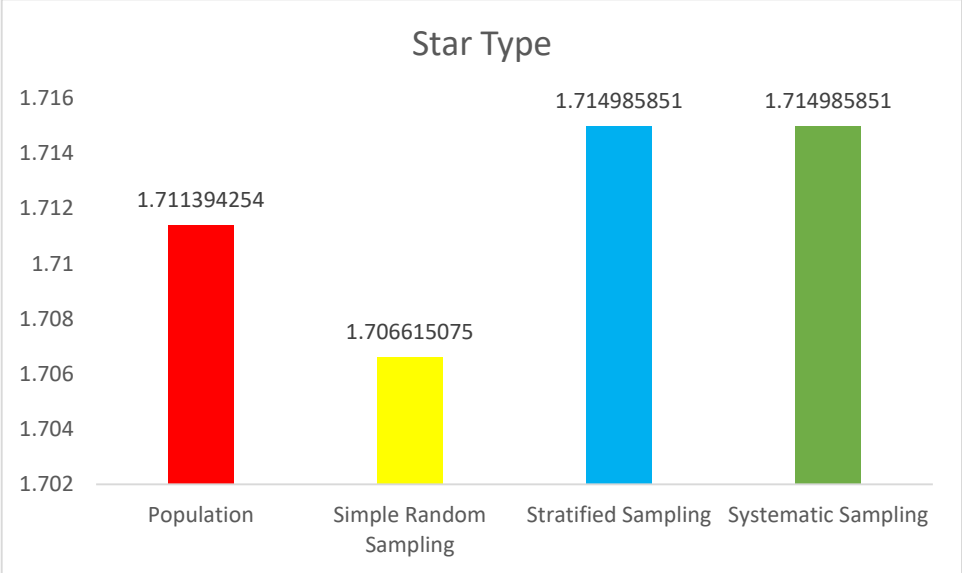

Star Type
1.716
1.714
1.712
1.71
1.708
1.706
1.704
1.702
1.711394254
1.706615075
1.714985851
1.714985851
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 3.5

	Descriptive Statistics	Temperature (K)	Luminosity(L/Lo)	Radius(R/Ro)	Absolute Magnitude(Mv)	Star Type
Population	**Sample Variance**	91248824.09	32195930524	267450.0836	110.9338164	2.928870293
Simple Random Sampling	**Sample Variance**	93775749.98	39748586407	243969.7878	112.2779679	2.912535014
Stratified Sampling	**Sample Variance**	83850614	29789938033	289894.4042	110.4109045	2.941176471
Systematic Sampling	**Sample Variance**	103811302.8	34457447945	278140.7301	110.8847005	2.941176471

TABLE 4

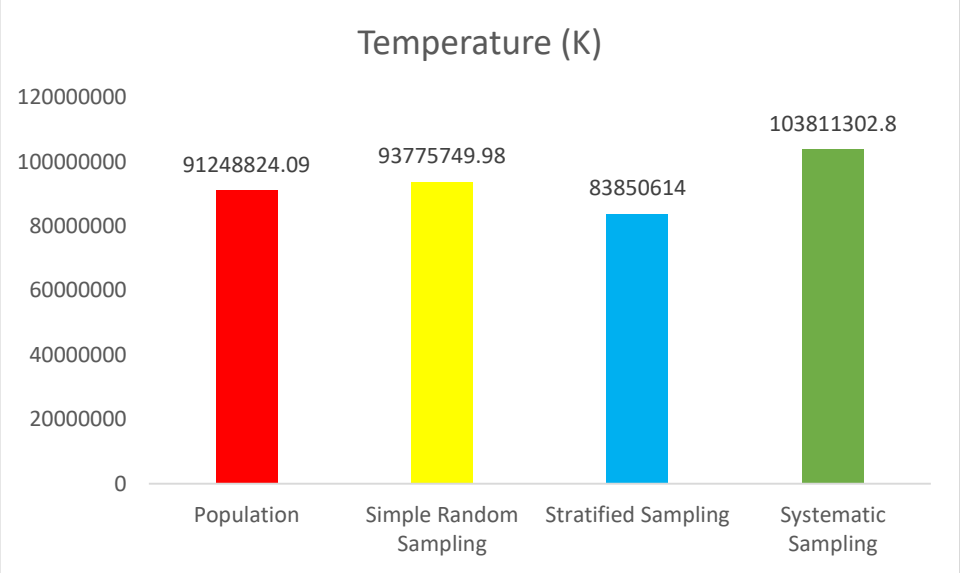


Fig 4.1

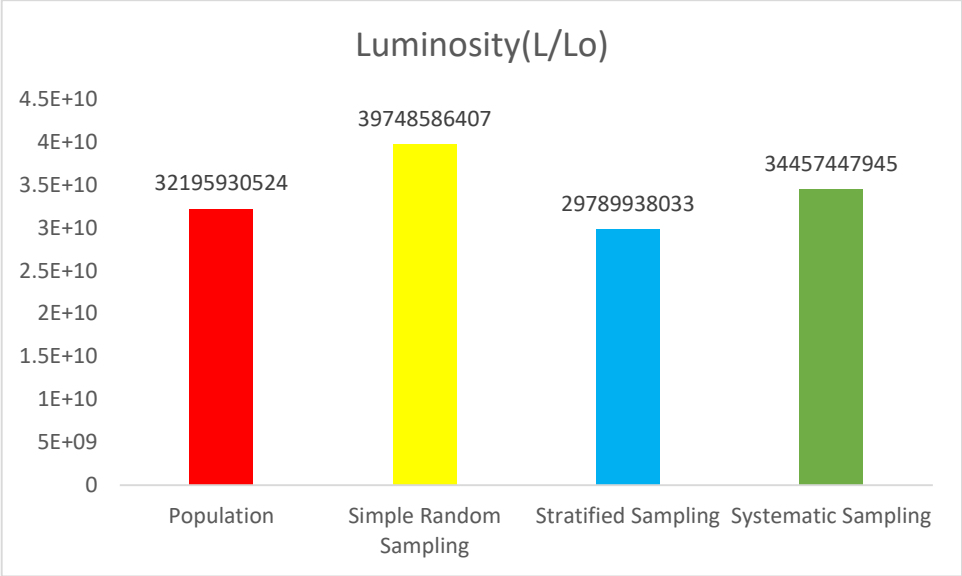

Luminosity(L/Lo)
4.5E+10
4E+10
3.5E+10
3E+10
2.5E+10
2E+10
1.5E+10
1E+10
5E+09
0
32195930524
39748586407
29789938033
34457447945
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 4.2

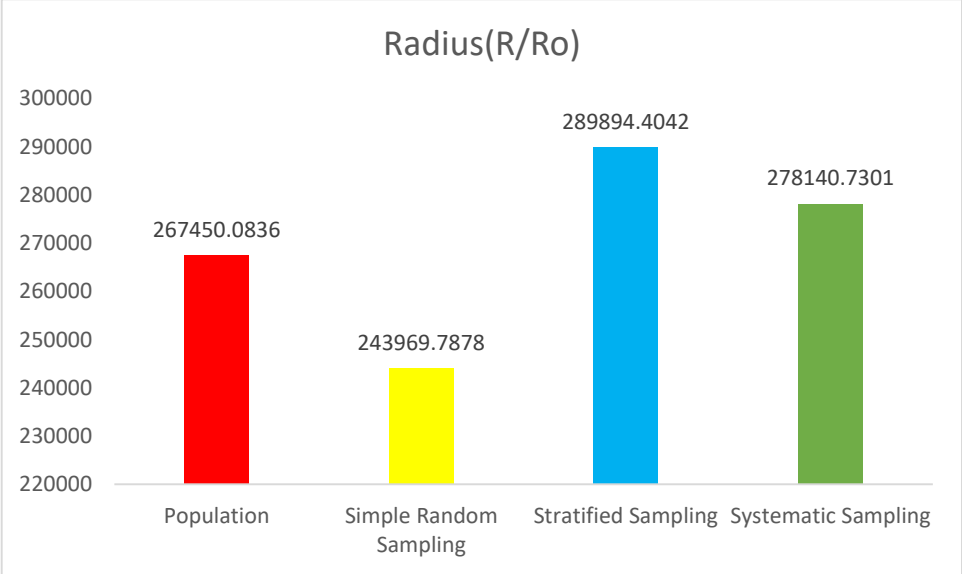

Radius(R/Ro)
300000
290000
280000
270000
260000
250000
240000
230000
220000
267450.0836
243969.7878
289894.4042
278140.7301
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 4.3

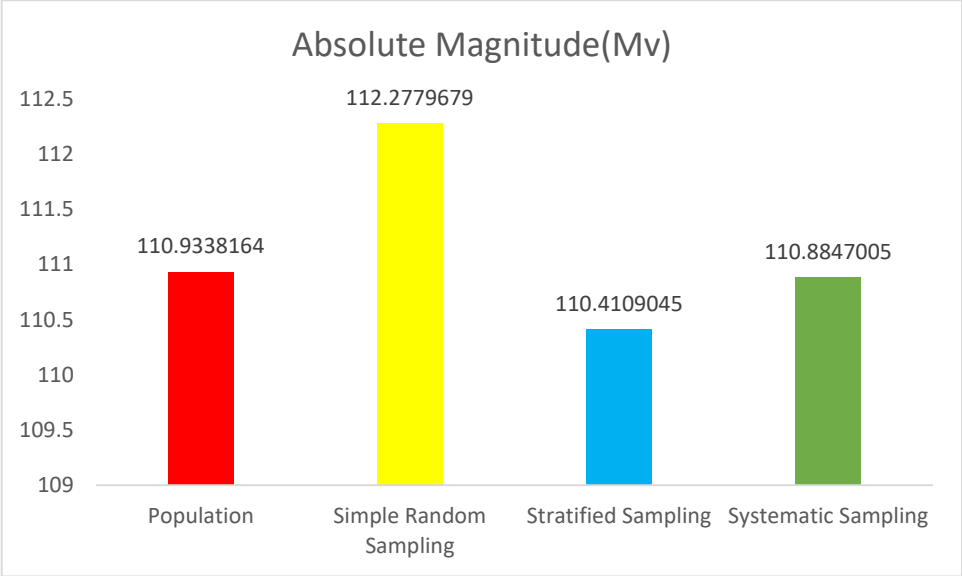

Absolute Magnitude(Mv)
112.5
112
111.5
111
110.5
110
109.5
109
110.9338164
112.2779679
110.4109045
110.8847005
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 4.4

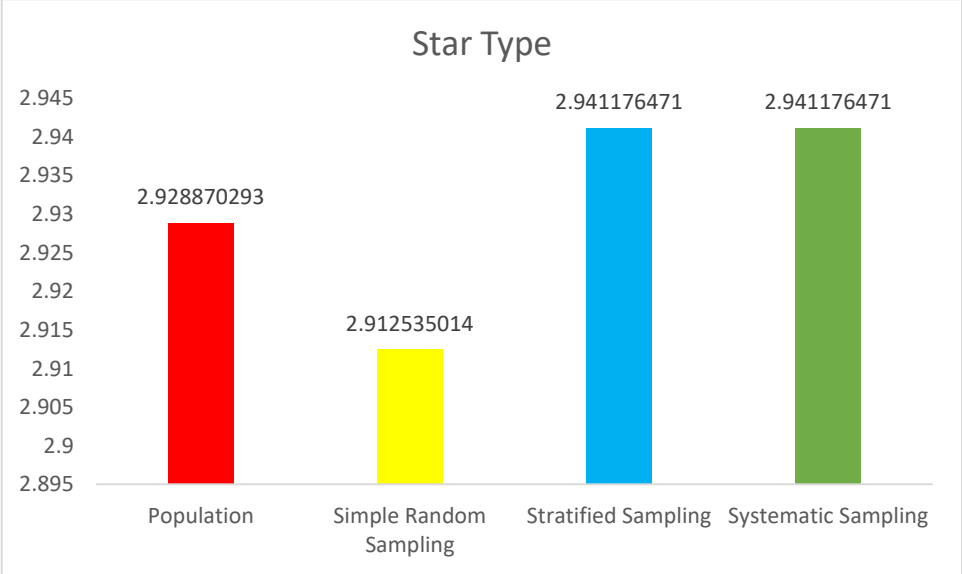

Star Type
2.945
2.94
2.935
2.93
2.925
2.92
2.915
2.91
2.905
2.9
2.895
2.928870293
2.912535014
2.941176471
2.941176471
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 4.5

	Descriptive Statistics	Temperature (K)	Luminosity(L/Lo)	Radius(R/Ro)	Absolute Magnitude(Mv)	Star Type
Population	**Confidence Level (95.0%)**	1214.677215	22816.43235	65.76103137	1.339304179	0.217619254
Simple Random Sampling	**Confidence Level (95.0%)**	1750.417904	36037.72093	89.28212079	1.915330593	0.308483579
Stratified Sampling	**Confidence Level (95.0%)**	1655.196644	31198.32469	97.32320216	1.899338876	0.30999666
Systematic Sampling	**Confidence Level (95.0%)**	1841.699602	33553.51541	95.32981908	1.903409741	0.30999666

TABLE 5

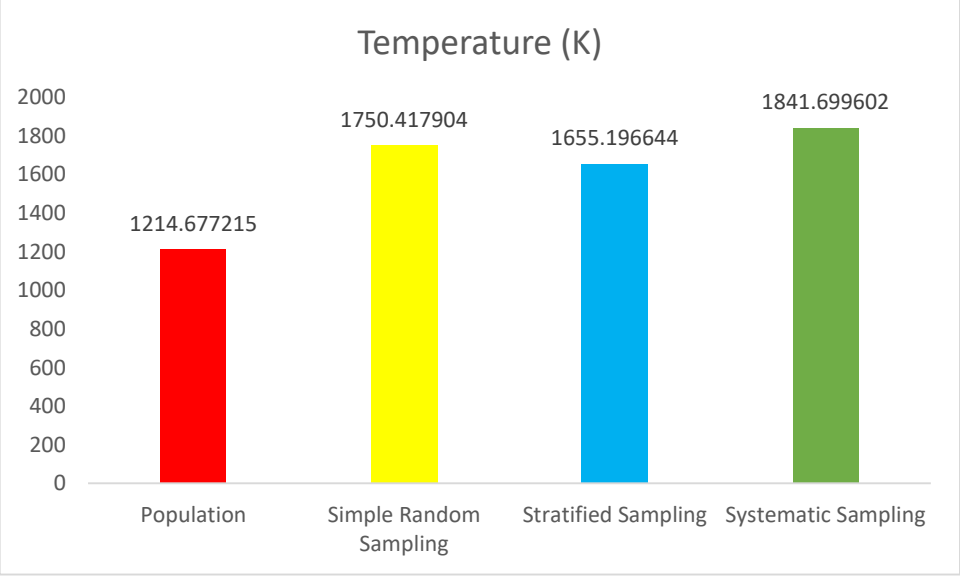


Fig 5.1

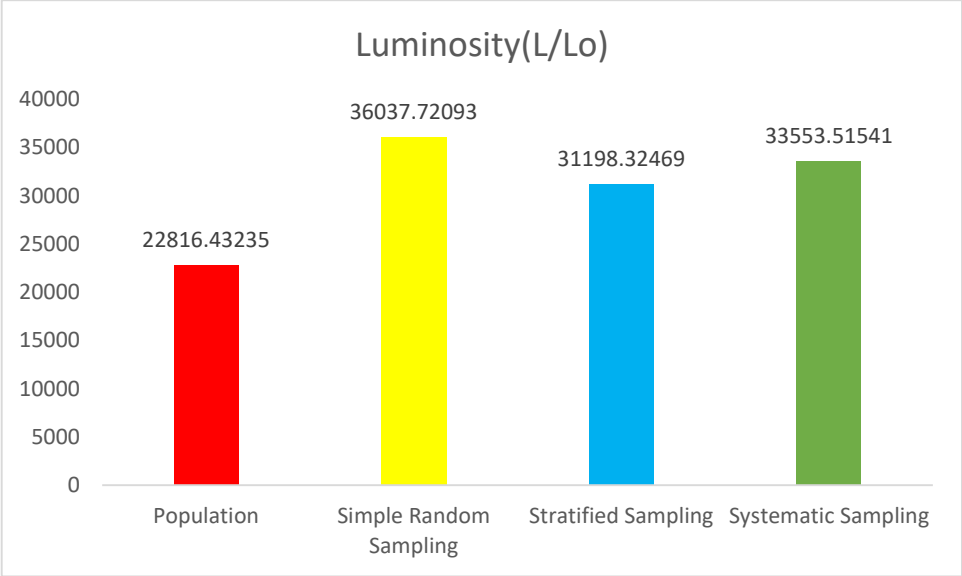

Luminosity(L/Lo)
40000
35000
30000
25000
20000
15000
10000
5000
0
22816.43235
36037.72093
31198.32469
33553.51541
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 5.2

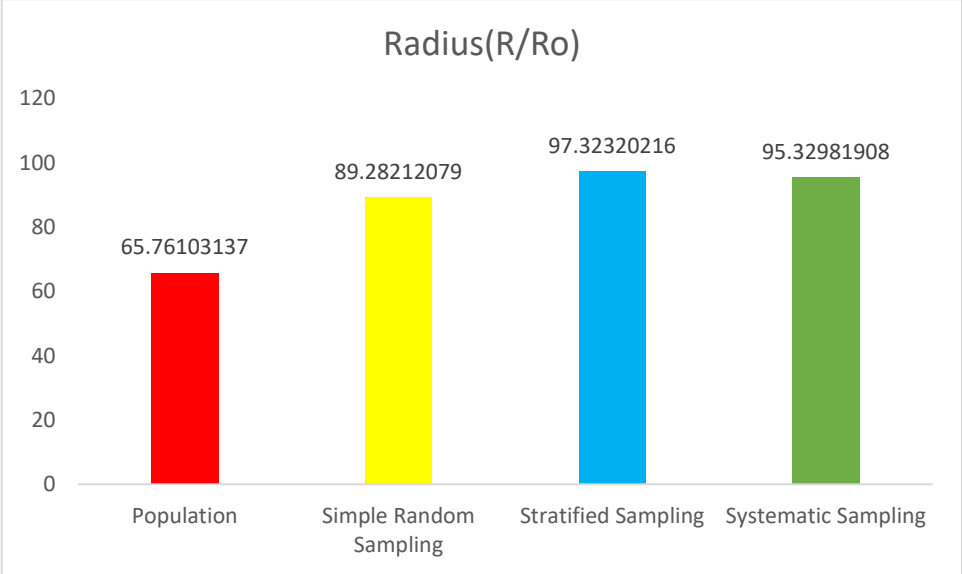

Radius(R/Ro)
120
100
80
60
40
20
0
65.76103137
89.28212079
97.32320216
95.32981908
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 5.3

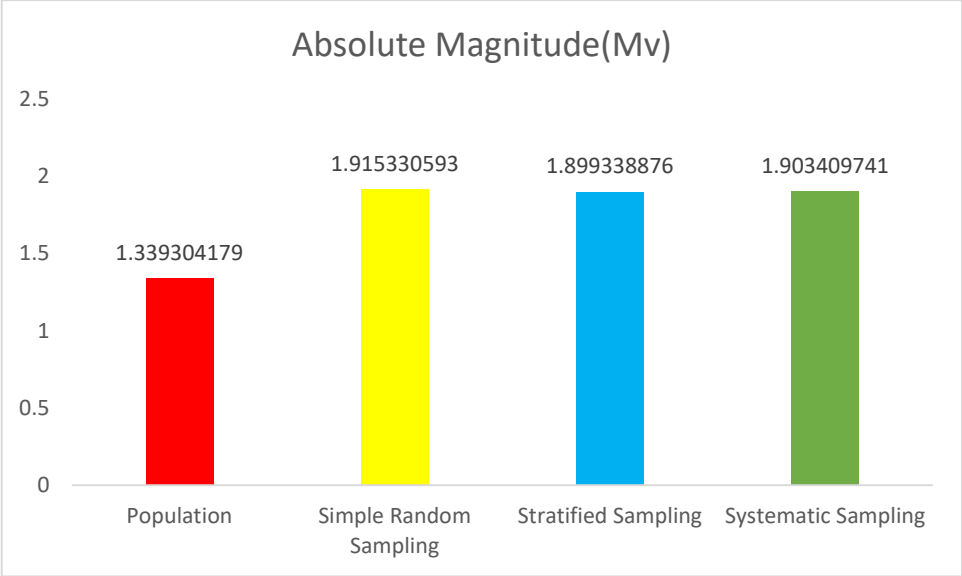

Absolute Magnitude(Mv)
2.5
2
1.5
1
0.5
0
1.339304179
1.915330593
1.899338876
1.903409741
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 5.4

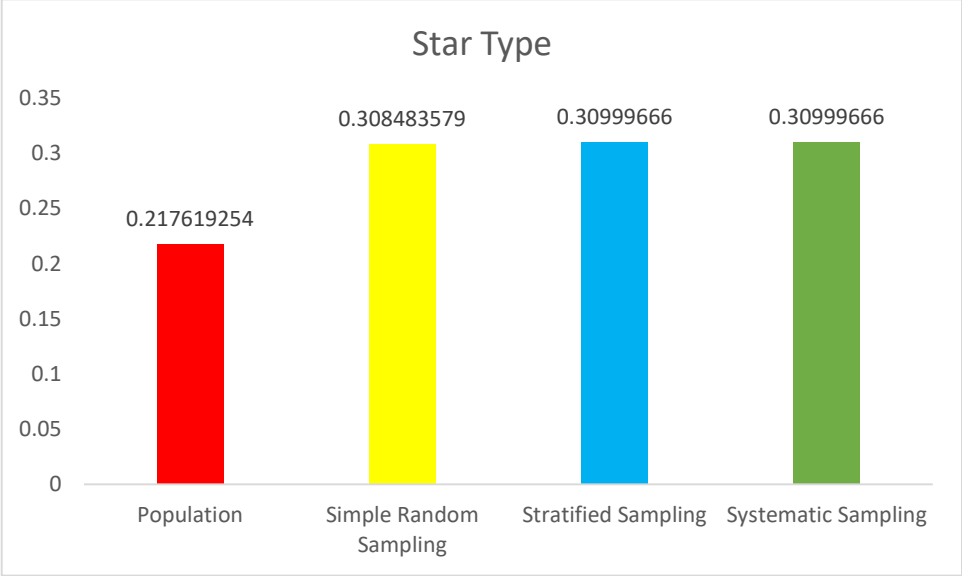

Star Type
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.217619254
0.308483579
0.30999666
0.30999666
Population
Simple Random Sampling
Stratified Sampling
Systematic Sampling

Fig 5.5

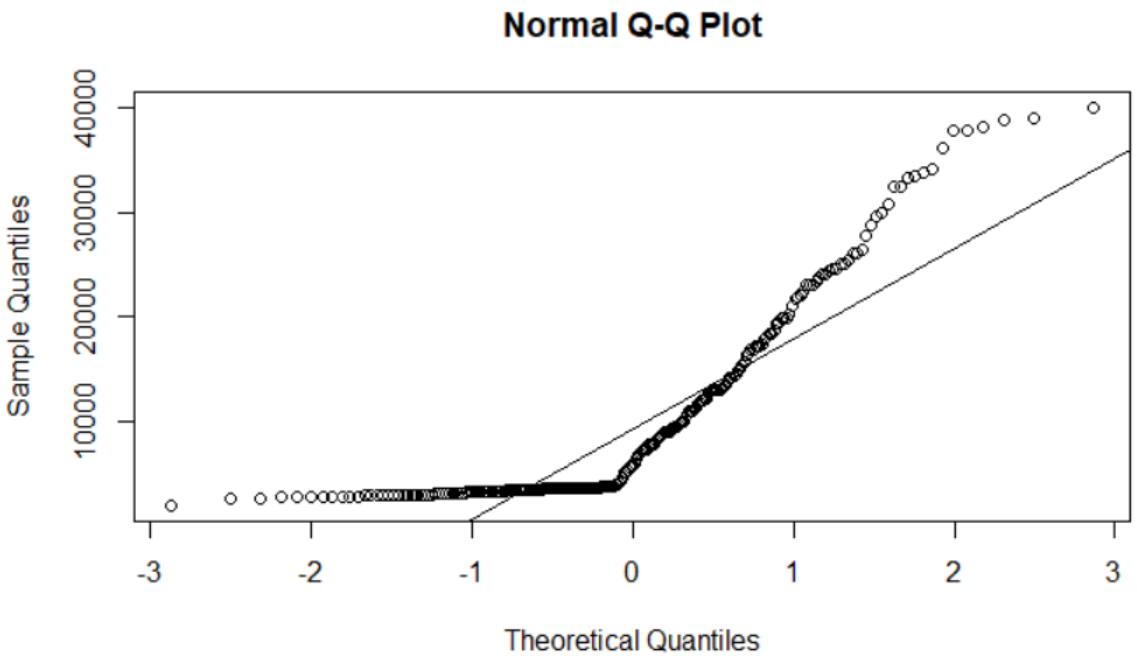
Hypothesis Testing:

H_0: Population mean of temperature = 10000

H_1: Population mean of temperature >10000 (Right Tailed Test)

Researchers are testing whether the population mean of temperature is greater than 10000 Kelvin (K)

```
>stars <- read.csv("C:/Users/Ishita/OneDrive/Desktop/r soft/stars.csv")
>   View(stars)
> temp.org=stars$Temperature..K.
> qqnorm(temp.org)
> qqline(temp.org)
```



The population follows **Normal distribution**

- **Simple Random Sampling**

```
srs <- read.csv("C:/Users/Ishita/OneDrive/Desktop/r soft/srs.csv")
>   View(srs)
```

H_0 : μ = 10000 H_1 : μ >10000

```
> z.test = function(x,mu,stdev){
+   one.tail.p <- NULL
+  z.score <- round((mean(x)-mu)/(stdev/sqrt(length(x))),3)
+   one.tail.p <- round(pnorm(abs(z.score),lower.tail = FALSE),3)
+  cat(" z =",z.score,"\n",
+       "one-tailed probability =", one.tail.p,"\n",
+       "two-tailed probability =", 2*one.tail.p )}
>z.test(x,mu,stdev)
 z = 1.375
 one-tailed probability = 0.085
 two-tailed probability = 0.17
```

- **Stratified Sampling**

```
> stratified <- read.csv("C:/Users/Ishita/OneDrive/Desktop/r soft/stratified.csv")
>   View(stratified)
> tempstra=stratified$Temperature..K.
```

H_0 : μ = 10000 H_1 : μ >10000

```
> z.test = function(x,mu,stdev){
```

```
+   #one.tail.p <- NULL
+  z.score <- round((mean(x)-mu)/(stdev/sqrt(length(x))),3)
+   one.tail.p <- round(pnorm(abs(z.score),lower.tail = FALSE),3)
+  cat(" z =",z.score,"\n",
      "one-tailed probability =", one.tail.p,"\n",
       "two-tailed probability =", 2*one.tail.p )}
> z.test(tempstra,mu,stdev)
 z = 0.336
 one-tailed probability = 0.368
 two-tailed probability = 0.736
```

- **Systematic Sampling**

```
systematic <- read.csv("C:/Users/Ishita/OneDrive/Desktop/r soft/systematic.csv")
>   View(systematic)
> tempsys=systematic$Temperature..K.
```

H_0 : μ = 10000 H_1 : μ >10000

```
> z.test = function(x,mu,stdev){
+   #one.tail.p <- NULL
+  z.score <- round((mean(x)-mu)/(stdev/sqrt(length(x))),3)
+   one.tail.p <- round(pnorm(abs(z.score),lower.tail = FALSE),3)
+  cat(" z =",z.score,"\n",
+       "one-tailed probability =", one.tail.p,"\n",
+       "two-tailed probability =", 2*one.tail.p )}
> z.test(tempsys,mu,stdev)
 z = 0.975
 one-tailed probability = 0.165
 two-tailed probability = 0.33
```

Result:

Descriptive Statistics analysis on all the three samples (Simple Random Sample, Stratified Sample and Systematic Sample) for Mean, Standard Error, Standard Deviation, Sample Variance and Confidence Level (95%) were conducted effectually. The results observed are as follows

	Mean	**Standard Error**	**Standard Deviation**	**Sample Variance**	**Confidence Level (95%)**
Temperature (K)	Stratified Sampling	Stratified Sampling	Simple Random Sampling	Simple Random Sampling	Stratified Sampling
Luminosity (L/L_O)	Simple Random Sampling	Stratified Sampling	Stratified Sampling	Systematic Sampling	Stratified Sampling
Radius (R/R_o)	Systematic Sampling	Simple Random Sampling	Systematic Sampling	Systematic Sampling	Simple Random Sampling
Absolute Magnitude (Mv)	Stratified Sampling	Stratified Sampling	Systematic Sampling	Systematic Sampling	Stratified Sampling
Star Type	Stratified Sampling / Systematic Sampling	Simple Random Sampling	Stratified Sampling / Systematic Sampling	Simple Random Sampling	Simple Random Sampling

The above table shows which sampling technique gives value closest to the population value.

Conclusion:

The purpose of this project was to efficiently use various sampling techniques on a given dataset and analyse various attributes of the data. Simple random sampling, systematic sampling and stratified sampling was successfully applied on the data using the excel software.

Drawbacks:

- Random sampling introduces certain arbitrary parameters which can cause over or under representation of data.
- Systematic sampling might cause data manipulation which might be inclined towards achieving targeted outcomes rather than letting random data produce a representative answer.
- In stratified sampling major attributes that fall into multiple groups have a higher selection chance which might cause misinterpreted analysis.
- The sample size plays an important role in deciding the descriptive analysis of the sample.
- The Z - test does not yield the hypothesis about the value of the population mean.

Future Prospects:

By adapting, developing and improving analytical methods and predictive models, statistics is helping astronomers to harness the full potential of their data. Statistics is a huge domain and every single astrophysical analysis could benefit from sophisticated methods and algorithms. Research and development in improvising the sampling techniques for better results can be done.

Reference:

The dataset is created based on several equations in astrophysics. They are given below:

- Stefan-Boltzmann's law of Black body radiation (To find the luminosity of a star).
- Wien’s Displacement law (for finding surface temperature of a star using wavelength).
- Absolute magnitude relation.
- Radius of a star using parallax.
- The missing data were manually calculated using those equations of astrophysics given above.
- Link for data

https://www.kaggle.com/deepu1109/star-dataset