Novelty Journals

ISSN 2394-9651

International Journal of Novel Research in Physics Chemistry & Mathematics
Vol. 5, Issue 1, pp: (1-9), Month: January - April 2018, Available at: www.noveltyjournals.com

The Statistical Sampling Methods and Their Effection to the Estimation

[1]Monzer Boubou, [2]Mohammad Dribati, [3]Hadel Darwish

[1,2,3] Tishreen University, Latakia, Syria

Abstract: **This research aims to study three methods for the random sampling (Simple, Stratified, Cluster), and the impact of these methods on the accuracy of the estimation of the statistical population parameters, as an example, we will study the estimation of the linear regression coefficients by using Least Squares Method, that we compare estimators according to the three methods of random sampling that we mentioned above by applying the mathematical equations for each method, where it is clear that the stratified random sampling with proportional allocation gives the most efficient estimators and less sampling error.**

Keywords: **Sampling Methods، Simple Random Sampling، Stratified Sampling، Cluster Sampling، Linear Regression، Least Squares Method**.

I. INTRODUCTION

Because it is difficult, costly and sometimes impossible to conduct research on the whole population, we choose a representative sample for the population to estimate population parameters. And because of the sensitivity of OLS estimators for non-normal distribution and outliers, we will try to improve these estimators depending on statistical sampling methods.

II. BODY OF ARTICLE

Basic Definitions in Sampling Theory:

Statistical Population: in statistics, population refers to the total set of observations that the elements from a set of data.

Statistical Sample: it's a part of population which consists one or more observations drawn from the population.

Sample Method: Is a procedure for selecting sample elements from a population.

Simple Random Sampling Refers to a sampling method that has the following properties:

- The population consists of N objects.
- The sample consists of n objects.
- All possible samples of n objects are equally likely to occur.

An important benefit of simple random sampling is that it allows researchers to use statistical methods to analyze sample results. For example, given a simple random sample, researchers can use statistical methods to define confidence interval around a sample mean. Statistical analysis is not appropriate when non-random sampling methods are used.

Sampling methods can be classified into one or two categories:

- **Probability Sampling:** Sample has a known probability of being selected.

- **Non-Probability Sampling:** Sample does not have known probability of being selected as in convenience or voluntary response surveys.

Probability Sampling:

In probability sampling it is possible to both determine which sampling units belong to which sample and the probability that each sample will be selected. In this research we had studied three types of random statistical sampling:

- Simple Random Sampling.
- Stratified Sampling.
- Cluster Sampling.

Simple Random Sampling:

A simple random sampling is a subset of a statistical population in which each member of the subset has an equal probability of being chosen.

Stratified Sampling:

Stratified sampling is possible when it makes sense to partition the population into groups based on a factor that may influence the variable that is being measured. These groups are then called strata.

An individual group is called a stratum.

With stratified sampling one should:

- Partition the population into groups (strata).
- Obtain a simple random sample from each group (stratum).
- Collect data on each sampling unit that was randomly sampled from each group (stratum).

Stratified sampling works better when a heterogeneous population is split into fairly homogeneous groups. Under these conditions, stratification generally produces more precise estimates of the population percents than estimates that would be found from a simple random sample.

The statistical (indicators) studied in the population in the random stratified sampling depend on partition the statistical sample on the strata.

Equal Allocation:

In this method, the stratified sample size is divided into strata equally.it means that the size in each stratum is equal:

$$n_1 = n_2 = n_3 = \cdots = n_L$$

The sample size in each stratum is:

$n_h = \frac{n}{L}$; L: the number of strata.

Proportional Allocation:

In this method, the stratified sample size is divided into strata evenly with the strata's sizes, this allocation is required:

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = \cdots = \frac{n}{N}$$

Cluster Sampling:

Cluster sampling refers to a type of sampling method. With cluster sampling, the researcher divides the population into separate groups, called clusters. Then, a simple random sample of clusters is selected from the population. The researcher conducts his analysis on data from sampled clusters.

Compared to simple random sampling and stratified sampling, cluster sampling has advantages and disadvantages. For example, given equal sample sizes, cluster sampling usually provides less precision than either simple random sampling or stratified sampling. On the other hand, if travel costs between clusters are high, cluster sampling maybe more cost-effective than the other methods.

Cluster Sampling is very different from Stratified Sampling with cluster sampling one should:

✓ Divide the population into groups (clusters).

✓ Obtain a simple random sample of so many clusters from all possible clusters.

✓ Obtain data on every sampling unit in each of the randomly selected clusters.

It is important to note that, unlike with strata in stratified sampling, the clusters should be microcosms, rather than subsections, of the population. Each cluster should be heterogeneous. Additionally, the statistical analysis used with cluster sampling is not only different, but also more complicated than that used with stratified sampling.

When we use this method, we should consider:

1- Cluster size should be small, and the number of clusters big.

2- When these clusters are formed, the elements of a contiguous society are taken or within a particular region where they are often similar to the studied character.

3- Each cluster should have an explanation and known for the data collector.

In this research, we had used Least Square Linear Regression Estimators as examples to compare between the previous three sampling methods to find the "Best Linear Unbiased Estimator", "Best" means "minimum variance" or "smallest variance".

Simple Linear Regression:

It is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables.

- One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.
- The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

The necessary OLS assumptions, which are used to derive the OLS estimators in linear regression models, are discussed below.

OLS Assumption 1: The linear regression model is "linear in parameters."

When the dependent variable (Y) is a linear function of independent variables (X's) and the error term, the regression is linear in parameters and not necessarily linear in X's.

OLS Assumption 2: There is a random sampling of observations

This assumption of OLS regression says that:

- The sample taken for the linear regression model must be drawn randomly from the population. For example, if you have to run a regression model to study the factors that impact the scores of students in the final exam, then you must select students randomly from the university during your data collection process, rather than adopting a convenient sampling procedure.
- The number of observations taken in the sample for making the linear regression model should be greater than the number of parameters to be estimated.

- The X's should be fixed (e. independent variables should impact dependent variables). It should not be the case that dependent variables impact independent variables. This is because, in regression models, the causal relationship is studied and there is not a correlation between the two variables
- The error terms are random. This makes the dependent variable random.

OLS Assumption 3: The conditional mean should be zero.

The expected value of the mean of the error terms of OLS regression should be zero given the values of independent variables.

Mathematically, $E = (\varepsilon/X) = 0$

This is sometimes just written as $E(\varepsilon) = 0$

In other words, the distribution of error terms has zero mean and doesn't depend on the independent variables X's. Thus, there must be no relationship between the X's and the error term.

OLS Assumption 4: There is no multi-collinearity (or perfect collinearity).

In a simple linear regression model, there is only one independent variable and hence, by default, this assumption will hold true. However, in the case of multiple linear regression models, there are more than one independent variable. The OLS assumption of no multi-collinearity says that there should be no linear relationship between the independent variables.

OLS Assumption 5: Spherical errors: There is homoscedasticity and no autocorrelation.

According to this OLS assumption, the error terms in the regression should all have the same variance.

Mathematically, $Var(\varepsilon/X) = \sigma^2$

If this variance is not constant (i.e. dependent on X's), then the linear regression model has heteroscedastic errors and likely to give incorrect estimates.

This OLS assumption of no autocorrelation says that the error terms of different observations should not be correlated with each other.

Mathematically, $Cov(\varepsilon_i \varepsilon_j/X) = 0 \; for \; i \neq j$

OLS Assumption 6: Error terms should be normally distributed.

This assumption states that the errors are normally distributed, conditional upon the independent variables. This OLS assumption is not required for the validity of OLS method; however, it becomes important when one needs to define some additional finite-sample properties. Note that only the error terms need to be normally distributed. The dependent variable Y need not be normally distributed.

Least Squares Estimators:

The aim of OLS is to find estimators to β_0 and β_1 in linear regression equations.

By using mathematical techniques, we calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ when minimizing Q to minimum.

$$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

We can calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ which make Q as small as possible by Partial derivation for the Q according to β_0 and β_1:

$$\frac{dQ}{d\beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\mathrm{d}Q}{\mathrm{d}\beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - n\,\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i\, y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

$$\sum_{i=1}^{n} y_i = n\,\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i\, y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2$$

$$\frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} {x_i}^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{S_x^2}$$

$$\hat{\beta}_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i\right) = \bar{y} - \hat{\beta}_1 \bar{x}$$

When

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$$

Collecting and Analyzing Population Data:

In this research we had got an experimental data of population consists of 2000 elements, which represent random results of tests of students from different schools of physics and mathematics.

Before starting to analyze population data, we conducted quality tests to reconcile the model using SPSS program (Statistical Package for Data Analysis).

We find that:

- Person correlation coefficient $r = 0.993$ and this indicates a strong correlation between the two variables.

- Determination coefficient $R^2 = 0.871$ and this indicates that the linear relationship explains 87% of data and total deviations in Y.

Therefore, we can say that the mode is good to reconcile and represents a linear relationship and the results can be generalized.

Novelty Journals

ISSN 2394-9651

International Journal of Novel Research in Physics Chemistry & Mathematics
Vol. 5, Issue 1, pp: (1-9), Month: January - April 2018, Available at: www.noveltyjournals.com

By conducting regression analysis using SPSS on the whole population, we get the following linear regression equation:

$$\hat{Y} = 0.115 + 0.987\,x$$

So: $\hat{\beta}_0 = 0.115$ and $\hat{\beta}_1 = 0.987$

As shown in figure (1)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.115	.191		.600	.548	-.261-	.490
	x	.987	.009	.933	116.059	.000	.971	1.004

a. Dependent Variable: y

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.933[a]	.871	.871	1.191

a. Predictors: (Constant), x
b. Dependent Variable: y

Figure (1)

We assumed that $n = 325$ and we will look for the sample which gives the nearest regression equation to the previous equation.

Simple Random Sampling:

By using SPSS, we randomly select a simple random sample of $= 325$, we analyze this sample we get this regression equation:

$$\hat{Y} = 0.191 + 0.98x$$

With standard errors: $\delta\beta_0 = 0.076$, $\delta\beta_1 = 0.007$

And average error: $\bar{\delta} = 0.0415$

As shown in figure (2)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.191	.475		.401	.688	-.744-	1.125
	x	.980	.021	.933	46.474	.000	.939	1.022

a. Dependent Variable: y

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.933[a]	.870	.870	1.239	.870	2159.876	1	323	.000

a. Predictors: (Constant), x
b. Dependent Variable: y

Descriptive Statistics

	Mean	Std. Deviation	N
y	22.03	3.429	325
x	22.28	3.262	325

Figure (2)

Novelty Journals

ISSN 2394-9651

International Journal of Novel Research in Physics Chemistry & Mathematics
Vol. 5, Issue 1, pp: (1-9), Month: January - April 2018, Available at: www.noveltyjournals.com

Stratified Sampling:

First of all, we divided the population to 5 clusters according to geographic regions, whereas the size of each stratum:

$N_h = 400 \ ; \ h = 1,2, \dots ,5$

We distribute the stratified sample to the strata according to the equal allocation:

$n_h = \frac{325}{5} = 65 \ ; \ h = 1,2, \dots ,5$

And by using SPSS we randomly select samples from each stratum using simple random sampling (because each stratum become homogeneous population), then we formed the stratified sample with size $n = 325$ and by applying the equations of stratified sampling we got the following regression equation:

$$Y = 0.12 + 1.003x$$

$$\beta_0 = 0.12 \quad \& \quad \beta_1 = 1.003$$

$$Std.\beta_1 = 0.017 \quad \& \quad Std.\beta_0 = 0.388$$

$$R^2 = 0.918$$

And the biased error:

$$\delta\beta_0 = 0.005 \ \delta\beta_1 = 0.016$$

$$\overline{\delta} = 0.0105$$

$$\overline{x} = 22.42 \quad \& \quad \overline{y} = 22.184$$

As population is divided to equals strata, the equal allocation is the same proportional allocation with strata sizes.

As a second step while we studying stratified sample, we divided the population to three strata according to the students results in physics, whereas the data of each stratum is close to each other, and thus less data deviation.

We assumed that the students who get results less than 20 in physics, got the rating "good", and the stratum size $N_1 = 624$.

And the students who get results between 20 and 25 in physics get the rating "very good", and the stratum size $N_2 = 1064$.

And the third stratum which represent students who get the rating "excellent" results more than 25 in physics, $N_3 = 312$.

We distribute the stratified sample to the three strata according to the equal allocation

$$n_1 = n_2 = n_3 = \frac{n}{3} \cong 108$$

Then we select samples from each stratum with size 108 using simple random sampling, and we formed the stratified sample with size $\cong 325$, and by applying the equations of stratified sampling we got:

$$Y = 0.11 + 1.002x$$

$$\beta_0 = 0.11 \quad \& \quad \beta_1 = 1.002$$

$$Std.\beta_1 = 0.019 \quad \& \quad Std.\beta_0 = 0.425$$

$$\delta\beta_1 = 0.015 \ \delta\beta_0 = 0.005$$

$$\overline{\delta} = 0.01$$

$$\overline{x} = 22.38 \quad \& \quad \overline{y} = 22.19$$

Then we distribute the stratified sample to the strata according to proportional allocation as follow:

$$\frac{325}{2000} = \frac{n_1}{624} = \frac{n_2}{1064} = \frac{n_3}{316}$$

$$n_1 = 101$$
$$n_2 = 173$$
$$n_3 = 51$$

We select the samples from the strata with the previous sizes using simple random sampling and formed the stratified sample $= 325$, and by applying the equations of stratified sampling we got:

$$Y = 0.116 + 1.002x$$

$$\beta_0 = 0.116 \quad \& \quad \beta_1 = 1.002$$

$$Std.\beta_1 = 0.019 \quad \& \quad Std.\beta_0 = 0.425$$

$$\delta\beta_0 = 0.015 \quad \& \quad \delta\beta_1 = 0.001$$

$$\overline{\delta} = 0.008$$

$$\overline{x} = 22.3 \quad \& \quad \overline{y} = 22.1$$

Cluster Sampling:

We divided the population which represent 2000 elements to clusters according to schools, whereas each cluster represents school, we got 20 clusters each one includes some of population units, and by using simple random sampling we select 5 clusters, i.e. the sample size $m = 5$.

$N_h = 10$; $h = 1,2,\dots,5$

Then we select preliminary sample from each cluster using simple random sampling with size

$n_j = 65$; $j = 1,2,\dots,5$

And then we formed the whole sample from this samples with size $n = 325$ and by applying the equations of cluster sampling we got:

$$Y = 0.3 + 0.974x$$

$$\beta_0 = 0.3 \quad \& \quad \beta_1 = 0.974$$

$$Std.\beta_1 = 0.021 \quad \& \quad Std.\beta_0 = 0.475$$

$$\delta\beta_0 = 0.185 \quad \& \quad \delta\beta_1 = 0.013$$

$$\overline{\delta} = 0.099$$

$$\overline{x} = 21.956 \quad \& \quad \overline{y} = 21.731$$

To confirm the result, we select the class sample with size $m = 5$ and then select new preliminary samples and form new sample and got the following results:

$$Y = 0.597 + 0.955x$$

$$\beta_0 = 0.597 \quad \& \quad \beta_1 = 0.955$$

$$Std.\beta_1 = 0.021 \quad \& \quad Std.\beta_0 = 0.471$$

$$\delta\beta_0 = 0.482 \quad \& \quad \delta\beta_1 = 0.028$$

$$\overline{\delta} = 0.225$$

$$\overline{x} = 21.99 \quad \& \quad \overline{y} = 21.686$$

III. CONCLUSION

1-The Simple Random Sample gives estimators that are not accurate enough, because the sample selection is completely random, and estimates are less accurate when variable increases and outlier values exist.

2-Stratified Sample gives the best and most accurate estimators to estimate linear regression coefficient using OLS, whereas the standard error of the sample generally affects by population elements dispersion which is selected from, so the standard error for the stratified sample is less than the standard error of simple random sample due to removing part of the statistical population dispersion by removing big difference in the same stratum. And when we design the sample according to proportional allocation, we get the best sampling results and the lowest standard error and biggest number of data on regression linear.

3-The results of Cluster Sampling are less accurate than Simple and Stratified sampling, due to the whole sample which represents cluster population is selected from specific categories (clusters) in the population, although the selection is random and simple, only number of categories (clusters) do not exist in the analysis, thus, its data are not taken while calculate the estimation of regression equation coefficient.

REFERENCES

[1] A. Vetro, H. Sun, P. DaGraca, and T. Poon, “Minimum drift architectures for three-layer scalable DTV decoding,” IEEE Transaction on Consumer Electronics, Vol. 44, No. 3, pp. 527-536, Aug. 1998.

[2] M. Young, the Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[3] Abdi, H, “The Method of Least Squares”, University of Texas, Dallas, USA, Neil Salkind, 2007.

[4] Abbott, M.G, “Statistical Properties of the OLS Coefficient Estimators”, 2002.

[5] Dufour, J. M, “Coefficients of determination”, McGill University, 2011.

[6] Hall, G, “Pearson’s correlation coefficient”, UK, 2015.

[7] Kaw, A, “Statistics Background of Regression Analysis”, University of South Florida, USA, 2000.

[8] Kuan, C. M, “INTRODUCTION TO ECONOMETRIC THEORY”, Institute of Economics Academia Sinica, Taipei, Taiwan, 2000.

[9] Neter, J. Wasserman, W. Kutner, M. “Applied Liner Statistical Models: Regression, Analysis of Variance and Experimental Designs”, Richars, D. Irwin Inc, 1990.

[10] Park, M, “Regression Estimation of The Mean In Survey Sampling”, USA, IOWA State University, 2012.

[11] Sen, P.K, “Estimates of the Regression Coefficient Based on Kendall's Tau”, University of North Carolina, Chapil Hill, USA, 2013.

[12] Singh, D. Chaudhry, F.S, “Theory and Analysis of Sample Survey Designs”, Caroga Singh, 2nd edition, John Wiley & Sons Inc, 2017.

[13] Xitao, F. Thompson, B. Lin, W, “The Effects of Sample Size, Estimation Methods, and Model, Specification On Structural Equation Modeling Fit Indxes”, USA, Routledge, 1999.