International Journal of Advanced Statistics and Probability, 12 (1) (2025) 12-16

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP
https://doi.org/10.14419/dp63md84
Research paper

Hidden properties in the main sampling techniques

Mohammad Fraiwan Al-Saleh [1]*, Hashim Abdallah Jarrah [1]

[1] *Department of Statistics, Yarmouk University-Jordan*
**Corresponding author E-mail: m-saleh@yu.edu.jo*

Abstract

Several important and hidden properties in the main sampling techniques, Simple, Stratified, Cluster and Systematic random samplings, are highlighted. A new result about Pearson correlation coefficient of all pairs of the values of the variable of interest for the population elements is explored. The highlighting of these properties may have a real contribution to educational statistics.

Keywords: *Correlation Coefficient; Ordered Population; Periodic Population; Random Sampling Techniques Similar within; Similar Among.*

1. Introduction

In simple words, Statistics is used to understand a population based on a portion of it. It is the science of collecting, organizing and summarizing information about the variable of interest in a representative sample from the underlying population, making inference about the population and provide the accuracy of the inference. The accuracy depends mainly on the sample size and the technique used to choose the sample. The main four well known random sampling techniques are the Simple, Stratified, Cluster and Systematic.
Casella and Berger (1990) discussed some properties of the sample variance and provided different ways of writing its formula. Al-Saleh (2005) discussed the confusion about the $(n-1)$-divisor in the formula of the sample variance. Al-Saleh (2007) highlighted some interesting facts about the Poisson distribution, based on the fact that its mean and variance are equal. Al-Saleh and Yousif (2009) discussed some important properties of the standard deviation such as its maximum value and its relation to the mean absolute deviation, they provided some supporting arguments for the $(n-1)$ divisor in the formula of the sample variance. Al-Saleh (2012, 2015) listed and discussed several statistical/mathematical concepts, definitions and facts that are usually not mentioned in classrooms or mentioned without enough details. The formula of the variance of the sample mean for a simple random sample can provide a useful upper or lower bound of the true variance of the sample mean of systematic random sample, (see Schaeffer, Mendenhall et al. (2006)). Heinbockel (2012) discussed some properties of the arithmetic mean, median, geometric mean and harmonic mean. Balakrishnan and Balasubramanian (2007) considered Sen's Inequality and some of its properties. The Digits of π and their Randomness was discussed by Heien and Rahman (2005). A rich learning statistical lesson in the Cauchy distribution is recently discussed by Al-Saleh and Maabreh (2025).
In this paper, the main concentration is on some of the hidden properties of the main four sampling techniques. Some of the considered properties are either never mentioned in classrooms or mentioned without giving enough details about them. Given more details about these properties may increase the motivation of students toward learning. A very interesting fact about the value of Pearson correlation coefficient ρ of all pairs of the values of the variable of interest is explored. A short description of the main sampling techniques and a close look at their properties is the content of section 2. Conclusion remarks are given in section 3.

2. Main sampling techniques and their important properties

The four main sampling techniques are the simple, stratified, cluster and systematic random sampling. The decision on the choice of the technique depends mainly on the population of interest and the objects of the study. Below is a description and a highlighting of the main properties of each technique.

a) Simple Random Sampling (SRS)

Assume that the population consists of N elements; $\Omega=\{e_1,e_2,...,e_N\}$. A sample is any non -empty subset of Ω . We have $\binom{N}{n}$ subsets(samples) of size n each. A sample of size n from a population of size N is called "simple random sample", SRS, if it is chosen in a method that gives all possible samples of size n the same chance of being chosen, i.e. each sample of size n has probability $=1/\binom{N}{n}$ of being the chosen one.

Some Properties of SRS:

i) If e is any element in the population then

$P(e \in chosen\ sample) = \binom{1}{1}\binom{N-1}{n-1} / \binom{N}{n} = \frac{n}{N}.$

Thus, the probability that any element in the population will be in the chosen SRS of size n is n/N =inclusion probability. Most of the students give the answer $1/N$; the correct answer is n/N because each element has n chances to be chosen.
This condition of equal inclusion probability of the elements is not a sufficient but only necessary condition for a sample to be SRS, see Systematic random sampling below.

ii) A sample is any nonempty subset of the population. Thus, the total number of samples is $2^N - 1$. Using the binomial theorem: For any positive integer N , we have $(a+b)^N = \sum_{k=0}^{N} \binom{N}{k} a^k b^{N-k}$ If $a = b = 1$ then $\sum_{k=1}^{N} \binom{N}{k} = 2^N - 1$Thus, the total number of samples is:

$\sum_{k=1}^{N} \binom{N}{k} = 2^N - 1 = \binom{N}{1} + \binom{N}{2} + ... + \binom{N}{N}$.

iii) If a sample of size n elements is taken from a population of size N as follows: A SRS of size m is taken from the population then a SRS of size n is taken from the m chosen elements. The obtained sample is SRS.

Proof: Let A be any subset of n elements in the population. A is the chosen sample if the elements of A are among the m chosen elements and the n chosen elements are from the m chosen elements:

P (A is the chosen sample) = P (A is a subset of the set of m elements of the first step and is the chosen sample from the m elements in the second step) which is

$\left[\binom{n}{n}\binom{N-n}{m-n} / \binom{N}{m}\right](1/\binom{m}{n}) = 1/\binom{N}{n}.$

For example: A SRS of size $n = 10$ can be chosen from a population of size 1000 by: taking a SRS of size 200 from the population and then taking a SRS of size 10 from the 200 chosen elements.
The above technique can be also used if we want to draw a SRS in 3 or more steps.

iv) If $X_1, X_2, ..., X_n$ are the values of the variable of interest for the members of the chosen sample, then the population mean μ is usually estimated by($\hat{\mu}$):

$\hat{\mu} = \bar{X}$ and $Var(\hat{\mu}) = \frac{\sigma^2}{n}(\frac{N-n}{N-1})$.
(See Schaeffer et. al. (2006)).

$$E(S^2) = E\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right) = \frac{\sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2}{n-1} = \frac{N}{N-1}\sigma^2$$

Thus, the sample variance, S^2 , is not an unbiased estimator of the population variance σ^2 ; $\frac{N-1}{N}S^2$ is an unbiased estimator of σ^2 . Therefore, $Var(\hat{\mu})$ is estimated by $\frac{S^2}{n}(1-\frac{n}{N})$. Thus, for finite population, the division of $\sum_{i=1}^{n}(X_i - \bar{X})^2$ by $(n-1)$ doesn't make it an unbiased estimator of σ^2 ; an argument which is usually used to justify the division by $(n-1)$. The following estimator is more suitable estimator of σ^2 :

$$\hat{\sigma}^2 = \left(\frac{N-1}{N}\right)\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}.$$

b) Stratified Random Sampling

In this method, the population is partitioned into $(L \geq 2)$ subgroups (pairwise disjoint and exhaustive subgroups). The subgroups are called Strata. The L subgroups are suitable Strata if their elements are similar within and different among. Similarity within of a subgroup is measured by its variance and the difference among subgroups is measured by their coefficient of variations, CVs. A stratified random sample consists of L random samples, one from each stratum. For stratified sampling to be better than simple random sampling, the variance of the elements in each stratum should be significantly smaller than that of the population and the strata have significantly different CVs.
For example: Suppose a population is partitioned in two subgroups(strata). Let $u_1, u_2, ..., u_{N_1}$ and $w_1, w_2, ..., w_{N_2}$ be the values of the variable of interest for the first stratum and second stratum respectively. Now, let

$$\mu_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} u_i,\ \mu_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} w_i,\ \mu = \frac{N_1\mu_1 + N_2\mu_2}{N}.$$

μ_1, μ_2, μ are respectively, the mean of the 1st stratum, the 2nd stratum and the population. Also, the variance of stratum 1 and 2 are:

$$\sigma^2_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} u^2_i - \mu_1^2 \ \&\ \sigma^2_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} w^2_i - \mu_2^2.$$

Thus, $\sum_{i=1}^{N_1} u^2_i = N_1\sigma^2_1 + N_1\mu_1^2 \ \&\ \sum_{i=1}^{N_2} w^2_i = N_2\sigma^2_2 + N_2\mu_2^2.$

Now, the variance of the population, σ^2, can be written as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N_1}\left(u_i^2 + w_i^2\right) - \mu^2 = \frac{1}{N}\sum_{i=1}^{N_1}\left(u_i^2 + w_i^2\right) - \frac{1}{N^2}\left(N_1\mu_1 + N_2\mu_2\right)^2 = \frac{N_1}{N}\sigma_1^2 + \frac{N_2}{N}\sigma_2^2 + \frac{N_1N_2}{N^2}(\mu_1 - \mu_2)^2.$$

- If $\mu_1 = \mu_2$ then $\sigma^2 = \frac{N_1}{N}\sigma_1^2 + \frac{N_2}{N}\sigma_2^2 \Rightarrow$ either σ_1^2 *or* σ_2^2 is larger than σ^2 (because it is a convex combination of σ_1^2 & σ_2^2. Thus, for this situation, this sampling technique may not be suitable to use.
- Since suitable strata are similar within and different among, it is reasonable for the chosen sample to be only a portion from each stratum.
- If a population consists of N elements and the numerical value of the variable of interest is "a" for N_1 of them and "b" of the remaining $N_2 = N - N_1$ elements then $\mu = (N_1a + N_2b)/N$, we have only two unknowns; a & b. If we take a SRS of size $n = \max(N_1, N_2) + 1$, then we can obtain μ exactly and the variance is zero. For example, if $N_1 = 35, N_2 = 40$, and $n = 41$then

$$\hat{\mu} = \frac{N_1a + N_2b}{N_1 + N_2} = \mu \Rightarrow \mathrm{var}(\hat{\mu}) = 0.$$

On the other hand, if the population is divided into 2 strata, one contains the elements with the 'a' values and the other contains the elements with the 'b' values then, using a stratified sample of size $n = 2$ ($n_1 = n_2 = 1$), we can obtain μ exactly and the variance is zero.

Examples of Suitable Stratification:

1) To study the smoking habit of adults, it is suitable to stratify the population of adults based on gender: Males stratum and Females Stratum.
2) "Poor People" and "Rich People" are two suitable strata, if the study is about income.
3) The following two examples were taken from The Holly Quran:

- " وَكَذَٰلِكَ أَوْحَيْنَا إِلَيْكَ قُرْآنًا عَرَبِيًّا لِتُنْذِرَ أُمَّ الْقُرَىٰ وَمَنْ حَوْلَهَا وَتُنْذِرَ يَوْمَ الْجَمْعِ لَا رَيْبَ فِيهِ ۚ فَرِيقٌ فِي الْجَنَّةِ وَفَرِيقٌ فِي السَّعِيرِ "
- "Wherein is no doubt - a party in Paradise, and a party in the Blaze".
- (فَأَصْحَابُ الْمَيْمَنَةِ مَا أَصْحَابُ الْمَيْمَنَةِ وَأَصْحَابُ الْمَشْأَمَةِ مَا أَصْحَابُ الْمَشْأَمَةِ وَالسَّابِقُونَ السَّابِقُونَ)وَكُنتُمْ أَزْوَاجًا ثَلَاثَةً" Three Categories"

For more details, see https://tanzil.net

c) Cluster Random Sampling

This method is used for a population that consists of N-groups called "clusters". Let N_i be the size of group i ; $i = 1,2,...,N$ (N_i are fixed but may not be known). For the groups to be suitable clusters, their elements should be different within and similar among. A cluster sample consists of all elements in a random sample of n clusters. Since the clusters are different among and similar within, it is reasonable to take all elements in some of the clusters of the population.

For example:

1) The fingers of our hands are suitable two clusters because each hand has different fingers and the two hands are similar.
2) Bunches of grapes are suitable cluster.
3) Regiments of graduates.
4) {3,4,5,2,6,8} and {300,304,305,302,306,308} are two suitable strata while, {30,70,100,175,188} and {31,75,105,177,185} are two suitable clusters.
5) If a population consists of 50 schools, then taking 20 students at random from each school is a stratified sample, while taking all students in 5 chosen schools is a cluster sample.

d) Systematic Random Sampling

A systematic sample is obtained by selecting a starting point randomly from the first k-elements, for a suitable k, and then include in the sample every k^{th} element thereafter. Usually k is obtained by dividing the population size N by the required sample size n. The obtained sample is called a 1 – in – k systematic sample. This method is suitable if the frame (a list of all elements) of the population are not available prior to sampling.

For example: if we have a population of size N=50 and we want to select a random sample of size n=10, then k=50/10= 5, therefore we choose a starting point from the first 5 elements and include every 5th element thereafter. If the starting point is 4 then the elements of the chosen 1 – in – 5 systematic sample are the:

4th, 9th, 14th, 19th, 24th, 29th, 34th, 39th, 44th, 49th

- Choosing a 1 – in – k systematic random sample is equivalent to dividing the population into k groups and choosing one of them at random. Thus, in this case a systematic sample is a cluster sample of size 1 cluster.
- The suitability of the use of systematic sampling technique depends on the type of the population's data. If the data is ordinal, then this method is better than SRS. If the data is periodic then SRS is better to use. If the data is random, then the two sampling techniques have similar performance.
- The probability of any of the k subsets in the population to be the chosen sample is $\frac{1}{k}$. The probability of any of the remaining $\left(\binom{N}{n} - k\right)$ subsets to be the chosen sample is zero. Therefore, Systematic sample is not a simple random sample. However, the probability of any element in the population to be in the chosen sample is $\frac{n}{N} = \frac{1}{k}$; the same as in SRS. Thus, a technique that guarantees equal inclusion probability of all elements of the population is not necessary a simple random sampling technique.
- For a systematic sample of size n, it is well known that: $Var(\bar{X}) = \frac{\sigma^2}{n}(1 + (n-1)\rho)$, where ρ is the Pearson correlation coefficient between all pairs within each possible sample: (a_i, a_j), $i \neq j$. Note that if $k = N/n$ then we will have k possible samples. (See Schaeffer et al. (2006), Page (249).

For example: Assume that an artificial population consists of: 9, 7, 10, 17, 20, 27, we want to select a systematic random sample of size n=3, then k=2. We have two possible samples: 9, 10, 20 or 7, 17, 27. Thus, $\bar{X}$ has 2 values: 13 and 17 each with prob. 0.5. Therefore,

$\mu = \sum_{i=1}^{6} u_i / 6 = 15$, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} u_i^2 - \mu^2 = \frac{1}{6}(1648) - 15^2 = \frac{149}{3}$.

$Var(\bar{X}) = 4 = \frac{\sigma^2}{n}(1+(n-1)\rho) \Rightarrow \rho = -0.3791946309.$

The possible pairs of (X,Y) are: $\{(9,10),(9,20),(10,9),(10,20),(20,9),(20,10),(7,17),(7,27),(17,7),(17,27),(27,7),(27,17)\}$

$\sum_{i=1}^{6}(x_i-\bar{x})(y_i-\bar{y}) = -226$, $\sum_{i=1}^{6}\left(x_i-\bar{x}\right)^2 = 596$, $\sum_{i=1}^{6}(y_i-\bar{y})^2 = 596$; Thus

$$\rho = \frac{\sum_{i=1}^{6}(x_i-\bar{x})^2(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{6}(x_i-\bar{x})^2\sum_{i=1}^{6}(y_i-\bar{y})^2}} = \frac{-226}{596} = -0.3791946309.$$

- In general, ρ is between -1 and 1, but in systematic sampling,

$\frac{\sigma^2}{n}(1+(n-1)\rho) \geq 0 \Rightarrow 1+(n-1)\rho \geq 0 \Rightarrow \rho \geq \frac{-1}{n-1}$. Thus, $\frac{-1}{n-1} \leq \rho \leq 1$.

- If $n = N$, then we have 1-in-1 systematic sample (it means a comprehensive survey (census)). In this case, $\bar{X} = \mu, ; Var(\bar{X}) = 0$. Thus,

$\frac{\sigma^2}{N}(1+(N-1)\rho) = 0 \Rightarrow 1+(N-1)\rho = 0 \Rightarrow \rho = \frac{-1}{N-1}$.

Based on the above, we have the following important and interesting fact:

Fact:

If a population consists of N elements with values of the variable of interest $u_1, u_2, ..., u_N$ then the correlation between the two variables (X,Y) , where the values of (X,Y) are $\{(u_i,u_j): 1 \leq i,j \leq N, i \neq j\}$ is $\rho = \frac{-1}{N-1}$.

In other words, the correlation between the $N(N-1)$ pairs is $\rho = \frac{-1}{N-1}$.

Example:

a) Based on the above fact, the correlation of all pairs of $\{3,7,11\}$ is $\frac{-1}{3-1} = -\frac{1}{2}$. The pairs of (X,Y) are $\{(3,7),(3,11),(7,3),(7,11),(11,7)(11,3)\}$; Using the formula of Pearson correlation:

$$\rho = \frac{\sum_{i=1}^{3}(x_i-\bar{x})^2(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{3}(x_i-\bar{x})^2\sum_{i=1}^{3}(y_i-\bar{y})^2}} = \frac{-32}{64} = -0.5$$

b) If $N = 100$, then correlation between the 9900 pairs is $\rho = -\frac{1}{99}$. If $N = 1000$ then the correlation between the 999000 pairs is $\rho = \frac{-1}{999}$.

3. Conclusions

Several interesting properties of the main sampling techniques (Simple, Stratified, Cluster and Systematic random sampling) are high-lighted. Some of these properties are hidden and probably rarely mentioned in classrooms. In addition to that, the Pearson correlation coefficient between all pairs of the N population values turned out to be $\rho = -1/(N-1)$. It is strongly believed that this very interesting fact never mentioned in any statistical paper or books. The main conditions for groups to be strata or clusters are discussed. Some examples about strata are taken from the Holly Quran. We believe that the topics discussed in this paper will have a real contribution to the area of educational statistics. Mentioning these topics in classrooms may motivate students to learn and to like the topics.

4. Funding and/or conflicts of interests/competing interests

The authors declare that there is no conflict of interest regarding the publication of this article. The article is not funded.

References

[1] AL-Saleh, M. F. (2005). On the confusion about ($n-1$)-divisor of the standard deviation. Journal of Probability and Statistical Science, 5, 139-144.

[2] AL-Saleh, M. F. (2007). A rich learning lesson using the Poisson distribution. Statistical Methodology, 4, 504-507. https://doi.org/10.1016/j.stamet.2007.01.005.

[3] Al-Saleh, M. F. and Yousif, A. E. (2009). Properties of the standard deviations that are rarely mentioned in classrooms. Austrian journal of statistics, 3, 193-202.

[4] Al-Saleh, M. F. (2012). Statistical Ideas that are Rarely Mentioned in Classrooms. 12th Islamic Countries Conference on Statistical Sciences, Qatar.

[5] Al-Saleh, M. F. (2015). Mathematical Ideas that are Rarely Mentioned in Classrooms. Second Statistical Conference, Jordan.

[6] Al-Saleh, M. F. and Maabreh, Arwa (2025). A Rich Learning Statistical Lesson in the Cauchy Distribution. Mutah Lil Buhuth wad-Dirasat Natural and Applied Sciences. To appear.

[7] Balakrishnan, N. and Balasubramanian, K. (2007). Revisiting Sen's inequalities on order statistics. Statistics and probability letters,78, 616-621. https://doi.org/10.1016/j.spl.2007.09.023.

[8] Casella, G. and Berger, R. (1990). Statistical Inference, Wadsworth & Brooks/Cole Advanced Book & Software, California, USAHeinbockel, J. H. (2012). Introduction to Calculus Volume II, Old Dominion University, VA, USA.

[9] Heinbockel, J. H. (2012). Introduction to Calculus Volume II, Old Dominion University, VA, USA.

[10] Heien, H. C. and Rahman, M. (2005). "Revisiting the Digits of PI and Their Randomness", International Journal of Statistical Science(IJSS), 4, 13-24.

[11] Sahoo, P. (2013). Probability and Mathematical Statistics. First Edition. Thompson, University of Lewisville, KY, USA.

[12] Schaeffer, R. L., Mendenhall, W. and Ott, R. L. (2006). Elementary Survey Sampling. Sixth edition, Duxbury Press.