Concept Paper

Topological Representation of Sample Size Calculation: The 4D Nature of Variables and 3D Nature of Surfaces

Dev Desai *

Posted Date: 9 April 2024

doi: 10.20944/preprints202403.0445.v2

Keywords: Sample Size, Topology, Variables, Z, Prevalence, Precision

Concept Paper

Topological Representation of Sample Size Calculation: The 4D Nature of Variables and 3D Nature of Surfaces

Dev Desai

Smt. NHLMMC, Ahmedabad, India; devhdesai01@gmail.com

Abstract: Sample size using prevalence has been one of the most important medical research methodologies and has been used in original articles more often than any other statistical tool available. The simple formula consists of variables interdependent and it is as important to understand how they influence each other as it is to learn to use the formula to calculate the sample size. With the choice of different amounts of precision and levels of confidence available to a medical researcher, it should be of paramount importance to understand what a slight change in choosing the parameter would mean at the end of the research. Understanding of different Z values, Levels of precision, and ranging Prevalence is required to understand the effect on the sample size and subsequently on the research. Topological representation using parameter ranges can be used to demonstrate this relationship.

Keywords: sample size; topology; variables; Z; prevalence; precision

Introduction

Sample size calculations using formulae have been the fundamental basis of research methodologies. The set precision standards and the Z-value are used with the known prevalence of the known disease is one of the simplest ways to calculate the size of the sample required in studying that disease at any given point in time [1] Prevalence is considered from the previous literature and is used in the formula. The formula of sample size using prevalence is simple yet has multiple unknown entities that need to be set by the researcher. It depends on the Z value which is determined by the level of confidence the author wants in the study and the precision of the study required [2] Using the formulae, yields a minimum number of patients required to give a result that at the level of confidence and p-value would have a corresponding probability of being by chance [3,4]

$$SS = \left(\frac{(Z^2)}{(a^2)}\right)((p)(1-p))$$

SS = Sample Size
Z = Z statistic for a level of confidence
p = Expected prevalence or proportion
a = Margin of Error (MOE) (Precision)

The Formula has multiple variables which can be set by the author and it's often difficult to comprehend the influence they carry and what changes would occur in the calculation when they are changed. It becomes important to understand the interrelationship between these variables and the effect they have [3,5].

Graphical representation of these variables is also impossible as for a 3D Graph, only 2 variables can be used and for a 3D surface plot, only 3 variables can be used. Using 4 variables would require a 4D plot or one of the variables needs to be determined. These 4 variables have a lower limit and upper limit and that can be used to determine different values and then they can be placed on the graph.

Z values are set mathematically for different levels of confidence. The usual value taken is 1.96 at a 95% level of confidence. Precision corresponds to the level of allowable false positive error. Prevalence is used from the literature and is used in the "out of 100 individuals" format in the formulae [6,7]

Table 1. For Z statistic or Z- value (z).

Level of confidence	Corresponding Z Statistic Value
50%	0.674
70%	1.036
75%	1.15
80%	1.282
85%	1.44
90%	1.645
95%	1.96
98%	2.326
99%	2.576
99.50%	2.807
99.73%	3.000
99.90%	3.291
99.99%	3.891
99.999%	4.417

For MOE (a) the value ranges from $0.0\overline{00}1\%$ to 20% of the prevalence its corresponding values are used in the formulae making the below expression represent the limits.

$0.0\overline{00}1 \leq \text{a} \leq 0.2$

For prevalence (p), the highest prevalence can be just $[0 \neq n \rightarrow 0]$ less than half of the population. Hence, the lower limit of prevalence can be explained by stating that it reaches up to 0 but is never 0. As the upper limit cannot be more than half, the below limit correctly represents the prevalence.

$$0.0\overline{00}1 \leq \text{p} \leq 0.4\overline{99}99$$

A disease prevalence higher than half of the population can be seen from the other side as less than half population is normal. Hence, the highest prevalence for anything stated is just less than half.

Firstly, the sample size can be calculated by taking Z at 1.96 (95% confidence) and MOE at 0.05 to demonstrate the most widely used status. Changing prevalence from 0.0001 to 0.49999, the following graph can be seen in 2D

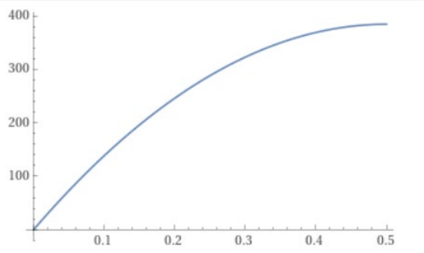

Figure 1. 2D Parabola graph of Sample Size at constant Z= 1.96 and MOE = 0.05.

It is evident here from the formula that the highest value of the sample size will not be more than 385. At any given prevalence due to the "(p)(1-p)" part in the formulae, the highest achievable value of the part is 0.25. Any other value of p higher or lower will result in a value lesser than 0.25

Using three values of Z, the lower and upper limit of Z, and the most commonly used value, Topological 3D surface plots can be created to represent the formula for changing prevalence and precision and how it changes the sample size.

The below images and link to the 3D graph are created using "**GeoGebra 3D surface plotter"**
On the Z axis – Sample Size
On the X axis - Precision ranging from $0.\overline{00}01 \leq a \leq 0.2$
On the Y axis - Prevalence ranging from $0.0\overline{00}1 \leq p \leq 0.4\overline{99}99$
The value of Z was used as a constant to create different surfaces
Z at 50% Confidence = 0.674 - Green Surface
Z at 95% Confidence = 1.96 - Blue Surface
Z at 99.999% Confidence = 4.417 - Orange Surface
Variable value from $0.647 \leq Z \leq 4.417$ - Pink Surface

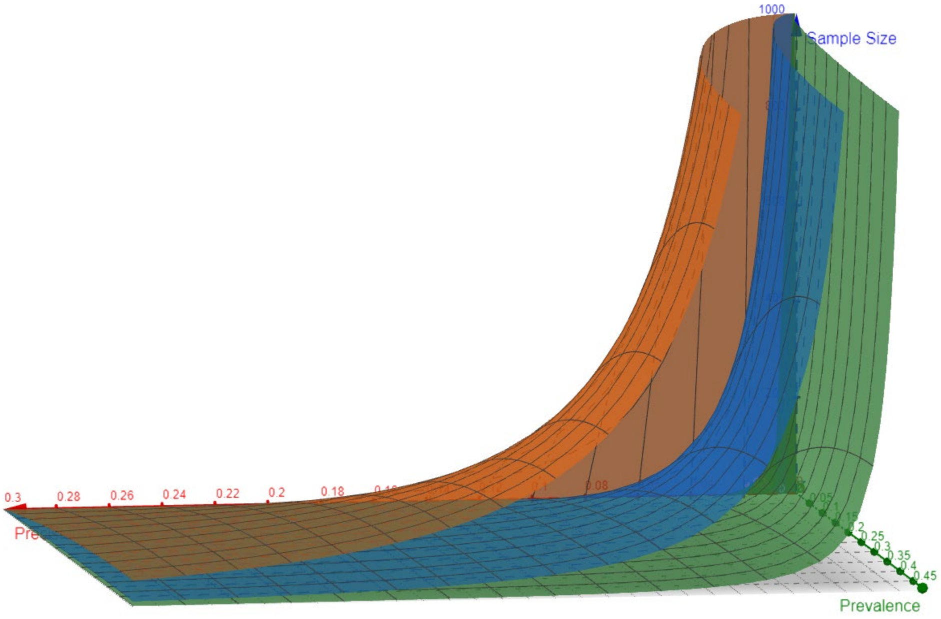

Figure 2. - XYZ view of the topological surfaces represents the curvature when precision on the X-axis moves from 0.0001 towards 0.3 and higher. The higher the value of Z, the further the surface from the XY plane suggests that a higher sample size is needed for a better level of confidence.

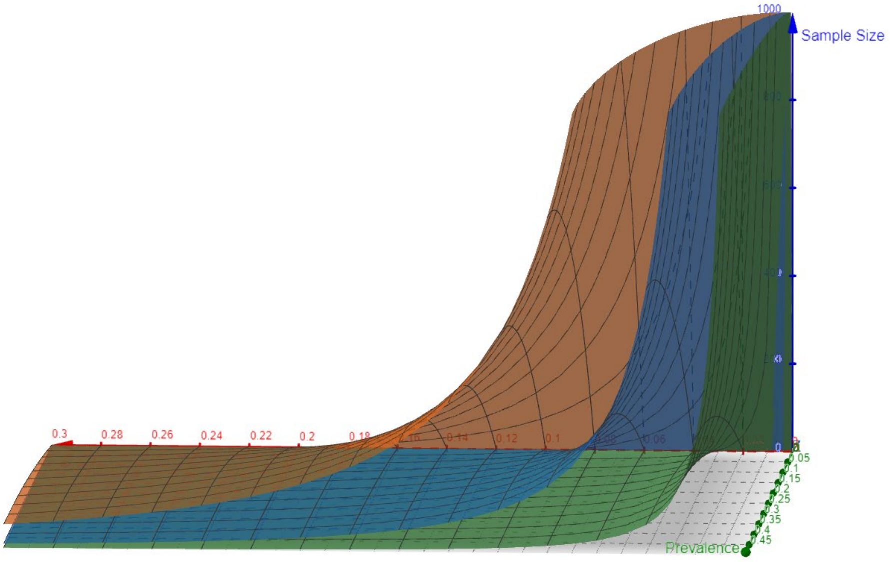

Figure 3. XYZ view of the topological surface showing the separation from the YZ plane created between Sample size and prevalence. It represents the end of the Z axis trying to show the curve is very near and adjacent to the YZ plane but it never touches that as it would mean that the precision

value is so high that the error is 0 making the result of the formula at Sample size of ∞. Hence, these surfaces will go along the YZ plane towards the Z axis but will never touch the plane.

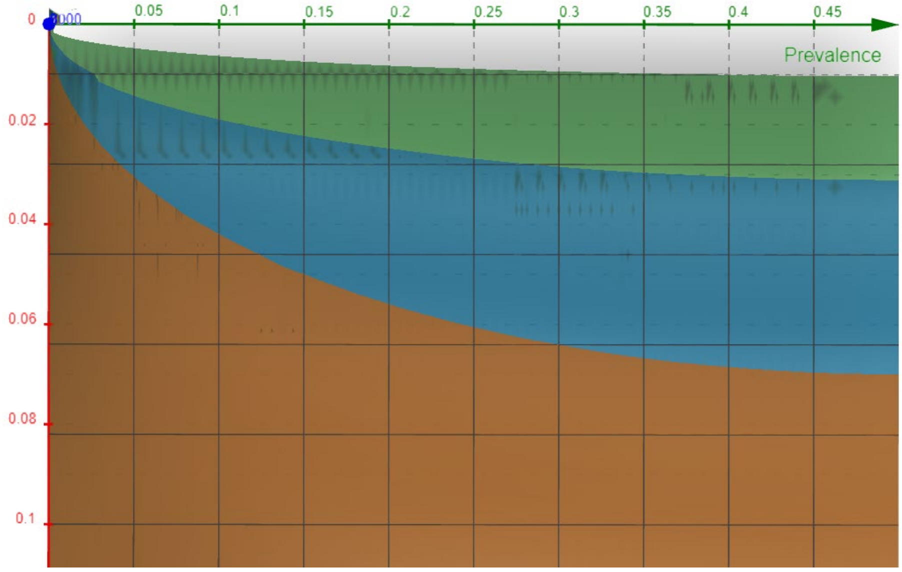

Figure 4. Contour plot of XY plane showing the top view. As the surfaces can be seen as solid areas under different curves, the relationship between precision and precision and prevalence can be seen corresponding to the respective Z value.

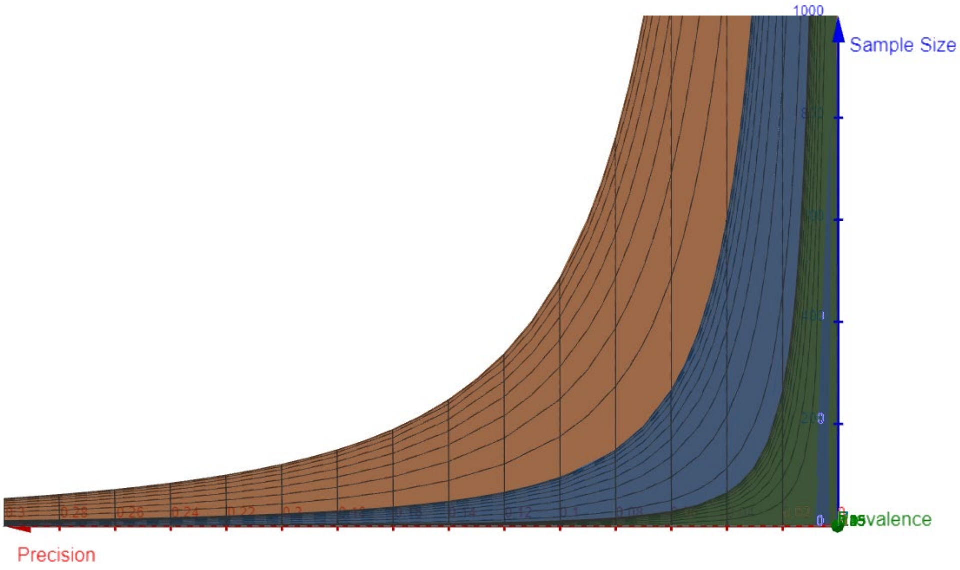

Figure 5. Contour plot of XZ plane showing the side view the curvature can be visualized in 2D here and the area under the surface can be visualized.

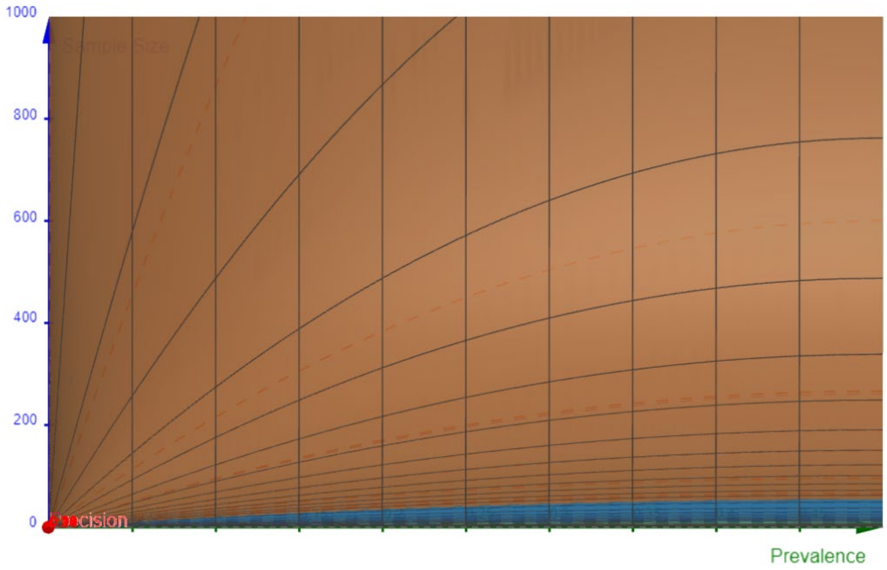

Figure 4. Contour plot of YZ plane showing the Front view. The Blue color (Z=1.96) at the bottom of the chart represents a minimal sample that will be required despite any precision value just because the Z value is very high and the prevalence is increasing. The grid lines can be visualized here showing the parabola nature of the formula evident in Figure 1.

Video 1:- Demonstrating the variable Z statistic value

The 3D plot can be accessed using this link - https://drive.google.com/drive/folders/1ZIdCHHPG8inVYP_nxuJZXLoxcRzJ80jB?usp=drive_link

These graphs represent the interrelationship between the variables such as Z value, Precision, and Prevalence and how it affects the sample size calculations.

More complex forms of sample size calculation formulae can be used similarly to understand how the parts of the formulae are related to each other and how a small change might affect the behavior and calculations and subsequently the research methodology and research process entirely.

Conclusion

Graphical representation of the sample size formulae using its variable can be done using 3D Surface plotting and limits of some variables. It gives important insights into the interrelation between these different parts of the formulae and how one affects the other. As a personnel from a medical and research background, just calculating sample size is not enough and it becomes important to understand what the leniency in choosing the parameters means.

References

1. **Mohamad Amin Pourhoseingholi, Mohsen Vahedi, Mitra Rahimzadeh.** Sample size calculation in medical studies. [Online] 2013. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4017493/.
2. **Hazra, Avijit.** Using the confidence interval confidently. [Online] 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5723800/.
3. **Lin Naing, Rusli Bin Nordin, Hanif Abdul Rahman, Yuwadi Thein Naing.** Sample size calculation for prevalence studies using Scalex and ScalaR calculators. [Online] 2022. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01694-7.
4. **Jaykaran Charan, Rimplejeet Kaur, Pankaj Bhardwaj, Kuldeep Singh, Sneha R. Ambwani.** Sample Size Calculation in Medical Research: A Primer. [Online] 2021. https://nams-india.in/anams/2021/NAMS57_2_article2.pdf.
5. **Sabyasachi Das, Koel Mitra, Mohanchandra Mandal.** Sample size calculation: Basic principles. [Online] 2016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5037946/.

6. **Hajian-Tilaki, Karimollah.** Sample size estimation in epidemiologic studies. [Online] 2011. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3895825/.
7. **Tenny, Steven and Hoffman, Mary R.** Prevalence. [Online] 2023. https://www.ncbi.nlm.nih.gov/books/NBK430867/.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.