Statistical distribution and categorization of continuous modifiable variables - A simulation approach

Afolabi Nathanael, Prof. Oyeyemi Gafar

Abstract— **This study examines the statistical distribution and categorization of continuous modifiable variables. The research focuses on recognizing the importance of statistical distribution properties needed for categorization in data analysis for decision making. By examining the predictive power, sensitivity and specificity of categorized and uncategorized factors, decisions on optimal cut-points can be made for important clinical interventions to improve health outcomes.**

Index Terms— **Categorization, Sensitivity, Predictive power, Cut-points.**

I. INTRODUCTION

Generally, improving health outcomes by targeting modifiable factors is extremely imperative as these variables constitute the elements of health intervention that can be altered to get the desired result. Modifiable factors, also known as modifiable risk factors, have greatly influenced the developments of health interventions, practices and policies that seek to better health outcomes in health research and epidemiology [1]. Depending on the type of health needs to be improved, modifiable variables include various components such as lifestyle factors (dietary intake, physical activity, smoking and alcohol habit, sleep quality/duration, safe sexual practices), medication adherence factors (taking prescribed medications, vaccinations schedule or the number of vaccinations received), psychosocial factors (stress management, social support received, mental health management, weight management), hygiene and sanitation factors (practices like handwashing and proper basic sanitation), environmental factors (exposure to toxins such as air and or smoke pollution), and education and health literacy factors (number of years of schooling or a specific educational intervention). Based on the statistical properties of variables, certain types of techniques are usually employed in analysis.

This happens with the exploration of the identified variables to determine the appropriate statistical approach to use. Normality or the concept of normal distribution may not always hold sway or be ideal in the application and utilization of modifiable factors, hence this poses a challenge and/or an opportunity for researchers. Whether modifiable variables are normally distributed depends on the specific risk factor and the group being studied. Accurate modeling, risk assessment, and intervention planning require a thorough understanding of the distribution of the modifiable components as this will provide detailed and subtle insights into the intricate relationship between individual attributes and health outcomes [1], [2].

This research paper delves into the statistical distribution and categorization of continuous modifiable variables [3]. This is needed to enhance how these variables influence disease risk, treatment response and overall well-being. Recognizing the distribution of risk factors is crucial for analysis and modeling since it can affect the selection of tests and methods used for data analysis. In addition, making informed decisions that positively impact health outcomes requires proper identification of important thresholds, optimal ranges, and potential nonlinear relationships. In essence although some modifiable risk factors may display distributions occasionally, they tend to adhere to distribution patterns more frequently. Therefore, it's essential to assess their distribution, within the study population context.

The main goal of this research is to explore how continuous modifiable variables are categorized and the distribution patterns using statistical analysis methods. By using techniques and conducting simulations, this study aims to improve knowledge of the distribution characteristics of these variables and find the best ways to categorize them effectively. The objective is to offer insights that can guide decision making in fields such as health, epidemiology and social sciences where accurately categorizing continuous modifiable variables is crucial for developing successful interventions and policies.

II. METHODOLOGY

Three continuous statistical distributions were chosen and data simulated, and these distributions include the normal (standard), gamma (most occurring) and beta (least occurring) distributions with their properties examined [4], [5]. The simulated data were thereafter categorized to investigate the optimal number of cut-points in the absence of expert opinion.

III. NORMAL DISTRIBUTION

The normal distribution is mathematically given as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ for all values of x and μ; while $\alpha > 0$ (1)

Mean = μ, Standard deviation = σ

Mean ($\boldsymbol{\mu}$) and standard deviation ($\boldsymbol{\sigma}$) are the distributional parameters.

Standard deviation > 0 and can be any positive value and Mean can be any value.

IV. Gamma Distribution

The gamma distribution mathematically given as:

$$f(x) = \frac{\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta} \text{ with any value of } \alpha > 0 \text{ and } \beta > 0 \quad (2)$$

Mean = αβ, Standard deviation = $\sqrt{(\alpha\beta)^2}$

Distributional parameters are (i) Shape parameter alpha (**α**) and (ii) scale parameter beta (**β**) while **Γ** is the gamma function.

Scale beta > 0 and can be any positive value,

Shape alpha ≥ 0.05 and any positive value while location can be any value.

V. Beta Distribution

The beta distribution mathematically given as:

$$f(x) = \frac{(x)^{(\alpha-1)}(1-x)^{(\beta-1)}}{\left[\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right]} \; for\; \alpha > 0;\; \beta > 0; x > 0 \quad (3)$$

$$\text{Mean} = \frac{\alpha}{(\alpha+\beta)}, \quad \text{Standard deviation} = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}}$$

Alpha (**α**) and beta (**β**) are the two distributional shape parameters, and **Γ** is the gamma function.

VI. Models and Parameter Settings for the Simulation Studies

Data was generated through simulation for normal distribution, given by N (0, 1), beta distribution (scale=7, shape=7), and gamma distribution (scale=7.5, shape=0.05). Sample sizes assumed for simulation are 100, 500 and 1000. Additionally, data was simulated for two scenarios - one continuous variable only and two continuous variables. Simulation was implemented using the R package called CatPredi [6], [7]. Derived cut-points scenarios for the studies are 1-cutpoint, 2-cutpoints and 3-cutpoints (k) corresponding to two, three and four categories respectively. The area under the receiver operating characteristics curve (AUC) performed within the logistic regression framework was used to classify the continuous predictor variables as well as fitted to create a confusion matrix. Diagnostics assessment obtained includes Akaike information criteria (AIC), sensitivity, specificity, and percentage correctly classified (accuracy or predictive power).

VII. Result

Table I shows the cut off points and their corresponding area under the curve (AUC) values, for three types of distributions (Normal, Beta and Gamma) across sample sizes (N=100, N=500 N=1000) for a single continuous modifiable variable. As observed in Table I, as the sample size increases, AUC tends to also increase and become stable across each of the distribution types. Comparing the AUC values in gamma and beta distributions to the normal distribution for the different cut-points, the normal distribution AUCs values were observed to be higher.

Table I: Table of statistical simulation (One continuous variable) for normal, gamma and beta **distributions at** different sample sizes

Distribution		**Optimal values**					
		N=100		**N=500**		**N=1000**	
	k	Cutpoint(s)	AUC	Cutpoint(s)	AUC	Cutpoint(s)	AUC
Normal (u=0 & δ=1)	1	-0.3451	0.6875	-0.0641	0.7286	-0.2304	0.6881
	2	-1.162	0.7112	-1.0442	0.7278	-0.9119	0.7261
		0.435		0.0654		0.3327	
	3	-0.4382	0.7456	-1.0450	0.7436	-1.0057	0.7421
		-0.1396		-0.8900		-0.2563	
		0.4352		0.0644		0.4238	
Beta (α=7, β=7)	1	0.6230	0.5852	0.5611	0.5677	0.5117	0.5528
	2	0.4011	0.6326	0.3974	0.5799	0.3353	0.5723
		0.4491		0.7028		0.5811	
	3	0.5304	0.6816	0.3975	0.6022	0.3358	0.5900
		0.5415		0.6243		0.4436	
		0.5953		0.7028		0.5811	
Gamma (α=7.5, β=0.05)	1	0.2347	0.6247	0.4318	0.5258	0.4031	0.5204
	2	0.3919	0.6761	0.2837	0.5739	0.3156	0.5938
		0.4863		0.4800		0.3734	
	3	0.2261	0.7133	0.2407	0.5998	0.2657	0.6086
		0.3913		0.2842		0.3156	
		0.4906		0.3256		0.3737	

Table II shows the cut off points and their corresponding area under the curve (AUC) values, for three distributions (Normal, Beta and Gamma) across sample sizes (N=100, N=500 N=1000) for two continuous modifiable variables. The AUCs in the gamma distribution were observed to be higher than the AUCs for the normal distribution, although the AUCs for beta distribution were lowest. Additionally, and like the single continuous modifiable variable, as the sample size increases, AUC tends to also increase and become stable across each of the distribution types.

Table II: Table of statistical simulation (two continuous variables) for normal, gamma and beta distributions at different sample sizes

Distribution	**Variables**		**Optimal values**					
			N=100		**N=500**		**N=1000**	
		k	Cutpoint(s)	AUC(s)	Cutpoint(s)	AUC(s)	Cutpoint(s)	AUC(s)
Normal	Xn21	1	-0.0355	0.7092	0.1177	0.6575	0.1449	0.6664
		2	-0.0379	0.7439	-0.6270	0.6962	-0.6882	0.7037
			0.7853		0.1190		0.2335	
		3	-0.3513	0.7695	-0.6269	0.7065	-0.6886	0.7162
			-0.0382		-0.5815		0.1454	
			0.8158		0.1198		0.9840	
	Xn22	1	-0.0022	0.6832	0.1200	0.7087	-0.0229	0.6666
		2	0.0065	0.7233	0.1227	0.7355	-0.02816	0.6969
			0.2302		0.9885		1.1033	
		3	-0.7554	0.7589	-0.2877	0.7452	-0.6340	0.7139
			-0.5955		0.1274		-0.0244	
			-0.0167		1.0076		1.1033	
Beta	Xb21	1	0.3930	0.5852	0.3486	0.5360	0.5369	0.5463
		2	0.3972	0.6442	0.3448	0.5772	0.4096	0.5565
			0.5638		0.5396		0.5410	
		3	0.3968	0.6610	0.3462	0.5933	0.4097	0.5716
			0.5658		0.5398		0.4760	
			0.6017		0.5507		0.5330	
	Xb22	1	0.5867	0.6741	0.4301	0.5427	0.5267	0.5206
		2	0.5001	0.7244	0.4299	0.5604	0.5909	0.5398
			0.5881		0.5521		0.6295	
		3	0.4887	0.7710	0.3145	0.5853	0.5908	0.5553
			0.5029		0.5520		0.6394	
			0.5872		0.6950		0.7259	
Gamma	Xg21	1	6.855	0.8276	6.428	0.7088	7.356	0.7037
		2	5.385	0.8573	5.462	0.7587	5.611	0.7412
			7.493		9.057		7.999	
		3	5.332	0.8677	4.851	0.7770	5.496	0.7588
			6.851		6.307		7.363	
			11.920		9.064		9.503	
	Xg22	1	7.724	0.7768	7.252	0.7550	7.213	0.7388
		2	6.094	0.8637	6.775	0.8077	6.532	0.7836
			8.851		9.103		9.411	
		3	6.093	0.8741	4.623	0.8211	5.734	0.8026
			8.209		7.254		7.216	
			9.857		9.079		9.468	

In Table III, the diagnostic results for one and two uncategorized and categorized variables on simulated data for the normal distribution with varying number of cut-points is presented. Using a logistic regression fit, sensitivity, specificity, and percentage correctly classified (accuracy or predictive power) of the variables were estimated. In both single and two continuous variables, accuracy percent for the varying cut-points were observed to be comparable to the uncategorized variable (0) across the different sample sizes. Sensitivity values were quite higher and doubling the specificity values across board irrespective of model or sample sizes and were likewise comparable between uncategorized and categorized models. Similarly, conclusions on significance were similar between categorized and uncategorized models across board.

Distribution/ Model	No of cutpoints(k)	AIC	Sensitivity	Specificity	Percent correctly classified	P-value
One continuous variable						
100	0	101.32	94.7%	16.7%	76.0%	-
	1	102.65	86.8%	45.8%	77.0%	0.868
	2	100.38	86.8%	45.8%	77.0%	0.868
	3	96.43	90.0%	45.8%	78.7%	0.648
500	0	445.23	96.4%	29.5%	81.4%	-
	1	453.09	90.2%	43.8%	79.8%	0.522
	2	452.40	89.7%	44.6%	79.6%	0.473
	3	437.28	97.4%	26.8%	81.6%	0.935
1000	0	898.39	95.6%	21.6%	79.5%	-
	1	942.79	100%	0.0%	78.2%	0.477
	2	904.98	92.6%	32.1%	79.4%	0.956
	3	881.39	97.3%	22.5%	81.0%	0.400
Two continuous variables						
100	0	117.28	84.4%	38.9%	68.0%	-
	1	105.17	90.6%	50.0%	76.0%	0.208
	2	92.97	84.1%	58.3%	72.5%	0.486
	3	87.36	86.4%	56.3%	73.7%	0.375
500	0	511.07	87.2%	55.6%	75.8%	-
	1	525.90	87.5%	57.2%	76.6%	0.766
	2	505.52	87.5%	57.2%	76.6%	0.766
	3	501.29	87.5%	57.2%	76.4%	0.824
1000	0	511.07	87.2%	55.6%	75.8%	-
	1	525.90	87.5%	57.2%	76.6%	0.674
	2	505.52	87.5%	57.2%	76.6%	0.674
	3	501.29	87.5%	57.2%	76.4%	0.753

Table III: Diagnostic results for uncategorized and categorized variables on simulated data for Normal distribution with varying number of cutpoints (k)

Table IV presents the diagnostic results for one and two uncategorized and categorized variables on simulated data for the beta distribution with varying number of cutpoints. Using the logistic regression fit, sensitivity, specificity, and percentage correctly classified (accuracy or predictive power) of the variables were estimated. Predictive power was observed to be similar between uncategorized and categorized continuous for both one and two variables as well as across the sample sizes. Sensitivity was consistently high across all categorized and uncategorized continuous variables while specificity was almost non-existent. The conclusion on significance were also observed to be similar between categorized and uncategorized continuous variables and across sample sizes.

Table IV: Diagnostic results for continuous and categorized models on simulated data for Beta distribution with varying number of cut-points (k)

Distribution/ Model	No of cut-points(k)	AIC	Sensitivity	Specificity	Percent correctly classified	P-value
One continuous variable						
100	0	95.85	100%	0%	82.0%	-
	1	95.22	100%	0%	82.0%	1.000
	2	92.46	100%	0%	82.0%	1.000
	3	87.96	98.8%	22.2%	85.0%	0.568
500	0	396.66	100%	0%	85.4%	-
	1	391.66	100%	0%	85.4%	1.000
	2	384.61	100%	0%	85.4%	1.000
	3	384.49	100%	0%	85.4%	1.000
1000	0	744.88	100%	0%	87.0%	-
	1	742.79	100%	0%	87.0%	1.000
	2	739.20	100%	0%	87.0%	1.000
	3	728.69	100%	0%	87.0%	1.000
Two continuous variables						
100	0	117.62	95.8%	13.8%	72.0%	-
	1	108.78	88.7%	44.8%	76.0%	0.519
	2	108.48	88.7%	44.8%	76.0%	0.519
	3	100.04	88.7%	50.0%	78.4%	0.295
500	0	564.98	100%	0%	75.0%	-
	1	561.57	100%	0%	75.0%	1.000
	2	559.93	100%	0%	75.0%	1.000
	3	551.59	96.8%	8.0%	74.6%	0.884
1000	0	564.98	100%	0%	75.0%	-
	1	561.57	100%	0%	75.0%	1.000
	2	559.93	100%	0%	75.0%	1.000
	3	551.59	96.8%	8.0%	74.6%	0.837

Table V presents the diagnostic results for continuous one and two uncategorized and categorized variables on simulated data for the gamma distribution with varying number of cut-points. Using the logistic regression fit, sensitivity, specificity, and percentage correctly classified (accuracy or predictive power) of the variables were estimated. Sensitivity was mostly 100% across board (uncategorized versus categorized) reflecting the degree of skewness to the right an in tandem with the characteristics of the gamma distribution on the one continuous variable. Specificity on the other hand, was nonexistent (0%) for all categorized and uncategorized one continuous variable. For the two continuous variables, both diagnostics were similarly high and comparable between categorized and uncategorized models. Across board, predictive power and conclusion on significance were similar between categorized and uncategorized models.

Table V: Diagnostic results for continuous and categorized models on simulated data for Gamma distribution with varying number of cutpoints (k)

Distribution/ Model	No of cutpoints(k)	AIC	Sensitivity	Specificity	Percent correctly classified	P-value
One continuous variable						
100	0	86.55	100%	0%	85.0%	-
	1	87.18	100%	0%	85.0%	1.000
	2	89.16	100%	0%	85.0%	1.000
	3	85.89	100%	0%	83.2%	0.728
500	0	365.26	100%	0%	87.8%	-
	1	362.03	100%	0%	87.8%	1.000
	2	359.46	100%	0%	87.8%	1.000
	3	359.26	100%	0%	87.8%	1.000
1000	0	816.43	100%	0%	85.1%	-
	1	811.44	100%	0%	85.1%	1.000
	2	809.52	100%	0%	85.1%	1.000
	3	808.92	100%	0%	85.1%	1.000
Two continuous variables						
100	0	50.31	88.7%	91.5%	90.0%	-
	1*	-	-	-	-	
	2	52.99	84.9%	95.7%	90.0%	1.000
	3	43.23	93.2%	85.7%	90.3%	0.943
500	0	297.30	89.7%	83.2%	87.2%	-
	1	418.92	97.4%	47.9%	78.6%	0.000#
	2	323.39	95.8%	68.4%	85.5%	0.434
	3	292.79	91.6%	85.3%	89.2%	0.327
1000	0	297.30	89.7%	83.2%	87.2%	-
	1	418.92	97.4%	47.9%	78.6%	0.000#
	2	323.39	95.8%	68.4%	85.4%	0.242
	3	292.79	91.6%	85.3%	89.2%	0.166

DISCUSSION

Review of current literature revealed the need to foster interaction between two or more modifiable factors [8], [9]. Effective interaction between factors is better assessed when variables are categorized [10]. Despite the argument against categorization of continuous variables due to loss of information and reduced statistical powers, the importance and use of categories in health research and disciplines is undeniable and continues to abound. Continued research on efficient way of categorizing modifiable continuous variables is essential to break through in meeting health care need in modern times.

REFERENCES

[1]N. A. Alwan *et al.*, "Risk factors for ill health: How do we specify what is 'modifiable'?," *PLOS Global Public Health*, vol. 4, no. 3, p. e0002887, Mar. 2024, doi: 10.1371/journal.pgph.0002887.

[2]A. A. Daly, R. Rolph, R. I. Cutress, and E. R. Copson, "A review of modifiable risk factors in young women for the prevention of breast cancer," *Breast Cancer: Targets and Therapy*, vol. 13. Dove Medical Press Ltd, pp. 241–257, 2021. doi: 10.2147/BCTT.S268401.

[3]I. Barrio, I. Arostegui, and J. M. Quintana, "Use of generalised additive models to categorise continuous variables in clinical prediction," *BMC Med Res Methodol*, vol. 13, no. 1, pp. 1–13, 2013, doi: 10.1186/1471-2288-13-83.

[4]R. Bono, M. J. Blanca, J. Arnau, and J. Gómez-Benito, "Non-normal distributions commonly used in health, education, and social sciences: A systematic review," *Frontiers in Psychology*, vol. 8, no. SEP. Frontiers Media S.A., Sep. 14, 2017. doi: 10.3389/fpsyg.2017.01602.

[5]P. Guzik and B. Więckowska, "Data distribution analysis – a preliminary approach to quantitative data in biomedical research," *J Med Sci*, Jun. 2023, doi: 10.20883/medical.e869.

[6]I. Barrio, M. X. Rodríguez-Álvarez, and I. Arostegui, "Categorisation of continuous variables in a logistic regression model using the R package CatPredi," vol. 1, p. e004, 2015, doi: 10.3390/mol2net-1-e004.

[7]I. Barrio, I. Arostegui, M. X. Rodríguez-Álvarez, and J. M. Quintana, "A new approach to categorising continuous variables in prediction models: Proposal and validation," *Stat Methods Med Res*, vol. 26, no. 6, pp. 2586–2602, 2017, doi: 10.1177/0962280215601873.

[8]A. A. Daly, R. Rolph, R. I. Cutress, and E. R. Copson, "A review of modifiable risk factors in young women for the prevention of breast cancer," *Breast Cancer: Targets and Therapy*, vol. 13. Dove Medical Press Ltd, pp. 241–257, 2021. doi: 10.2147/BCTT.S268401.

[9]L. Abudaqa *et al.*, "Analysis of Potentially Modifiable and Unmodifiable Risk Factors of Myocardial Infraction: Systematic Review," *International Journal of Public Health Excellence (IJPHE)*, vol. 2, no. 2, Apr. 2023, doi: 10.55299/ijphe.v2i2.382.

[10] Q. A. Khoiry, S. D. Alfian, and R. Abdulah, "Modifiable and Non-modifiable Factors Associated with Low Awareness of Hypertension Treatment in Indonesia: A Cross-Sectional Population-Based National Survey," *Glob Heart*, vol. 17, no. 1, 2022, doi: 10.5334/gh.1143