ISSN impreso: 0258-5936
ISSN digital: 1819-4087
Ministerio de Educación Superior. Cuba
Instituto Nacional de Ciencias Agrícolas
http://ediciones.inca.edu.cu

ALTERNATIVES OF DATA ANALYSIS WITH BINOMIAL DISTRIBUTION IN RANDOM BLOCK DESIGN

Alternativas de análisis de datos con distribución binomial en diseño de bloques al azar

Edison Ramiro Vásquez[1]✉, Alberto Caballero Núñez[2] and Magaly Herrera Villafranca[3]

ABSTRACT. The objective of the research was to evaluate the statistical techniques ANOVA, Proportions and Friedman as alternatives to analyze data with binomial distribution in random block design. Using the Monte Carlo method, 100 experiments were simulated with 3, 5 and 9 treatments (t); 4 and 8 replicates (r); with 5, 10 and 30 observations per experimental unit (n) and probability of success of the event (p) of 0,10; 0,20; ... 0,90. The alternatives of analysis: Comparison of Proportions and nonparametric procedure of Friedman, as for the indicators, they do not surpass those obtained in the classic ANOVA of the binomial data. It should be mentioned that in recent years few contributions have been made related to this type of research.

Key words: simulation, Monte Carlo, ANOVA assumptions, Friedman, statistics, linear models

RESUMEN. La investigación tuvo como objetivo valorar las técnicas estadísticas ANAVA, Proporciones y Friedman como alternativas para analizar datos con distribución binomial en diseño de bloques al azar. Mediante el método de Monte Carlo, se simularon 100 experimentos con tres, cinco y nueve tratamientos (t); cuatro y ocho réplicas (r); con 5, 10 y 30 observaciones por unidad experimental (n) y probabilidad de éxito del evento (p) de 0,10; 0,20; … 0,90. Las alternativas de análisis: Comparación de Proporciones y procedimiento no paramétrico de Friedman, en cuanto a los indicadores, no superan a los obtenidos en el ANAVA clásico del dato binomial. Es preciso mencionar, que en los últimos años se han realizado pocos aportes relacionados con este tipo de investigación.

Palabras clave: simulación, Monte Carlo, supuestos ANAVA, Friedman, estadística, modelos lineales

INTRODUCTION

The joint work between the statistician and the researcher (1) is essential when defining a statistical model, reflecting as much as possible, what is wanted to be evidenced through experimentation. In these considerations, one of the most widespread models is the Analysis of Variance, which when used efficiently, becomes a powerful tool for analysis. However, this technique requires compliance with certain requirements of the random error terms of the linear model, as independent errors, normally distributed and with homogeneous variances for all observations, conditions that are often not met (2-5).

In research practice, the presence of variables that, in some way, do not satisfy the requirements that the ANAVA demands (6,7) is frequent; such is the case, of variables of counts, which due to their discrete nature can move away from normality. In this sense, some authors point out (8-11) that given the "robustness" of the F test in this analysis procedure, its failure does not have serious consequences in the analysis; which is practically irrelevant in relation to the probability of committing a type I error (6); then, it does not deviate from the α value determined by the experimenter. However, the "robustness" of the test can be affected when this breach is severe, since the probability of exceeding the nominal value of the test increases (12,13).

[1] Universidad Nacional de Loja, Ecuador
[2] Universidad Técnica de Manabí, Ecuador
[3] Instituto de Ciencia Animal, Cuba
✉ edison.ramiro.vasquez@gmail.com

Given their nature and frequent existence in many branches of science, are important variables of counts that come from dichotomous variables or binomial distribution, which establishes a close relationship of dependence between variance and average treatments; aspect that may be present in other types of variables (14). Therefore, it is assumed that if there are differences between the means in each variant that are being tested, differences between their respective variances are possible and, therefore, the non-fulfillment of this assumption.

Indicators such as the percentage in which the null hypothesis is rejected, the minimum difference that can be detected between treatment means, observed power of the ANAVA, number of rejection of equality of treatment means (1,15); they can receive the unfavorable impact when the assumptions are not met; so it is important to identify, take into account and know their degree of involvement.

In this virtue, in the present article the statistical techniques ANAVA, Comparison of Proportions and nonparametric test of Friedman are evaluated as alternatives to analyze data with Binomial distribution in random block design.

MATERIALS AND METHODS

The Monte Carlo Simulation process (16-20) was used to generate populations of random variables with Binomial distribution, with homogeneous and heterogeneous variances, according to Levene's test $p < 0.05$ for 5, 10 and 30 observations per experimental unit (n) and probability of event success of 0.10, 0.20, ..., 0.90 (p). Experiments were designed in randomized blocks design with three, five and nine treatments (t); four and eight replicas (r). The combination of means of the treatments was defined in such a way that the differences between these means were detectable by the Minimum Significant Difference test (MDS) at a significance level of 0.05 (Table 1); for each combination, treatment-replication-observations per experimental unit, 100 experiments were generated.

The data with Binomial distribution with heterogeneous and homogeneous variances were processed with the statistical techniques ANAVA, Comparison of Proportions and the non-parametric Friedman test.

The Proportion Comparison test was used to compare the difference between the percentage of experiments in which the H0 is rejected with the ANAVA, the Comparison of Proportions and Friedman for experiments with homogeneous and heterogeneous treatment variance.

Table 1. Structure of means and variances of treatments for the different analysis variants

Treatments	p	n=5		n=10		n=30	
		μ	σ^2	μ	σ^2	μ	σ^2
1	0,10	0,5	0,5	1,0	0,9	3,0	2,7
2	0,30	1,5	1,1	3,0	2,1	9,0	6,3
3	0,50	2,5	1,3	5,0	2,5	15,0	7,5
1	0,10	0,5	0,5	1,0	0,9	3,0	2,7
2	0,20	1,0	0,8	2,0	1,6	6,0	4,8
3	0,30	1,5	1,1	3,0	2,1	9,0	6,3
4	0,40	2,0	1,2	4,0	2,4	12,0	7,2
5	0,50	2,5	1,3	5,0	2,5	15,0	7,5
1	0,10	0,5	0,5	1,0	0,9	3,0	2,7
2	0,20	1,0	0,8	2,0	1,6	6,0	4,8
3	0,30	1,5	1,1	3,0	2,1	9,0	6,3
4	0,40	2,0	1,2	4,0	2,4	12,0	7,2
5	0,50	2,5	1,3	5,0	2,5	15,0	7,5
6	0,60	3,0	1,2	6,0	2,4	18,0	7,2
7	0,70	3,5	1,1	7,0	2,1	21,0	6,3
8	0,80	4,0	0,8	8,0	1,6	24,0	4,8
9	0,90	4,5	0,5	9,0	0,9	27,0	2,7

RESULTS AND DISCUSSION

The behavior of statistical indicators is discussed, which allows to evaluate the quality of the analysis procedures that are related to the theoretical assumptions of the ANAVA rejection of the null hypothesis and number of differences detected.

Rejection of the null hypothesis

In Table 2, it is observed that the percentage of rejection declared significant, turned out to be superior with the procedure of Comparison of Proportions that obtained by the ANAVA and Friedman, for three and five treatments and number of observations per small experimental unit (5 and 10), but not when the number of experimental units is large (30); which may be associated with an approximation of the variable to normality and a closer approach to this assumption that requires this analysis technique.

Table 2. Rejection percentage of H_0 in the ANOVA, Comparison of Proportions and Friedman

Technique: r	n	ANAVA (%)	Proportion (%)	Friedman (%)	Sign. α=0,05	ES_x
Heterogeneous variances					3 treatments	
4	5	52 a	75 b	37 c	*	5
	10	89 a	99 b	73 c	*	3
	30	100	100	97	ns	1
8	5	93 a	97 a	87 b	*	3
	10	100	100	99	ns	1
	30	52	75	37	ns	0
Homogeneous variances					3 treatments	
4	5	52 a	83 b	36 c	*	5
	10	83 a	100 b	66 c	*	4
	30	98	100	98	ns	1
8	5	92 a	99 a	83 b	*	3
	10	100	100	99	ns	1
	30	100	100	100	ns	0
Heterogeneous variances					5 treatments	
4	5	52 a	75 b	37 c	*	5
	10	89 a	99 b	73 c	*	3
	30	100	100	97	ns	1
8	5	93 a	97 a	87 b	*	3
	10	100	100	99	ns	1
	30	100	100	100	ns	0
Homogeneous variance					5 treartments	
4	5	54 a	67 a	37 b	*	5
	10	88 a	94 a	79 b	*	3
	30	100	100	100	ns	0
8	5	92 a	97 a	87 b	*	3
	10	100	100	100	ns	0
	30	100	100	100	ns	0
Heterogeneous and homogeneous variances					9 treatments	
4	5	100	100	99	ns	1
	10	100	100	100	ns	0
	30	100	100	100	ns	0
8	5	100	100	100	ns	0
	10	100	100	100	ns	0
	30	100	100	100	ns	0

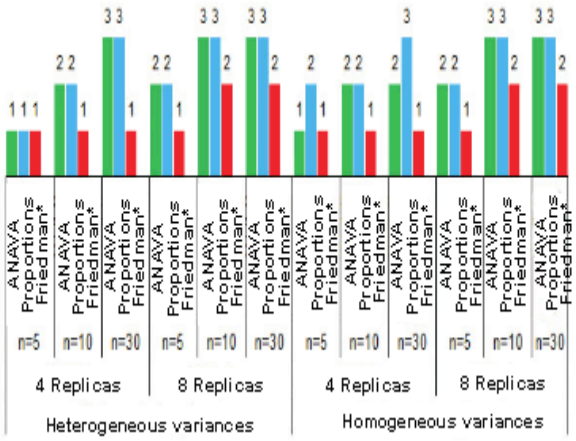
ns: significance level greater than 0.05
*: significance level less than 0.05
a, b: comparisons are made only with the ANAVA procedure

Another aspect that cannot be underestimated in the previous result is the fact that the average value of the probability of success of the event of these simulated experiments with three and five treatments is 0.30; and, the Comparison of Proportions procedure is based on the Chi-square distribution, which is more precise as the parameter p of the Binomial distribution moves away from 0.50, at which point the variance becomes maximum; this is explained for nine treatments, where the average values of the p parameter of these experiments is 0.50. The results of the rejection indicator of the H0 hypothesis are equalized in the three analysis procedures.

Friedman's nonparametric procedure showed a low behavior with respect to the other procedures and, even more emphasized, when the number of observations per experimental unit and number of replicas are small (n = 5 and 10; r = 4), these results corroborate what was raised by other researchers (5,21), when they argue that the parametric procedure is always more effective than its nonparametric counterpart.

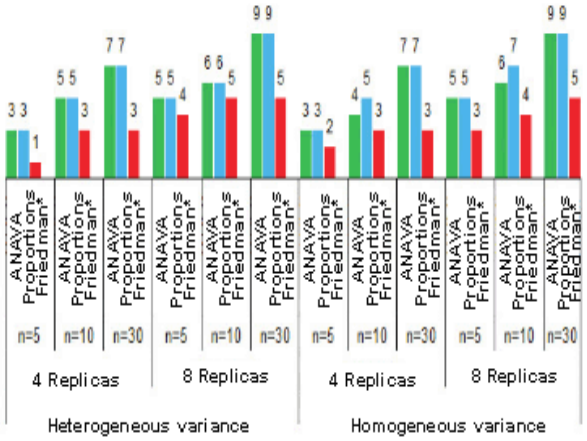
Number of differences detected

In Figures 1, 2 and 3, through the three analysis procedures and in all the analyzed variants, a significant increase was observed in the number of differences detected, as the number of observations per experimental unit increases and the number of observations increases number of replicas. This aspect is more evident for five and nine treatments, given that the number of possible comparisons are 10 and 36, respectively.



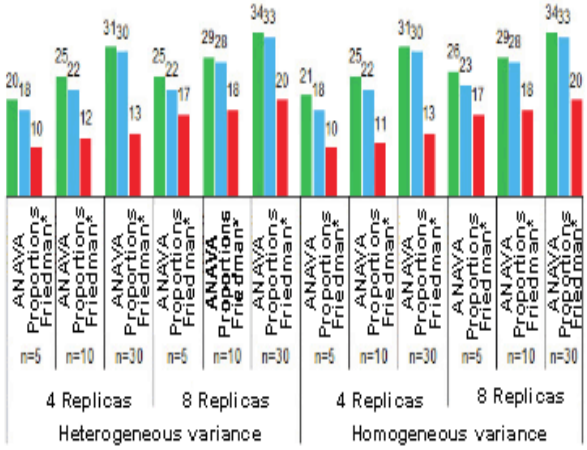
* The difference between treatments was made through their medians

Figure 1. Number of differences between treatment means, detected by the MDS test (α=0.05) for three treatments



* The difference between treatments was made through their medians

Figure 2. Number of differences between treatment means, detected by the MDS test (α=0.05) for five treatments



* The difference between treatments was made through their medians

Figure 3. Number of differences between treatment means, detected by the MDS test (α=0.05) for nine treatments

In presence or absence of homogeneity of variance of the Binomial variable, through all the combinations of number of observations per experimental unit and number of replicas, with the ANAVA and Comparison of Proportions procedures, a greater number of differences between treatments were found than with the Friedman procedure, which is associated with the fact that in this analysis procedure, the observation in it is replaced by the range that this observation occupies through the set of treatments, which leads to a loss of the essence of the amount or magnitude of the data, very important for this type of variable.

CONCLUSIONS

In general, there were no advantages of the two analysis alternatives: Comparison of proportions and non-parametric Friedman procedure, in terms of indicators that reflect the effectiveness of the ANAVA, which expresses that they showed no advantages with respect to the classic ANAVA of the data binomial, which seems to be a reasonable option for this type of data.

BIBLIOGRAPHY

1. Vásquez E, Caballero A. Cuando falla el supuesto de homocedasticidad en variables con distribución binomial. Cultivos Tropicales. 2011;32(3):46–51.
2. Di Rienzo J, Casanoves F, González L, Tablada E, Díaz M. Estadística para las ciencias agropecuarias. 7maed. Córdoba, AR: Edit. Brujas; 2009. 372 p.
3. Herrera M, Bustillos C, Sarduy L, García Y, Martínez C. Diferentes métodos estadísticos para el análisis de variables discretas. Una aplicación en las ciencias agrícolas y técnicas. Revista Ciencias Técnicas Agropecuarias. 2012;21(1):58–62.
4. Wiedenhofer S H. Pruebas no paramétricas para las ciencias agropecuarias : muestras pequeñas. 2da. Maracay , Venezuela; 2013. 261 p.
5. Bustillo C, Herrera M, Vázquez Y, Bueno A. Contribución de la estadística al análisis de variables categóricas: aplicación del análisis de regresión categórica en las ciencias agropecuarias. Revista Ciencias Técnicas Agropecuarias. 2014;23(1):68–73.
6. Sokal R, Rohlf F. Biometry: the principles and practice of statistics in biological research. 4th ed. Vol. 133. 2012. 880 p. doi:10.2307/2343822
7. Pedrosa I, Juarros J, Robles A, Basteiro J, García-Cueto E. Pruebas de bondad de ajuste en distribuciones simétricas, ¿qué estadístico utilizar? Universitas Psychologica. 2015;14(1):245–54. doi:10.11144/Javeriana.upsy14-1.pbad
8. Wetherill G. Intermediate statistical methods [Internet]. Springer Science & Business Media; 2012. 406 p. Available from: https://www.google.com/books?hl=es&lr=&id=dcLoCAAAQBAJ&oi=fnd&pg=PR13&dq=intermediate+statistical+methods&ots=kO6fC_RyPn&sig=5tUWuX-WSErFvaNQGnAbilDkoxU
9. Schmider E, Ziegler M, Danay E, Beyer L, Bühner M. Is It Really Robust? Reinvestigating the Robustness of ANOVA Against Violations of the Normal Distribution Assumption. Methodology. 2010;6(4):147–51. doi:10.1027/1614-2241/a000016
10. Ostertagová E, Ostertag O. Methodology and Application of Oneway ANOVA. American Journal of Mechanical Engineering, American Journal of Mechanical Engineering. 2013;1(7):256–61. doi:10.12691/ajme-1-7-21
11. Mendeş M, Yiğit S. Comparison of ANOVA-F and ANOM tests with regard to type I error rate and test power. Journal of Statistical Computation and Simulation. 2013;83(11):2093–104. doi:10.1080/00949655.2012.679942

12. Arnau J, Bendayan R, Blanca M, Bono R. Efecto de la violación de la normalidad y esfericidad en el modelo lineal mixto en diseños split-plot. 2012;24(3):449–54.
13. Hecke T. Power study of anova versus Kruskal-Wallis test. Journal of Statistics and Management Systems. 2012;15(2–3):241–7. doi:10.1080/09720510.2012.10701623
14. McDonald J. Handbook of Biological Statistics. 3ra ed. Baltimore, Maryland: Sparky House Publishing; 2014. 299 p.
15. Vásquez E, Caballero A, Herrera M. Transformación De Variables Binomiales Para Su Análisis Según Un Diseño De Bloques Al Azar. Cultivos Tropicales. 2017;38(1):108–14.
16. Rubinstein R, Kroese D. Simulation and the Monte Carlo Method. John Wiley & Sons; 2011. 401 p.
17. Kroese D, Taimre T, Botev Z, Rubinstein R. Student Solutions Manual to Accompany Simulation and the Monte Carlo Method , Student Solutions Manual. 2daed. John Wiley & Sons; 2012. 205 p.
18. Robert C, Casella G. Monte Carlo Statistical Methods. Springer Science & Business Media; 2013. 522 p.
19. Ortiz J, Moreno E. ¿Se necesita la prueba t de Student para dos muestras independientes asumiendo varianzas iguales? Comunicaciones en Estadística. 2011;4(2):139–57. doi:10.15332/s2027-3355.2011.0002.05
20. Peña D. Fundamentos de estadística. Alianza editorial; 2014. 688 p.
21. Siegel S, Castellan N. Estadística no paramétrica: aplicada a las ciencias de la conducta. 4taed. México: Editorial Trillas; 2009.

Received: October 17th, 2017
Accepted: September 5th, 2018

Revista Cultivos Tropicales
Instituto Nacional de Ciencias Agrícolas