

ALTERNATIVAS DE ANÁLISIS DE DATOS CON DISTRIBUCIÓN BINOMIAL EN DISEÑO DE BLOQUES AL AZAR

Alternatives of data analysis with binomial distribution in random block design

Edison Ramiro Vásquez^{1✉}, Alberto Caballero Núñez²
y Magaly Herrera Villafranca³

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

INTRODUCCIÓN

El trabajo conjunto entre el estadístico y el investigador (1) es imprescindible al momento de definir un modelo estadístico, que refleje en lo posible, lo que se quiere evidenciar a través de la experimentación. En estas consideraciones, uno de los modelos más difundidos constituye el Análisis de Varianza, que al ser utilizado de manera eficiente, se convierte en una poderosa herramienta de análisis. No obstante, esta técnica exige del cumplimiento de ciertos requerimientos de los términos de error aleatorio del modelo lineal, como errores independientes, normalmente distribuidos y con

varianzas homogéneas para todas las observaciones, condiciones que muchas veces no se cumplen (2–5).

En la práctica investigativa, es frecuente la presencia de variables que, de alguna manera, no satisfacen los requerimientos que el ANAVA exige (6,7); tal es el caso, de variables de conteos, que por su naturaleza discreta pueden alejarse de la normalidad. En tal sentido, algunos autores señalan (8–11) que dada la “robustez” de la prueba F en este procedimiento de análisis, su incumplimiento no tiene graves consecuencias en el análisis; que resulta prácticamente irrelevante en lo referente a la probabilidad de cometer un error tipo I (6); pues, no se aparta del valor α determinado por el experimentador. Sin embargo, la “robustez” de la prueba puede afectarse cuando este incumplimiento es severo, ya que se incrementa la probabilidad de exceder el valor nominal de la prueba (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

Dada su naturaleza y frecuente existencia en muchas ramas de la ciencia, son de importancia aquellas variables de conteos que provienen de variables dicotómicas o distribución binomial, en las que se establece una estrecha relación de dependencia entre varianza y media de tratamientos; aspecto que puede estar presente en otro tipo de variables (14). Por tanto, es de suponer que de presentarse diferencias entre las medias en cada variante que se están ensayando, sean posibles diferencias entre sus respectivas varianzas y, con ello, el no cumplimiento de este supuesto.

Indicadores como el porcentaje en que se rechaza la hipótesis nula, diferencia mínima que se puede detectar entre medias de tratamientos, potencia observada del ANAVA, número de rechazo de igualdad de medias de tratamientos (1,15); pueden recibir el impacto desfavorable cuando se incumplen los supuestos; por lo que es importante identificar, tener en cuenta y conocer su grado de afectación.

En tal virtud, en el presente artículo se valoran las técnicas estadísticas ANAVA, Comparación de Proporciones y prueba no paramétrica de Friedman, como alternativas para analizar datos con distribución Binomial en diseño de bloques al azar.

MATERIALES Y MÉTODOS

Se utilizó el proceso de Simulación de Monte Carlo (16–20) para generar poblaciones de variables aleatorias con distribución Binomial, con varianzas homogéneas y heterogéneas, según prueba de Levene a $p < 0,05$ para 5, 10 y 30 observaciones por unidad experimental (n) y probabilidad de éxito del evento de 0,10, 0,20, ..., 0,90 (p). Se conformaron experimentos en diseño de Bloques al Azar con tres, cinco y nueve tratamientos (t); cuatro y ocho réplicas (r). La combinación de medias de los tratamientos se definió de modo tal, que las diferencias entre estas medias fueran detectables por la prueba Mínima Diferencia Significativa (MDS) a un nivel de significación del 0,05 (Tabla I); para cada combinación, tratamiento-réplica-observaciones por unidad experimental, se generaron 100 experimentos.

A los datos con distribución Binomial con varianzas heterogéneas y homogéneas, se los procesó con las técnicas estadísticas ANAVA, Comparación de Proporciones y la prueba no paramétrica de Friedman.

Se utilizó la prueba de Comparación de Proporciones para contrastar la diferencia entre el porcentaje de experimentos en los que se rechaza la H_0 con el ANAVA, la Comparación de Proporciones y Friedman para experimentos con varianza entre tratamientos homogénea y heterogénea.

Tabla I. Estructura de medias y varianzas de tratamientos para las diferentes variantes de análisis

Tratamiento	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

RESULTADOS Y DISCUSIÓN

Se discute el comportamiento de indicadores estadísticos, que permiten evaluar la calidad de los procedimientos de análisis que se relacionan con los supuestos teóricos del ANAVA rechazo de la hipótesis nula y número de diferencias detectadas.

RECHAZO DE LA HIPÓTESIS NULA

En la Tabla II, se observa que el porcentaje de rechazo declarado significativo, resultó ser superior con el procedimiento de Comparación de Proporciones que los obtenidos por el ANAVA y Friedman, para tres y cinco tratamientos y número de observaciones por unidad experimental pequeños (5 y 10), no así cuando el número de unidades experimentales es grande (30); lo cual puede estar asociado con una aproximación de la variable a la normalidad y un mayor acercamiento a este supuesto que exige esta técnica de análisis.

Tabla II. Porcentaje de rechazo de H_0 en el ANAVA, Comparación de Proporciones y Friedman

Técnica:	r	n	ANAVA (%)	Proporción (%)	Friedman (%)	Sign. $\alpha=0,05$	ES _x
Varianzas heterogéneas							
3 tratamientos							
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
Varianzas homogéneas							
3 tratamientos							
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
Varianzas heterogéneas							
5 tratamientos							
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
Varianzas homogéneas							
5 tratamientos							
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
Varianzas heterogéneas y homogéneas							
9 tratamientos							
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: nivel de significación mayor a 0,05

*: nivel de significación menor a 0,05

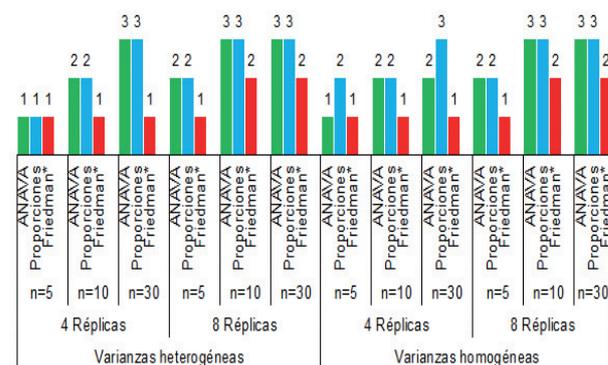
a, b: las comparaciones se realizan únicamente con el procedimiento ANAVA

Otro aspecto que no puede desestimarse en el resultado anterior, es el hecho de que el valor promedio de la probabilidad de éxito del evento de estos experimentos simulados con tres y cinco tratamientos, es de 0,30; y, el procedimiento de Comparación de Proporciones se sustenta en la distribución Chi-cuadrado, que resulta más preciso en la medida que el parámetro p de la distribución Binomial se aleja de 0,50, punto en el cual la varianza se hace máxima; esto se explica para nueve tratamientos, donde los valores promedio del parámetro p de estos experimentos es 0,50. Los resultados del indicador rechazo de la hipótesis H_0 , se iguala en los tres procedimientos de análisis.

El procedimiento no paramétrico de Friedman, mostró un bajo comportamiento, respecto a los demás procedimientos y, aún más acentuado, cuando el número de observaciones por unidad experimental y número de réplicas son pequeños ($n=5$ y 10 ; $r=4$), estos resultados corroboran lo planteado por otros investigadores (5,21), cuando sostienen que el procedimiento paramétrico es siempre más eficaz que su homólogo no paramétrico.

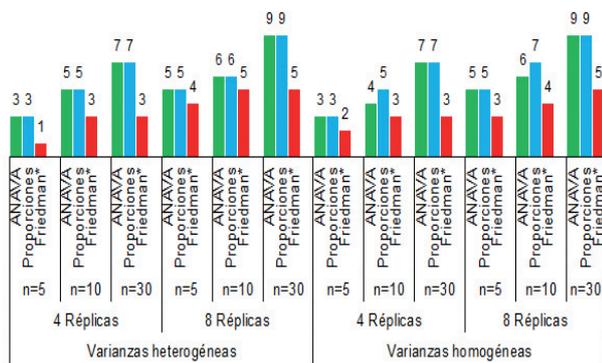
NÚMERO DE DIFERENCIAS DETECTADAS

En las Figuras 1, 2 y 3, a través de los tres procedimientos de análisis y en todas las variantes analizadas, se observó un incremento significativo en el número de diferencias detectadas, en la medida que aumenta el número de observaciones por unidad experimental y el número de réplicas. Este aspecto es más evidente para cinco y nueve tratamientos, dado que el número de comparaciones posibles son de 10 y 36, respectivamente.



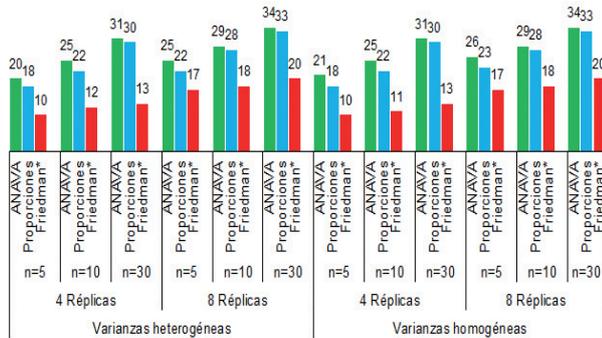
* La diferencia entre tratamientos se realizó a través de sus medianas.

Figura 1. Número de diferencias entre medias de tratamientos, detectadas por la prueba MDS ($\alpha=0,05$) para tres tratamientos



* La diferencia entre tratamientos se realizó a través de sus medianas.

Figura 2. Número de diferencias entre medias de tratamientos, detectadas por la prueba MDS ($\alpha=0,05$) para cinco tratamientos



* La diferencia entre tratamientos se realizó a través de sus medianas

Figura 3. Número de diferencias entre medias de tratamientos, detectadas por la prueba MDS ($\alpha=0,05$) para nueve tratamientos

En presencia o no de homogeneidad de varianza de la variable Binomial, a través de todas las combinaciones de número de observaciones por unidad experimental y número de réplicas, con los procedimientos ANAVA y Comparación de Proporciones, se encontraron un mayor número de diferencias entre tratamientos que con el procedimiento de Friedman, lo cual está asociado al hecho de que en este procedimiento de análisis, la observación en él es reemplazada por el rango que dicha observación ocupa a través del conjunto de tratamientos, lo que conlleva a una pérdida de la esencia de la cuantía o magnitud del dato, muy importante para este tipo de variable.

CONCLUSIONES

En sentido general, no se evidenció ventajas de las dos alternativas de análisis: Comparación de Proporciones y procedimiento no paramétrico de Friedman, en cuanto a los indicadores que reflejan la eficacia del ANAVA, lo que expresa que no mostraron ventajas respecto al ANAVA clásico del dato binomial, la cual parece ser una opción razonable para este tipo de datos.

BIBLIOGRAFÍA

- Vásquez E, Caballero A. Cuando falla el supuesto de homocedasticidad en variables con distribución binomial. *Cultivos Tropicales*. 2011;32(3):46–51.
- Di Rienzo J, Casanoves F, González L, Tablada E, Díaz M. *Estadística para las ciencias agropecuarias*. 7^{ma}ed. Córdoba, AR: Edit. Brujas; 2009. 372 p.
- 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.
- Wiedenhof S H. *Pruebas no paramétricas para las ciencias agropecuarias : muestras pequeñas*. 2^{da}. Maracay , Venezuela; 2013. 261 p.
- Bustillo C, Herrera M, Vásquez 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.
- 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
- 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
- Wetherill G. *Intermediate statistical methods* [Internet]. Springer Science & Business Media; 2012. 406 p. Available from: https://www.google.com/books?hl=es&lr=&id=dcLoCAAQBAJ&oi=fnd&pg=PR13&dq=intermediate+statistical+methods&ots=kO6fC_RyPn&sig=5tUWuX-WSErFvaNQGnAbilDkoxU
- 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
- Ostertagová E, Ostertag O. Methodology and Application of One-way ANOVA. *American Journal of Mechanical Engineering*, *American Journal of Mechanical Engineering*. 2013;1(7):256–61. doi:10.12691/ajme-1-7-21
- Mendes 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. *Psicothema*. 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*. 3^{ra} 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. 2^{da}ed. 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*. 4^aed. México: Editorial Trillas; 2009.

Recibido: 17 de octubre de 2017

Aceptado: 5 de septiembre de 2018

