

Vásquez, Edison Ramiro, Caballero Núñez, Alberto

CUANDO FALLA EL SUPUESTO DE HOMOCEDASTICIDAD EN VARIABLES CON
DISTRIBUCIÓN BINOMIAL

Cultivos Tropicales, vol. 32, núm. 3, julio-septiembre, 2011, pp. 63-68

Instituto Nacional de Ciencias Agrícolas

La Habana, Cuba

Disponible en: <http://www.redalyc.org/src/inicio/ArtPdfRed.jsp?iCve=193222357009>



Cultivos Tropicales

ISSN (Versión impresa): 0258-5936

revista@inca.edu.cu

Instituto Nacional de Ciencias Agrícolas

Cuba

CUANDO FALLA EL SUPUESTO DE HOMOCEDASTICIDAD EN VARIABLES CON DISTRIBUCIÓN BINOMIAL

When the assumption of homocedasticidad in variables with binomial distribution fails

Edison Ramiro Vásquez[✉] y Alberto Caballero Núñez

ABSTRACT. The process of Monte Carlo simulation was used to generate populations of random variables with Binomial distribution with homogeneous or heterogeneous variance; for 5, 10 and 30 observations by experimental unit (n) and likelihood of success of the event of 0.10, 0.20, ..., 0, 90 (p). It consisted of experiments in randomized block design with 3, 5 and 9 treatments (t); 4 and 8 replics (r); for each combination $t-r-n$, generated 100 experiments. By way of a reference point of discussion to have, included the variant of other 100 experiments of variables with Normal distribution, with similar mean and variance of the experiments of data with Binomial distribution. It was found that the behavior of the indicators: percentage of experiments in which there is a rejection of the hypothesis H_0 ; the power in the ANOVA; minimum difference detected in the experiment, as well as the number of differences between treatments is similar within each alternative analysis through three variants; showing a marked influence the number of observations per experimental unit and the number of replics in these indicators.

Key words: homogeneity of variance, homocedasticidad, variables binomiales, assumptions of ANOVA

INTRODUCCIÓN

El trabajo del estadístico, de manera conjunta con el investigador, consiste en conseguir un modelo que refleje en lo posible, la situación que tiene planteada y a partir de aquí, aplicar los procedimientos de análisis que más se adecue (1).

Sin duda, uno de los modelos más difundidos lo constituye el Análisis de Varianza (2); descrito por (3, 4) como una colección de modelos estadísticos y sus procedimientos asociados, apropiado para evaluar la igualdad entre poblaciones o tratamientos, en la cual, la varianza está fraccionada en ciertos componentes, debido a

Dr.C. Edison Ramiro Vásquez, Profesor Principal, Universidad Nacional de Loja (UNL), Ecuador y Dr.C. Alberto Caballero Núñez, Investigador Titular del departamento de Matemática Aplicada, Instituto Nacional de Ciencias Agrícolas (INCA), gaveta postal 1, San José de las Lajas, La Habana, Mayabeque, Cuba. CP 32 700.

✉ eramirov@hotmail.com

RESUMEN. Se utilizó el proceso de Simulación de Monte Carlo para generar poblaciones de variables aleatorias con distribución Binomial con varianzas homogéneas y heterogéneas; para cinco, 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 Bloques al Azar con tres, cinco y nueve tratamientos (t); cuatro y ocho réplicas(r); para cada combinación $t-r-n$, se generaron 100 experimentos. A modo de disponer de un referente de discusión, se incluyó la variante de otros 100 experimentos de variables con distribución Normal, con similares medias y varianzas de los experimentos de datos con distribución Binomial. Se comprobó que el comportamiento de los indicadores: porcentaje de experimentos en que se produce un rechazo de la hipótesis H_0 ; la potencia en el ANOVA; diferencia mínima detectada en el experimento, así como el número de diferencias entre medias de tratamientos, es similar dentro de cada alternativa de análisis a través de las tres variantes; evidenciando una marcada influencia el número de observaciones por unidad experimental y el número de réplicas en estos indicadores.

Palabras clave: homogeneidad de varianza, homocedasticidad, variables binomiales, supuestos del ANOVA

diferentes factores o variables explicativas, incluidas en el modelo.

No obstante, como señala (5), 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 a veces no se cumplen (6, 7, 8, 9).

En la práctica investigativa, se presentan innumerables situaciones de variables que pueden de alguna manera no satisfacer los requerimientos que esta técnica de análisis exige; tal es el caso de variables de conteos, que por su naturaleza discreta pueden alejarse de la normalidad. En tal sentido señalan (10) que dada la «robustez» de la prueba F en este procedimiento de análisis, su incumplimiento no tiene graves consecuencias en el análisis; de igual modo (11) señalan que resulta prácticamente irrelevante en lo referente a la probabilidad de cometer un error tipo I; pues, no se aparta del valor

α determinado por el experimentador. Sin embargo (12) advierte que dicha «robustez» de la prueba puede perderse cuando este incumplimiento es severo, ya que se incrementa la probabilidad de exceder el valor nominal de la prueba.

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 (13). 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.

Son muchos los aspectos que puede recibir el impacto desfavorable o no cuando se incumplen estos supuestos, entre los que pueden enumerarse: el porcentaje en que se rechaza H_0 ; la diferencia mínima que se puede detectar entre medias de tratamientos; la potencia observada del ANOVA; número de rechazo de igualdad de medias de tratamientos; entre otros. De aquí que identificar, tener en cuenta y conocer su grado de afectación, revista gran importancia.

En este contexto, en el presente artículo se pretende valorar el impacto que puede tener la presencia de heterogeneidad de varianza en variable con distribución Binomial.

MATERIALES Y MÉTODOS

Para cumplir con el objetivo propuesto se utilizó el proceso de Simulación de Monte Carlo para generar poblaciones de variables aleatorias con distribución Binomial con varianzas homogéneas y heterogéneas, según prueba de Levene a $p < 0,05$ (14, 15, 18); para cinco, diez y treinta observaciones por unidad experimental (n) y probabilidad de éxito del evento de 0,10, 0,20, ..., 0,90 (p). Se conformaron experimentos en Diseño 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 (DMS) a un nivel de significación del 0,05 (Tabla I); para cada combinación t - r - n , se generaron 100 experimentos. A modo de disponer de un referente de discusión, se incluyó la variante de otros 100 experimentos de variables con Distribución Normal, con similares medias y varianzas de los experimentos de datos con distribución Binomial.

Se utilizó la prueba de Comparación de proporciones con el fin de contrastar la diferencia entre el porcentaje de experimentos con distribución Normal de referencia y con distribución Binomial (varianza homogénea y heterogénea).

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

Tratamiento	p	n	5		10		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

Porcentaje en que se rechaza H_0 . En la Tabla II se observa que la capacidad del ANOVA para producir el rechazo de la igualdad de medias de tratamientos, son iguales entre las variantes de variables binomiales aunque estas cumplan o no el supuesto de homogeneidad de varianza, ello se debe a la robustez de la prueba F ante la violación de este supuesto aspecto que ha sido señalado por (10, 11).

De igual modo en estos experimentos con variables binomiales, comparados con los de variables con distribución normal que cumplen con todos los supuestos, ofrecen resultados similares; lo cual está asociado, al hecho que este tipo de variables son fácilmente aproximables a una distribución normal bajo determinadas condiciones, como señalan algunos autores (16).

El porcentaje de experimentos en los que se rechaza la igualdad entre medias de tratamientos aumentó de manera significativa de cuatro a ocho el número de réplicas, lo cual evidencia lo señalado por (17), acerca del marcado efecto que tiene el número de réplicas en el diseño de un experimento, en cuanto a la precisión en las estimaciones de las medias de tratamientos, incidiendo en la reducción de los intervalos de confianza de los parámetros y mayor facilidad para detectar una diferencia significativa.

El aumento del número de observaciones por unidad experimental mostró un efecto marcado en el porcentaje de rechazo de H_0 en el ANOVA, lográndose el 100 % de los rechazos de igualdad de medias de tratamientos a partir de 10 repeticiones y de ocho réplicas en el diseño, esto está asociado con varios factores, entre otros, a una mejor aproximación de la binomial a la normal por el aumento del número de unidades experimentales y de la precisión de las estimaciones por un incremento del número de réplicas en el diseño.

Tabla II. Porcentaje de rechazo de H_0 en el ANOVA, de datos con distribución: Normal y Binomial

t	r	n	Distribución Normal %	Distribución Binomial con varianza		Significación $\alpha=0,05$	ES \pm
				Homogénea %	Heterogénea %		
3	4	5	55	52	52	ns	5
		10	85	89	83	ns	4
		30	100	100	98	ns	1
	8	5	92	93	92	ns	3
		10	100	100	100	ns	0
		30	100	100	100	ns	0
5	4	5	49	54	54	ns	5
		10	91	88	85	ns	3
		30	100	100	100	ns	0
	8	5	95	92	92	ns	3
		10	100	100	100	ns	0
		30	100	100	100	ns	0
9	4	5	100	100	100	ns	0
		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

Potencia observada. De manera general, las variables binomiales aunque incumplan o no el supuesto de homogeneidad de varianza, su potencia es similar a la observada en la variable normal que cumple todos los supuestos del AVOVA, esto evidencia que el incumplimiento de estos supuestos no tiene efectos desfavorables en el error Tipo II, es decir, aceptar como igual diferencia entre medias de tratamientos que debiera ser declarada significativamente diferente (Tabla III).

Dentro de cada alternativa y a través de las tres variantes de análisis se evidencia lo señalado por (17) y toda la literatura especializada respecto al marcado efecto que ejerce el número de réplicas sobre la potencia de la prueba estadística. De este modo con número de réplicas igual a ocho se alcanzan potencias superiores a 85 %, para cualquier número de unidades experimentales (n) (Tabla III).

Diferencia mínima a detectar entre medias de tratamientos. En la Tabla IV se puede observar que teniendo en cuenta que la media en variables binomiales es función directa del número de unidades experimentales (n), cualquier análisis comparativo entre ellas puede conducir a conclusiones erradas, debido a que son generadas con diferentes medias (Tabla I); de aquí que para efectuar comparaciones justas entre alternativas y variantes de trabajo se emplee el indicador por ciento de diferencia mínima a detectar expresada como porcentaje de la media general del experimento.

Resulta significativo el aumento de la capacidad de la prueba DMS para detectar diferencias más pequeñas cuando se incrementa del número de réplicas (r) por tratamiento y el número de observaciones en las unidades experimentales (n); ello está asociado al efecto marcado que estos factores ejercen en la precisión en las estimaciones, dado por una reducción del intervalo de confianza del parámetro señalado por (16), lo cual puede advertirse a través del valor $z s_{\bar{x}}$.

No se observa diferencias notables en este indicador, entre las variables con distribuciones normal y binomial cuando el supuesto de homogeneidad de varianza es satisfecho; sin embargo, en la variable con distribución binomial donde se incumple este supuesto, la capacidad de esta prueba para detectar diferencia significativa entre medias de tratamientos se ve afectada considerablemente sobre todo si el número de observaciones por unidad experimental y el número de réplicas por tratamiento son pequeños ($n=5$ y $r=4$).

Número de diferencias entre medias de tratamientos detectadas en cada experimento. Dado que el número de comparaciones posibles entre medias de tratamientos varía con el número de tratamiento, cualquier análisis entre experimentos con diferentes números de tratamiento carecería de sentido, debido a que en la medida que el número de comparaciones aumenta, la posibilidad de detectar diferencias verdaderas entre medias disminuye (Tabla V).

Tabla III. Potencia observada en el ANOVA con $\alpha=0,05$, de datos con distribución Normal y Binomial

t: Tratamiento r: Réplica n: Observaciones			Distribución Normal	Distribución Binomial con varianza		Significación $\alpha=0,05$	ES \pm
t	r	n	%	Homogénea	Heterogénea		
3	4	5	63	60	59	ns	5
		10	74	74	72	ns	4
		30	99	98	94	ns	2
	8	5	87	88	87	ns	3
		10	99	98	98	ns	1
		30	100	100	100	ns	0
5	4	5	64	66	66	ns	5
		10	88	87	84	ns	3
		30	100	100	99	ns	1
	8	5	91	90	93	ns	3
		10	99	99	99	ns	1
		30	100	100	100	ns	0
9	4	5	100	100	100	ns	0
		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

Tabla IV. Intervalo de confianza al 95 % para la diferencia mínima detectada entre medias de tratamientos con la prueba DMS ($\alpha=0,05$)

t: Tratamiento r: Réplica n: Observaciones			Distribución Normal			Distribución Binomial con varianza		
t	r	n	\bar{x}	\pm	$z s_{\bar{x}}$	\bar{x}	\pm	$z s_{\bar{x}}$
3	4	5	1,6		0,1	1,6		0,1
		10	2,1		0,1	2,1		0,1
		30	3,6		0,2	4,1		0,3
	8	5	1,0		0,0	1,0		0,0
		10	1,4		0,1	1,4		0,1
		30	2,4		0,1	2,4		0,1
5	4	5	1,4		0,1	1,4		0,1
		10	2,1		0,1	2,1		0,1
		30	3,7		0,1	3,5		0,2
	8	5	1,0		0,0	1,0		0,0
		10	1,4		0,0	1,4		0,0
		30	2,4		0,1	2,5		0,1
9	4	5	1,3		0,0	1,4		0,0
		10	2,0		0,1	1,9		0,0
		30	3,5		0,1	3,3		0,1
	8	5	0,9		0,0	1,0		0,0
		10	1,3		0,0	1,3		0,0
		30	2,3		0,0	2,3		0,0

Tabla V. Intervalo de confianza al 95 % para el número de diferencias entre medias de tratamientos detectadas por la prueba MDS ($\alpha=0,05$)

t: Tratamiento r: Réplica n: Observaciones			Distribución Normal Varianza homogénea			Distribución Binomial con varianza					
t	r	n	\bar{x}	\pm	$z s_{\bar{x}}$	Homogénea			Heterogénea		
						\bar{x}	\pm	$z s_{\bar{x}}$	\bar{x}	\pm	$z s_{\bar{x}}$
3	4	5	1	\pm	0	1	\pm	0	1	\pm	0
		10	2	\pm	0	2	\pm	0	2	\pm	0
		30	3	\pm	0	3	\pm	0	2	\pm	0
	8	5	2	\pm	0	2	\pm	0	2	\pm	0
		10	3	\pm	0	3	\pm	0	3	\pm	0
		30	3	\pm	0	3	\pm	0	3	\pm	0
5	4	5	3	\pm	0	3	\pm	0	3	\pm	0
		10	5	\pm	0	5	\pm	0	4	\pm	0
		30	7	\pm	0	7	\pm	0	7	\pm	0
	8	5	5	\pm	0	5	\pm	0	5	\pm	0
		10	7	\pm	0	6	\pm	0	6	\pm	0
		30	9	\pm	0	9	\pm	0	9	\pm	0
9	4	5	21	\pm	1	20	\pm	1	21	\pm	1
		10	25	\pm	0	25	\pm	1	25	\pm	0
		30	30	\pm	0	31	\pm	0	31	\pm	0
	8	5	25	\pm	0	25	\pm	0	26	\pm	0
		10	29	\pm	0	29	\pm	0	29	\pm	0
		30	34	\pm	0	34	\pm	0	34	\pm	0

Para cada número de tratamiento el número de comparación en la que se rechaza la diferencia entre medias de tratamientos tanto para la variable normal como las variables binomiales que cumplan o no con la igualdad de varianza, el número de diferencias detectadas es similar.

Sin embargo, es notorio el aumento de este indicador con el aumento del número de observaciones por unidad experimental (n) y el número de réplicas (r), lo cual está asociado a una mayor precisión en la estimación de este parámetro, que hace más fácilmente detectable una diferencia entre medias como se señala en la discusión correspondiente a la Tabla II.

REFERENCIAS

- Khan, A. y Rayner, D. Robustness to Non-Normality of Common Tests for the Many-Sample Location Problem. *J. Appl. Mathematics & Decision. Sci.*, 2003, vol. 7, no. (18)
- Norell, L. ANOVA Estimators under imbalance in the one-way random model. *Communications in Statistics: Theory & Methods*. 2003, vol. 32, no. 601.
- Tejedor, F. J. Análisis de varianza: introducción conceptual y diseños básicos. Madrid: La Muralla. 1999, 132 p.
- Spiegel, M. R.; Schiller, J. J. y Srinivasan, R. A. Teoría y problemas de probabilidad y estadística. 2 ed. México D.F.: McGraw-Hill. 2003, 416 p.
- Vilar, J. Curso de Estadística 2. La Coruña: Universidad de La Coruña. [Consultado: 15 febrero 2011]. Disponible en: <http://www.udc.es/dep/mate/estadistica2/estadistica_2.htm>. 2009.
- Cochran, W. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 1947, vol. 3, no. 1, p. 22-38
- Eisenhart, C. The assumptions underlying the analysis of variance. *Biometrics*, 1947, vol. 3, no. 1, p. 03-21.
- Díaz, A. Diseño estadístico de experimentos. 2 ed. Antioquia, Colombia: Editorial Universidad de Antioquia. ISBN: 978-958-714-264-8. [Consultado: 15 octubre 2009] Disponible en: <http://books.google.es/books?id=0x0DW6dNiyAC&pg=PA12&dq=estadistica+dise%C3%B1o+experimental&hl=es&ei=vOijTY3TMJTbtgfh5TYCw&sa=X&oi=book_result&ct=result&resnum=1&ved=0CDQ6AEwAA#v=onepage&q&f=trueSTAT> 2009.

9. INFOSTAT. Infostat versión 1.1 Manual del usuario. Grupo Infostat, FCA. Universidad Nacional de Córdoba. Primera Edición. Ed. Brujas. 2002, p. 61-90.
10. Steel, R. y Torrie, J. Bioestadística. Principios y Procedimientos. 2ª ed. McGraw-Hill Interamericana. México. 1988.
11. Sokal, R. y Rohlf, J. Biometry. The Principles and Practice of Statistics in Biological Research. 3ª ed. W. H. Freeman and Company. New York. 1995.
12. Wetherill, G. B. Intermediate Statistical Methods. Chapman and Hall. London. 1981.
13. Little, T. y Jackson, F. Métodos Estadísticos para la Investigación en la Agricultura. 2ª ed. Editorial Trillas. México. 1991.
14. Levene, H. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin *et al.* eds., Stanford University Press, 1960. p. 278-292.
15. Brown, M. y Forsythe. Robust Test for the Equality of Variances. *Journal of the American Statistical Association*, 1974, vol. 69, no. 346, p. 364-367.
16. Martín, A. y Álvarez, Roxana. Aproximación de la Distribución Binomial a la Normal, la calculadora y las TIC. SIGMA 2006.
17. Cochran, W. y Cox, G. Diseños Experimentales. 3ª ed. Editorial Trillas. México. 1997.
18. O'Neil, E. y Mathews, L. Levene tests of homogeneity of variance for general block and treatment designs. *Biometrics*, 2002, vol. 58, p. 216-224.

Recibido: 18 de marzo de 2010

Aceptado: 8 de junio de 2011