ANÁLISIS DE LA PRECISIÓN DE MANOVA EN HOJAS DE CÁLCULO DE RECIENTE DISPONIBILIDAD

ANALYSIS OF THE ACCURACY OF A MANOVA IN SPREADSHEETS OF RECENT AVAILABILITY

Juan Pablo González Morales
Tecnológico Nacional de México en Celaya, México
jpabloglezm@hotmail.com

Manuel Darío Hernández Ripalda
Tecnológico Nacional de México en Celaya, México
dario.hernandez@itcelaya.edu.mx

José Alfredo Jiménez García
Tecnológico Nacional de México en Celaya, México
alfredo.jimenez@itcelaya.edu.mx

Recepción: 22/noviembre/2019 **Aceptación:** 20/diciembre/2019

Resumen

Existen diferentes alternativas para la realización de cálculos estadísticos en el ámbito académico y laboral. Sin embargo, poco se conoce acerca de la precisión que actualmente tienen las hojas que se ubican en la nube y de la confiabilidad que estas podrían brindar en cálculos que requieran de gran precisión. El propósito de este artículo es el de brindar una revisión de precisión bajo los estandartes establecidos por el Instituto Nacional de Estándares y Tecnología (NIST) para el análisis de varianza multivariante. Este artículo provee un estudio de la precisión en matrices varianza-covarianza utilizando la metodología MANOVA con el método de cuadrados media, en conjunto con la matriz de correlación.

Para este propósito, se utilizan los softwares libre acceso LibreOffice, Excel, Google Spreadsheet y Gnumeric para determinar y conocer las cifras significativas en donde se ve afectada la precisión. En este artículo muestra cuales softwares tienen una mejor precisión al realizar un análisis multivariante.

Palabras Clave: Hojas de cálculo, MANOVA, Precisión del software, Softwares estadísticos

Abstract

There are different alternatives for performing statistical calculations in the academic and labor field. Little is known about the accuracy of the leaves that are located in the cloud and the reliability they could provide in calculations that require high precision. The purpose of this article is to provide a precision review under the standards established by the National Institute of Standards and Technology (NIST) for multivariate analysis of variance. This article provides a study of the precision in variance-covariance matrices using the MANOVA methodology with the mean square method, in conjunction with the correlation matrix.

For this purpose, LibreOffice, Excel, Google Spreadsheet and Gnumeric free access software are used to determine and know the significant figures where accuracy is affected. This article shows which softwares have better accuracy when performing a multivariate analysis.

***Keywords*:** MANOVA, Software Accuracy, Spreadsheets, Statistical Software.

1. Introducción

Existen hoy en día una extensa gama hojas de cálculo, empleadas en distintos campos de la estadística, los cuales proporciona una vasta variabilidad entre sus resultados, es de suma importancia reconocer cuales son estas discrepancias que se presentan al determinar sus valores. [Knüsel, 1995] se cuestiona ¿Cuál es promedio requerido en las distribuciones estadísticas en programas computacionales? Si la respuesta esperada es 0.0000 y el valor obtenido 5.0×10^{-5}, se considera correcta y confiable. Sin embargo, si el valor esperado es 4.820×10^{-15} y el obtenido es 4.073×110^{-15}, es considerado como inaceptable [Yalta, 2008].

La hoja de cálculo más utilizada es Microsoft Excel, paralelamente existen una extensa rama de opciones de hojas de cálculo con funciones semejantes, pero con una mayor precisión. Las diferencias observadas en la precisión de las hojas fueron en todos los ámbitos de la estadística.Las primeras pruebas realizadas fueron en Excel 97 sobre todas las distribuciones de probabilidad [Knusel, 1998]. Se realizaron una serie de pruebas con las distintas hojas de cálculo, y bastantes diferencias se encontraron y fueron reportadas [McCullough & Wilson, 1998].

Microsoft no realizó las correcciones pertinentes, ya que Excel 97 y todas sus versiones posteriores hasta llegar a Excel 07 presentan errores [McCullough & Wilson, 2002] [McCullough & Wilson, 2004]. Yalta [2008] lo contrasta y lo publica en The Accuracy of Statistical Distributions in Microsoft Excel 2007, comparando los softwares Calc y Gnumeric. McCullough & Heiser [2008] continuaron con las pruebas en Excel, utilizando experimentación en pruebas de regresión lineal y los errores residuales, concluyendo que era un error asumir que los resultados que se obtienen como válidos e invitaban a usar otros paquetes para un resultado correcto. Continuaron las pruebas en otros aspectos de la estadística, ahora aplicadas a los temas concernientes a los métodos de regresión lineal y no lineales, análisis de correlación, mediciones de la media y desviación estándar [Almiron, Almeida, & Miranda, 2009].

Versiones de Microsoft Excel comparadas contra Gnumeric, Google Docs, Numbers y Open Office, fueron utilizadas en la prueba del Logaritmo del error relativo [McCullough & Yalta, 2013].

Los análisis fueron concretos, la hoja de cálculo Gnumeric presentó la mayor confiabilidad, y la mejor precisión.

El caso de medición multivariada

En la década de los ochentas en la industria automotriz se inició con el análisis de estudios de sistemas de medición (gauge R&R) e índices de señal de ruido. Posteriormente se ha propuesto el análisis multivariante de sistemas de medición usando MANOVA para manejar la dependencia de varios factores, ya sean cualitativos o cuantitativos (Majeske, 2008).

El análisis de precisión que se realiza en este artículo determina cual es la mejor opción al momento de realizar con MANOVA un análisis multivariante para un sistema de medición, y muestra donde está ocurriendo cambios en las cifras significativas. Dependerá de la precisión de cada estudio y de que software se utiliza. Sin embargo, un software confiable en el aspecto de la precisión nos la brinda Gnumeric.

2. Métodos

Los errores numéricos se generan cuando en los estudios se requiere hacer aproximaciones, producto de las operaciones aritméticas y se precisa obtener un resultado exacto. Los tipos de error que se generan se incluyen dentro de los errores de truncamiento, estos resultan del uso de aproximaciones como un procedimiento matemático en el cual se busca un resultado exacto. Y los errores de redondeo que se producen cuando se usan valores que tienen un límite de cifras significativas para representar números exactos [Chapra & Raymond, 2007].

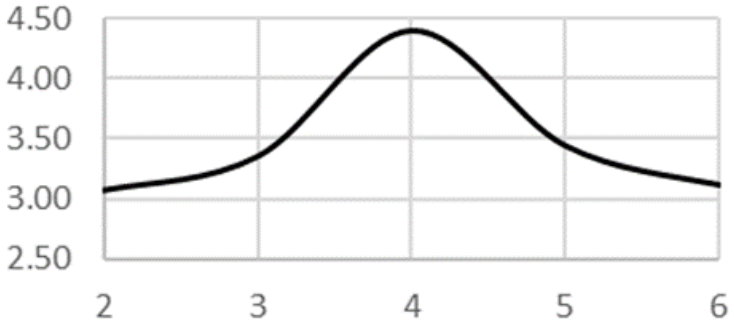
Para la determinación de la precisión en el presente artículo, se habrá de emular el proceso que estableció [McCullough, 1998] en el cual se establece el uso de los logaritmos del error, para determinar cuál es la cifra significativa donde está presentándose la diferenciación de los valores obtenidos.

Logaritmo del error relativo (LRE)

Los parámetros que utilizará el presente trabajo serán los errores, relativo y error absoluto. El error relativo (LRE) es un logaritmo de base 10, ecuación 1.

$$LRE\ (x, c) = -\log_{10}\left(\frac{|x - c|}{|c|}\right) \tag{1}$$

Donde el valor x es el resultado de la evaluación en la prueba (valor estimado) y c es el valor certificado correspondiente (valor correcto). Una representación (figura 1) particular de como la función de logaritmo indica la cifra significativa donde ocurre la diferencia de los valores se presenta en la tabla 1.



Fuente: Elaboración propia.

Figura 1 Cifra significativa de la función LRE, LRE contra cifra significativa.

Tabla 1 Valores certificados “c” y estimados “x”

C	*X*	*LRE (X,C)*
2.5111	2.508	2.9085
2.5111	2.509	3.0776
2.5111	2.51	3.3585
2.5111	2.511	4.3999
2.5111	2.512	3.4456
2.5111	2.513	3.1211
2.5111	2.514	2.9375

Logaritmo del error absoluto (LRA)

En los casos, cuando el valor estimado y el valor correcto son exactamente iguales $x = c$, el error relativo se indetermina. Bajo estas circunstancias se da pie al uso del error absoluto (LRA) que será aplicado cuando el valor objetivo sea 0, el cual está definido por ecuación 2.

$$LRA(x) = -\log_{10}|x| \tag{2}$$

MANOVA

Se ha utilizado el análisis de sistemas de medición en la industria manufacturera, sus principales sistemas de medición se enfocan en la evaluación de la medición de repetitibilidad y reproducibilidad en estudios GRR.

Cuando las múltiples mediciones de interés de un objeto siguen una distribución normal multivariante, Se ha probado distintas técnicas para determinar cuál provee de una mejor actuación, para su implementación. Una comparativa entre MANOVA y análisis de componentes principales (PCA), para un estudio GRR, determinó que es el estudio de MANOVA el mejor soporte en un estudio con múltiples características [Wang, 2013].

Este artículo utiliza la metodología que estipulo Almirón (2009) sobre el criterio de aprobación de Majeske (2008) para un sistema de medición multivariado, aquí se realiza un estudio de la precisión en la matriz de varianza-covarianza calculada utilizando los softwares: Excel, LibreOffice y Google Spreadsheet, Se utilizó como valor de referencia el calculado con Gnumeric.

3. Resultados

La tabla 2 muestra los valores concernientes a un análisis de cinco variables para 6 objetos, para el análisis multivariante que habrá de proveer las matrices varianza-covarianza y medir su precisión.

Tabla 2 Valores de las 5 variables para distintos 6 objetos medidos.

	Variables				
Objetos	**V1**	**V2**	**V3**	**V4**	**V5**
1	182	94	28.5	29	1
2	174	85	27.5	36	1
3	174	82	27.5	25	1
4	192	99	30.0	23	1
5	179	63	27.5	25	1
6	150	48	23	40	0

La matriz de la varianza está determinada por ecuación 3.

$$s = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix} = \frac{1}{n-1}\left[Y'Y - Y'\left(\frac{1}{n}J\right)Y\right] \quad (3)$$

Donde la matriz de correlación está dada por ecuación 4.

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix} = D_s^{-1} s D_s^{-1} \quad (4)$$

Donde $D_S = diagonal\ (\sqrt{s_{11}}, \sqrt{s_{22}}, \ldots \sqrt{s_{pp}}.)$

El procedimiento se utiliza en todas las hojas de cálculo, en todas se obtiene la matriz de varianza-covarianza y la matriz de correlación, la tabla 3 muestra el valor donde ocurre el cambio de la cifra significativa, entre el valor calculado y el valor de referencia, indica además a que elemento de la matriz corresponde: Las celdas con "-" indican que no hubo diferencia. Se incluye en el análisis de precisión, el estudio de la matriz correlación, con los mismos softwares, los resultados se muestran en la tabla 4 cabe resaltar que, en esta matriz, el valor de correlación que es compartido $(r_{xy} = r_{yx})$ en los términos $r_{15}, r_{23,} r_{25}$, tiene variación en la cifra significativa.

Tabla 3 Valores de la cifra significativa contra valor de referencia obtenido con Gnumeric.

Elemento de la matriz	Software utilizado vs Valor de referencia		
	Excel	**LibreOffice**	**GoogleSpreadsheet**
s_{11}	14	14	14
s_{22}	-	-	-
s_{33}	14	14	14
s_{44}	14	14	14
s_{55}	14	14	14
s_{12}	-	-	-
s_{13}	13	13	13
s_{14}	15	15	14
s_{15}	14	14	14
s_{23}	14	14	14
s_{24}	14	14	14
s_{25}	-	-	-
s_{34}	14	14	14
s_{35}	14	14	14
s_{45}	14	14	14

Tabla 4 Valores matriz de correlación contra valor de referencia obtenido con Gnumeric..

Elemento de la matriz correlación	Software utilizado vs Valor de referencia		
	Excel	**LibreOffice**	**GoogleSpreadsheet**
r_{11}	15	15	15
r_{22}	-	-	-
r_{33}	-	-	-
r_{44}	-	-	-
r_{55}	15	15	15
r_{12}	15	15	15
r_{13}	13	13	13
r_{14}	15	15	15
r_{15}	0/15	0/15	0/15
r_{23}	15/14	15/14	15/14
r_{24}	14	14	14
r_{25}	15/0	15/0	15/0
r_{34}	-	-	-
r_{35}	14	14	14
r_{45}	-	-	-

4. Discusión

Al realizar un estudio de matriz de varianza-covarianza multivariante, los programas Excel y LibreOffice presentaron una mejor consistencia en la precisión, al momento de trabajo con respecto a la nube con Google Spreadsheet, se presentaron varios aspectos: primero es complicado el uso al no estar en línea, y presenta una diferencia en una unidad más respecto a los demás. Y para resaltar los resultados tenemos la matriz de correlación que debería tener simetría y el análisis de la precisión nos muestra magnitudes distintas para correlaciones que se suponían iguales, situación que se presenta en varias ocasiones.

El uso de Excel y LibreOffice nos proporciona el mismo sesgo en comparativa contra los valores proporcionados por Gnumeric. Hasta donde se tiene explorado no se conoce en la literatura un estudio comparativo de precisión del cálculo con análisis multivariado en programas de hojas de cálculo.

En perspectiva serán: el usuario, manejo de la información y la necesidad de precisión para el análisis de la información quienes marquen la pauta para el uso de estos softwares.

6. Bibliografía y Referencias

[1] Almiron, M. G., Almeida, E. S., & Miranda, M. N. (2009). The reliability of statical functions in four software packages freely used in numerical computation. (B. S. Association, Ed.) Brazilian Journal of Probability and Statistics, 23(2), 107-119.

[2] Chapra, S. C., & Raymond, P. C. (2007). Métodos numéricos para ingenieros (5ta Edición ed.). McGraw Hill.

[3] Hair, J. F., Anderson, R. E., & Tatham, R. L. (1999). Análisis Multivariante (5ta ed.). Prentice Hall.

[4] Knüsel, L. (1995). On the Accuracy of the Statiscal Distributions in Gauss. Computational Statistics and Data Analysis(20), 699-702.

[5] McCullough, B., & Heiser, D. a. (2008). On the accuracy of statistical procedures in Microsoft Excel 2007. Computational Statistics and Data Analysis(52), 4570-4578.

[6] Knusel, L. (1998). On the Accuracy of Statistical Distributions in Microsoft Excel 97.

[7] Majeske, K. D. (2008). Approval Criteria for Multivariate Measurement Systems. Journal of Quality Technology, 140-153.

[8] McCullough. (1998). Assessing the Reliability of Statistical Software: Part I. The American Statistician, 52(4), 358-366.

[9] Yalta, T. A. (Enero de 2013). Spreadsheets in the Cloud - Not Ready Yet. Journal of Statistical Software, 52(7), 1-14.

[10] McCullough, B., & Wilson, B. (1998). On the accuracy of statistical procedures in Microsoft Excel 97. Computational statistics, 4(3), 27-37.

[11] McCullough, B., & Wilson, B. (2002). On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. Computational Statistics & Data Analysis, 713-721.

[12] McCullough, B., & Wilson, B. (2002). On the accuracy of Statistical procedures in Microsoft Excel 2000 and Excel XP. Computational Statistics & Data Analysis, 713-721.

[13] McCullough, B., & Wilson, B. (2004). On the accuracy of statistical procedures in Microsoft Excel 2003. Computational Statistics & Data Analysis, 1244-1252.

[14] McCullough, B. D., & Yalta, A. T. (2013). Spreadsheets in the cloud− not ready yet. Journal of Statistical Software, 52(7), 1-14.

[15] NIST. (2003). National Institute of Standars and Technology : The statistical reference datasets. Gaithersburg, USA: https://www.itl.nist.gov/div898/strd/.

[16] Yalta, A. T. (January 2008). The Accuracy of Statiscal Distributions in Microsoft Excel 2007. 1-13.