UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2014, Vol. 7, No. 2, pp. 191–201

Una nota de cuidado sobre el efecto de datos parcialmente faltantes en la prueba de independencia χ^2

A cautionary note on the effect of partially-missing data in the χ^2 test of independence

Juan Carlos Correa[a]
jcorrea@unal.edu.co

Jorge Iván Vélez[b]
jorgeivanvelez@gmail.com

Resumen

El análisis de tablas de contingencia se utiliza ampliamente en muchas disciplinas, siendo el principal interés la determinación de posibles asociaciones entre dos variables categóricas. Una de las pruebas más utilizadas para este fin es la prueba de independencia basada en el estadístico χ^2. Con frecuencia, los investigadores enfrentan situaciones en las que una de las dos variables (o, en el peor de los casos, ambas) es parcialmente observada (es decir, presenta algunos valores faltantes). Por lo general, el procedimiento en estos casos es excluir de los análisis aquellas observaciones (i.e., sujetos) en los que por lo menos para una de las variables no se tiene información. En esta nota analizamos el efecto de no considerar observaciones parcialmente observadas en la prueba cuando ajustamos modelos loglineales, concentrándonos principalmente en la prueba de independencia χ^2.

***Palabras clave*:** datos faltantes, modelos loglineales, prueba de independencia, tablas de contingencia.

Abstract

The analysis of contingency tables is widely used in many areas, being its main interest to determine any potential associations between two categorical variables. To disclose these associations, it is common to use a test of independence based on the χ^2 statistic. However, it is often the case that researchers face situations in which

[a]Profesor Asociado, Escuela de Estadística, Universidad Nacional de Colombia, Sede Medellín. Grupo de Investigación en Estadística, Universidad Nacional de Colombia, Sede Medellín.

[b]Genomics and Predictive Medicine Group, Department of Genome Biology, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia. Grupo de Investigación en Estadística, Universidad Nacional de Colombia, Sede Medellín. Grupo de Neurociencias, Universidad de Antioquia, Medellín, Colombia.

at least one of the variables is partially observed. In general, to perform the χ^2 test, the procedure has been to exclude observations with incomplete information. In this cautionary note, we analyze the effect of not considering partially-missing information on the χ^2 test of independence when loglinear models are fitted to the data, focusing our attention on the test of independence.

***Keywords*:** contingency tables, loglinear models, missing data, test of independence.

1. Introducción

Una de las pruebas que más se utilizan en el trabajo diario los analistas de datos es la prueba χ^2 para verificar la independencia entre dos variables categóricas (McHugh 2013). En muchas ocasiones, es común enfrentarse a conjuntos de datos parcialmente observados, es decir, aquellos en los que no se tiene información para algunas variables. En el caso especial en el que las variables son parcialmente perdidas al azar (MCAR por sus siglas en inglés) (Rubin 1976, Little & Rubin 2002), al no considerar esta información parcial estaríamos desaprovechando información que podría reforzar (o refutar) las conclusiones que se obtienen cuando esta se descarta.

Tabla 1: *tabla de contingencia para dos variables categóricas X e Y en presencia de información faltante (F) en las filas y/o en las columnas. Fuente: elaboración propia.*

	Variable Y				Totales	
Variable X	1	2	$\cdots$	J	Observados	Faltantes
1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,J}$	n_{1+}	n_{1F}
2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,J}$	n_{2+}	n_{2F}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
I	$n_{I,1}$	$n_{I,2}$	$\cdots$	$n_{I,J}$	n_{I+}	n_{IF}
Observados	n_{+1}	n_{+2}	$\cdots$	n_{+J}	n_{++}	$-$
Faltantes	n_{F1}	n_{F2}	$\cdots$	n_{FJ}	n_{F+}	$-$

Los modelos loglineales (Agresti 1990, Andersen 1997, Grizzle et al. 1969, Haberman 1972), un caso especial del modelo lineal generalizado Poisson (McCullagh & Nelder 1983, Zeileis et al. 2008, Vélez & Correa 2013), pueden ser vistos como la extensión de una tabla de contingenca de doble entrada donde la relación entre las variables categóricas que la componen, y que se analiza tomando del logarítmo natural de la frecuencia en cada una de las celdas (Jeansonne 2014). La variable respuesta corresponderá al resultado de transformar las entradas de una tabla de contingencia que, en el caso más simple, estaría conformada por dos variables aleatorias categóricas y tendría una estructura similar a la presentada en la Tabla 1. Si bien los modelos loglineales pueden utilizarse para determinar la asociación

que existe entre dos variables categóricas, es mucho más frecuente utilizarlos para tablas de mayor dimensión.

La estimación por máxima verosimilitud de modelos loglineales, entre los cuales se incluye el modelo de independencia, depende del conjunto de estadísticos suficiente minimal; estos estadísticos se reducen a conjuntos de subtablas con conteos que permiten estimar tanto los valores sus esperados, como los parámetros del modelo (Agresti 1990). Para efectos de estimación de las probabilidades requeridas por el modelo loglineal, es necesario incluir la información de los datos incompletos en estas subtablas.

En la práctica, quienes están a cargo del análisis de los datos utilizan programas estadísticos a fin de facilitar la construcción de tablas de contingencia y posteriormente realizar la prueba de independencia χ^2. Desafortunadamente, la mayoría de estos programas, incluido R (R Core Team 2014), descarta la información parcialmente observada, por lo que es necesario verificar si la conclusión es la misma al hacer la corrección usando todos los datos.

En esta nota ilustramos el efecto de no incluir observaciones parcialmente observadas en la prueba de independencia cuando se utilizan modelos loglineales. Adicionalmente, se presentan los estadísticos suficientes minimales y se desarrollan los estimadores para las probabilidades marginales de la Tabla 1 cuando incluimos la información faltante. Finalmente presentamos un ejemplo como ilustración y, en los Apéndices A y B, proporcionamos funciones en R para realizar la prueba de independencia χ^2 considerando la información faltante; la función `PruebaChi2FaltantesTabla` recibe como argumento una tabla de frecuencias similar a la Tabla 1, mientras la función `PruebaChi2FaltantesVariables` tiene como argumentos dos variables categóricas. Como resultado ambas funciones entregan el valor de los estadísticos χ^2 y los valores p cuando se incluye y excluye la información faltante.

2. Efecto de datos faltantes

Consideremos dos variables categóricas X e Y, la primera con I categorías y la segunda con J categorías, como se muestra en la Tabla 1. Cuando la tabla de 2×2 es observada completamente, el conjunto de estadísticos suficiente minimal para la prueba de independencia está dado por el conjunto de datos observado marginalmente, es decir, por $\{n_{i+}, n_{+j}\}$, $i = 1, 2, \ldots, I, j = 1, 2, \ldots, J$. Sin embargo, cuando existen observaciones parcialmente faltantes, estos estadísticos son

$$n_{i+T} = n_{i+} + n_{iF} \qquad i = 1, 2, \ldots, I \tag{1}$$

$$n_{+jT} = n_{+j} + n_{Fj} \qquad j = 1, 2, \ldots, J. \tag{2}$$

Bajo un esquema de muestreo multinomial, es posible determinar el efecto de las observaciones parcialmente faltantes sobre la prueba de independencia χ^2. Asu-

miendo independencia entre las variables X e Y, la probabilidad de cada una de las marginales está dada por

$$\hat{\pi}_{i+} = \hat{\pi}^*_{i+}\frac{n_{++}}{n^{\text{Fila}}_{++T}} + \frac{n_{iF}}{n^{\text{Fila}}_{++T}} \quad (3)$$

$$\hat{\pi}_{+j} = \hat{\pi}^*_{+j}\frac{n_{++}}{n^{\text{Columna}}_{++T}} + \frac{n_{Fj}}{n^{\text{Columna}}_{++T}}, \quad (4)$$

con $\hat{\pi}^*_{i+} = \frac{n_{i+}}{n_{++}}$, $\hat{\pi}^*_{+j} = \frac{n_{+j}}{n_{++}}$, $n^{\text{Fila}}_{++T} = \sum_{i=1}^{I} n_{i+T}$, $n^{\text{Columna}}_{++T} = \sum_{j=1}^{J} n_{+jT}$ y $n_{++} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$.

El estadístico χ^2 será entonces

$$\chi^2_c = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}, \quad (5)$$

con o_{ij} y $e_{ij} = n_{++}\hat{\pi}_{i+}\hat{\pi}_{+j}$ respectivamente, el valor observado y esperado para la ij-ésima celda de la tabla de contingencia al corregir por los datos parcialmente faltantes ($i = 1, 2, \ldots, I; j = 1, 2, \ldots, J$). Finalmente, para un nivel de significancia predeterminado $\alpha \in (0, 1)$, rechazamos la hipótesis de independencia si $p = P\left\{\chi^2_c > \chi^2_{(I-1)(J-1)}\right\} < \alpha$.

3. Distribución del estadístico χ^2_c

3.1. Relación entre χ^2_c y χ^2

Supongamos que se tiene una tabla de contingencia de $I \times J$ con vector de probabilidades $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \ldots, \pi_{IJ})$, y sean $\mathrm{S}_c = n_{++}\boldsymbol{\pi}$ y $\mathrm{S} = (n_{++} - m)\boldsymbol{\pi} = \mathrm{S}_c - m\boldsymbol{\pi}$ las tablas de contingencia incluyendo observaciones parcialmente faltantes y solo información completa, respectivamente. Aquí, m es el número total de datos faltantes en la tabla de contingencia. Bajo H_0, los estadísticos para la prueba de independencia están dados por

$$\chi^2_c = \sum_{i=I}^{I}\sum_{j=J}^{J}\frac{(\mathrm{S}_{c_{ij}} - e_{ij})^2}{e_{ij}} \quad (6)$$

y

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(\mathrm{S}_{ij} - e'_{ij})^2}{e'_{ij}} \quad (7)$$

con $e_{ij} = n_{++}\hat{\pi}_{i+}\hat{\pi}_{+j}$ y $e'_{ij} = (n_{++} - m)\pi_{i+}\pi_{+j}$. Observe que cuando $m = 0$, los estadísticos son equivalentes. Sin embargo, cuando $m > 0$, $\mathrm{S}_c > \mathrm{S}$ y $e_{ij} > e'_{ij}$ $\forall i, j$. Por lo tanto, $\chi^2_c > \chi^2$. Como se mencionará más adelante, esta desigualdad se cumple toda vez que $m/N \to 0$.

3.2. Convergencia de χ^2_c

Al final de la sección §2 se mencionó que la prueba de independencia en presencia de información parcialmente faltante se rechazaría si $p < \alpha$. Implícitamente, este cálculo asume que $\chi^2_c \sim \chi^2_{(I-1)(J-1)}$ bajo H_0. Es fácil mostrar que esto es cierto si y sólo si el número de datos parcialmente faltantes en la tabla de contingencia es cero (*véase* sección §3.1). Sin embargo, cuando la proporción de faltantes es pequeña comparada con el tamaño muestral, es decir, $m/N \to 0$, el teorema de Slutsky garantiza la convergencia en distribución a la misma distribución límite (Serfling 1980). A continuación mostramos el porqué.

En la sección anterior se mostró $\chi^2_c > \chi^2$ para $m > 0$. Observe que la expresión $\chi^2_c > \chi^2$ es equivalente a $\chi^2_c = \chi^2 + k$, con k una constante que depende de m. Si el número de datos parcialmente observados es pequeño en relación con el tamaño muestral, entonces $m \to 0$, $m/N \to 0$ y $k \to 0$. Dado que $\chi^2 \xrightarrow{d} \chi^2_1$, se tiene entonces que $\chi^2_c \xrightarrow{d} \chi^2_1$ por el teorema de Slutsky.

Tabla 2: *resultados de la prueba Kolomogorov-Smirnov; N el número total de observaciones, m es el número de observaciones parcialmente faltantes, D el estadístico de prueba y p el valor p. Fuente: elaboración propia.*

N	χ^2_c	m	χ^2	D	p
100	0.227	10	0.215	0.033	0.648
		15	0.210	0.047	0.219
		20	0.204	0.056	0.087
200	0.454	10	0.442	0.014	1.000
		15	0.437	0.019	0.994
		20	0.431	0.020	0.988
300	0.680	10	0.669	0.012	1.000
		15	0.663	0.012	1.000
		20	0.658	0.017	0.999
500	1.134	10	1.122	0.011	1.000
		15	1.117	0.015	1.000
		20	1.111	0.013	1.000

Como ilustración de este resultado y el presentado en §3.1, se generaron $B = 1000$ tablas de contingencia de 2×2 con vector de probabilidades $\boldsymbol{\pi} = (0.1, 0.2, 0.2, 0.5)$, $m = \{10, 15, 20\}$ y $N = \{100, 200, 300\}$ bajo un esquema de muestreo multinomial, y se calcularon los estadísticos χ^2_c y χ^2. Finalmente se realizó la prueba de Kolmogorov-Smirnov para evaluar si la distribución de probabilidad límite de los

estadísticos es equivalente[1]. Los resultados obtenidos se presentan en la Tabla 2.

Para N fijo, el estadístico χ^2_c no varía, mientras el estadístico χ^2 disminuye a medida que m aumenta. En la práctica, esto podría implicar no detectar independencia al utilizar el estadístico χ^2 puesto que no se tiene en cuenta la información extra (i.e, la parcialmente faltante). Por otro lado, $p \to 1$ cuando $m <<< N$ y $N \to \infty$, es decir, $\chi^2_c \xrightarrow{d} \chi^2_1$. Sin embargo, una proporción de valores parcialmente faltantes relativamente alta comparada con N puede no garantizar la convergencia del estadístico para valores de α relativamente altos (i.e., $N = 100, m = 20, m/N = 0.2, p = 0.087$).

4. Ejemplo

4.1. Datos completos

La onicofagia, o hábito compulsivo de comerse las uñas, afecta a entre el 28 % y el 33 % de niños entre 7 y 10 años, y al 45 % de los adolescentes (Leung & Robson 1990). A continuación presentamos el número de personas que muerden objetos, por género, entre estudiantes que no sufren onicofagia (Andersen 1997):

Tabla 3: *número de personas que muerden objetos en una muestra de 40 individuos. Fuente: Andersen (1997).*

	¿Muerde objetos?		
Sexo	Faltantes	Sí	No
Masculino	7	10	12
Femenino	2	3	6

Observe que al considerar solo la información completa, el número de individuos incluídos en la prueba de independencia χ^2 entre género y morder objetos es $n = 31$, y que este número incrementa a $n = 40$ cuando se incluye la información faltante. Adicionalmente, la proporción de datos faltantes es 22.5 %, un número relativamente alto para el total de individuos en la muestra. Como se mostrará en la sección §4.3, esta alta proporción de datos faltantes no garantiza los resultados presentados en §3, i.e., el estadístico χ^2_c no es siempre mayor que el estadístico χ^2. El valor del estadístico χ^2 al considerar los datos faltantes es 0.4348, mientras que al considerar sólo la información completa este se reduce a 0.3854 (ver Apéndice A). A pesar de que ninguno de los dos valores permite rechazar la hipótesis de independencia entre género y morder objetos, el incluir los datos adicionales en la prueba incrementa el estadístico χ^2 en ≈ 12 %.

[1]El programa en R se encuentra disponible bajo solicitud expresa del lector.

4.2. Cambio en el número de observaciones faltantes

¿Qué resultados se obtendrían si el número de datos faltantes por género cambiara? Y ¿por moder objetos? Cómo afectaría este cambio el estadístico χ^2? En la Figura 1 presentamos dichos resultados.

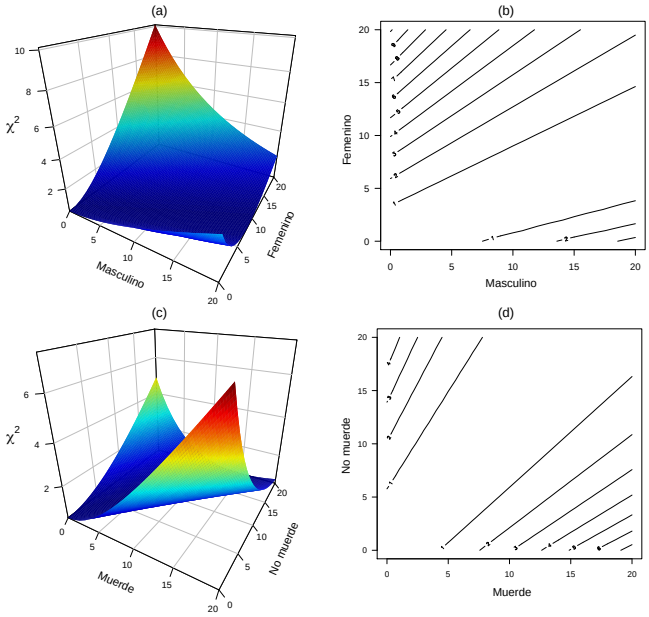

Figura 1: *representación 3D y curvas de nivel para el estadístico χ^2 obtenido con los datos presentados en la Tabla 3, en función del número de datos parcialmente faltantes por género (a, b), y si muerde o no muerde objetos (c, d). Valores del estadístico $\chi^2 > 3.841$ indican que la asociación entre género y morder objetos es estadísticamente significativa. Fuente: elaboración propia.*

Cuando el número de datos faltantes en la categoría *femenino* es diez o más, y simultáneamente el número de datos faltantes en la categoría *masculino* es superior a uno, la asociación entre género y onicofagia es estadísticamente significativa para una probabilidad de error tipo I del 5 % puesto que $\chi^2 > 3.841$ (ver esquina superior izquierda en la Figura 1(a) y 1(b) para más información). Por otro lado, la asociación es nuevamente estadísticamente significativa cuando el número de datos faltantes en la categoría *no muerde objetos* es inferior a tres y el número de datos faltantes en la categoría *muerde* es mayor a 13. Un resultado similar se obtiene cuando el número de datos faltantes en la categoría *no muerde* objetos es superior a 13 y el número de datos faltantes en la categoría *muerde objetos* es inferior a 3 (ver esquina superior izquierda en la Figura 1(c) y 1(d) para más información).

4.3. Experimento de simulación

Utilizando un esquema de muestreo multinomial y a partir de la Tabla 3, se realizó un experimento de simulación con el objetivo de evaluar el comportamiento del estadístico χ^2 cuando se incluyen o no los datos parcialmente observados. En este experimento, se generaron $B = 10000$ tablas de contingencia y a cada una se le aplicó la función `PruebaChi2FaltantesTabla` del Apéndice A.

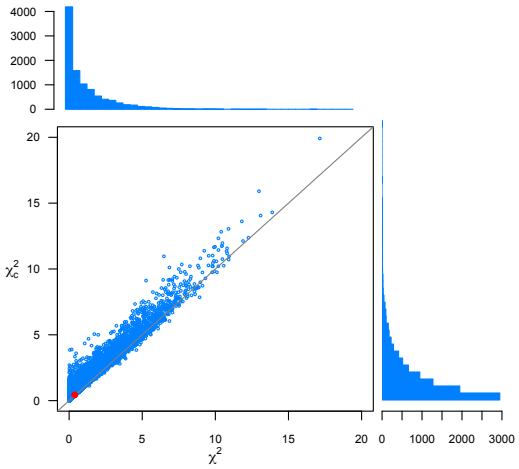

Figura 2: *comportamiento del estadístico χ^2 cuando se incluyen o no datos parcialmente observados en la prueba de indepedencia para la Tabla 3. El punto de color rojo representa los estadísticos $\chi^2 = 0.3854$ y $\chi^2_c = 0.4348$ calculados en §4.1. Fuente: elaboración propia.*

Los resultados obtenidos se presentan en la Figura 2. Debido a la alta proporción de datos faltantes (i.e., $9/40 = 22.5\,\%$), sólo en el $88.83\,\%$ de las tablas de contingencia generadas, $\chi^2_c > \chi^2$. Adicionalmente, la prueba χ^2 clásica sólo rechazó la hipótesis de independencia entre género y morder objetos en $8.92\,\%$ de los casos, frente al $11.04\,\%$ obtenido cuando se incluyó la información parcialmente faltante. Al comparar estos porcentajes utilizando una prueba de igualdad de proporciones se obtiene que la diferencia es estadísticamente significativa al $5\,\%$ ($\chi^2_1 = 24.778, p < 10^{-5}$).

5. Conclusiones

Hemos descrito con detalle el efecto que tiene en la prueba de independencia χ^2 el no considerar la información parcialmente observada cuando se dispone de esta. Por lo tanto, el analista debe verificar si en problemas concretos debe realizar la corrección que le incluya toda la información disponible y no usar indiscriminadamente los resultados que se obtienen automáticamente mediante el uso de paquetes

estadísticos. En los Apéndices A y B proporcionamos código en R para realizar la prueba de independencia en presencia de valores parcialmente faltantes.

Agradecimientos

Los autores agradecen los comentarios y sugerencias de un revisor anónimo. JIV agradece el apoyo incondicional de la señorita Yolima Espinosa Jaramillo, y la financiación por parte de The Eccles Scholarship in Medical Sciences, The Fenner Merit Scholarship y The Australian National University (ANU) High Degree Research Scholarship.

Recibido: 25 de abril de 2014
Aceptado: 20 de julio de 2014

Referencias

Agresti, A. (1990), *Categorical Data Analysis*, New York: Wiley.

Andersen, E. B. (1997), *Introduction to the Statistical Analysis of Categorical Data*, Springer-Verlag: Berlin.

Grizzle, J. E., Starmer, C. F. & Koch, G. G. (1969), 'Analysis of categorical data by linear models', *Biometrics* **25**(3), 489–504.

Haberman, S. (1972), 'Log-linear fit for contingency tables—Algorithm AS51', *Applied Statistics* **21**, 218–225.

Jeansonne, A. (2014), 'Loglinear Models '. Consultado Marzo 25, 2014. URL = `http://goo.gl/e0Y3aE`.

Leung, A. & Robson, W. (1990), 'Nailbiting', *Clin Pediatr (Phila)* **29**(12), 690–2.

Little, R. & Rubin, D. (2002), *Statistical Analysis With Missing Data*, 2nd edn, New York: Wiley.

McCullagh, P. & Nelder, J. A. (1983), *Generalized Linear Models*, Chapman & Hall, London.

McHugh, M. L. (2013), 'The χ^2 test of independence', *Biochemia Medica* **23**(2), 143–9.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL = `http://www.R-project.org/`.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley: New York.

Vélez, J. I. & Correa, J. C. (2013), 'Comparación de procedimientos FDR para la selección de parámetros en Regresión Poisson', *Comunicaciones en Estadística* **6**(1), 45–57.

Zeileis, A., Kleiber, C. & Jackman, S. (2008), 'Regression Models for Count Data in R', *Journal of Statistical Software* **27**(8), 1–25.

A. Prueba χ^2 a partir de tablas de contingencia

```
## descargar funciones
if(!require(devtools)) install.packages("devtools")
```

```
## Warning: package 'devtools' was built under R version 3.1.1
```

```
devtools:::source_url("http://bit.ly/1rzX8JY")
```

```
## Ejemplo onicofagia (ver Tabla 2)
tabla <- matrix(c(10, 12, 3, 6), ncol = 2, byrow = TRUE)
colnames(tabla) <- c('Si','No')
rownames(tabla) <- c('Masculino','Femenino')

# datos faltantes
falta.col <- c(7, 2)   # faltantes "Masculino" y "Femenino"
falta.fil <- c(0, 0)   # faltantes "Si", "No"

#  prueba (ver http://bit.ly/1rzX8JY para los argumentos)
PruebaChi2FaltantesTabla(tabla, falta.col, falta.fil)
```

```
##        X2         p       X2c       p_c
## 0.3853924 0.5347314 0.4347721 0.5096566
```

B. Prueba χ^2 a partir de dos variables categóricas

```
## descargar funciones
if(!require(devtools)) install.packages("devtools")
devtools:::source_url("http://bit.ly/1rzX8JY")

## Ejemplo con variables categoricas
set.seed(1)
x <- sample(c(1:3, NA), 1000, prob = c(.4, .3, .2, .1), replace = TRUE)
```

```
y <- sample(c(1:3, NA), 1000, prob = c(.4, .3, .25, .05), replace = TRUE)

#  prueba (ver http://bit.ly/1rzX8JY para los argumentos)
PruebaChi2FaltantesVariables(x, y)
```

```
##        X2         p       X2c       p_c
## 1.3471043 0.8533344 1.7827076 0.7756445
```