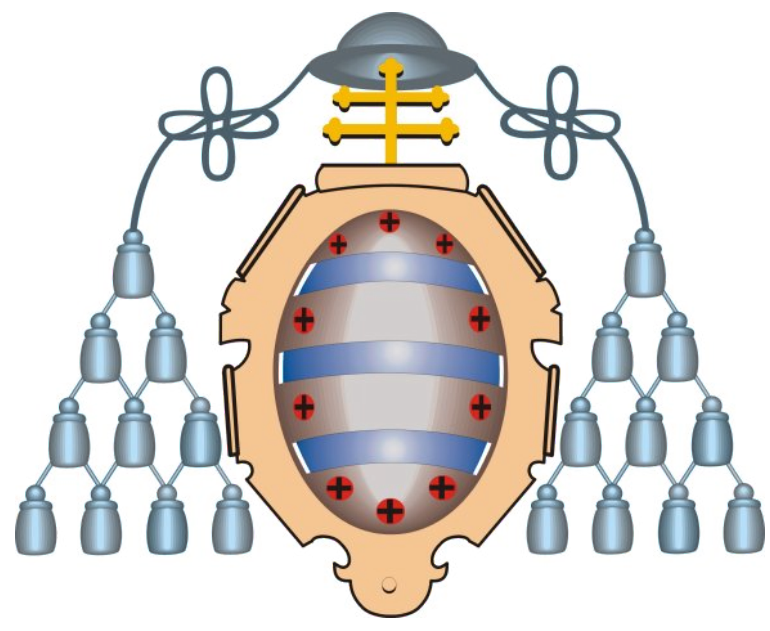


ESTADÍGRAFO DE CRAMÉR Y VON MISES PARA MUESTRAS APAREADAS



Pablo MARTÍNEZ CAMBLOR
CARLOS CARLEOS ARTIME
NORBERTO CORRAL BLANCO

DEPARTAMENTO DE ESTADÍSTICA
E INVESTIGACIÓN OPERATIVA
UNIVERSIDAD OVIEDO



RESUMEN

SE PROPONE UN CONTRASTE NO PARAMÉTRICO PARA COMPROBAR LA IGUALDAD DE LAS DISTRIBUCIONES MARGINALES DE UNA VARIABLE ALEATORIA PLURIDIMENSIONAL. EL ESTADÍGRAFO EMPLEADO PARA EL CONTRASTE APLICA EL CRITERIO DE CRAMÉR Y VON MISES A LAS FUNCIONES DE DISTRIBUCIÓN INVOLUCRADAS. SE CALCULA LA DISTRIBUCIÓN ASINTÓTICA DE DICHO ESTADÍGRAFO Y, MEDIANTE SIMULACIÓN, SE COMPARA CON EL COMPORTAMIENTO DE TÉCNICAS DE REMUESTREO APLICADAS AL MISMO. SE ESTUDIA LA POTENCIA DEL CONTRASTE PROPUESTO FRENTE AL REFERENTE CLÁSICO, EL CONTRASTE DE FRIEDMAN. CUANDO LA DIFERENCIA ENTRE LAS DISTRIBUCIONES AFECTA A LA TENDENCIA CENTRAL, AMBOS CONTRASTES PRESENTAN POTENCIAS SIMILARES. CUANDO LA DIFERENCIA CONSISTE SÓLO EN LA DISPERSIÓN O LA FORMA, EL NUEVO CONTRASTE RESULTA NOTABLEMENTE SUPERIOR AL DE FRIEDMAN.

INTRODUCCIÓN

Igualdad de marginales

SEA $\mathbf{X} = (X_1, \dots, X_k)$ UNA VARIABLE CONTINUA k -DIMENSIONAL. PARA $i \in \{1, \dots, k\}$ SEA F_i LA FUNCIÓN DE DISTRIBUCIÓN DE X_i . INTERESA CONTRASTAR

$$H_0: F_i = F_j \quad \forall i, j$$

$$H_1: F_i \neq F_j \quad \exists i, j$$

EL ESTADÍGRAFO CLÁSICO NO PARAMÉTRICO PARA RESOLVER DICHO CONTRASTE ES EL DE FRIEDMAN (1937). OTROS MÉTODOS HAN SIDO PROPUESTOS POR PURI Y SEN (1972); HOLLANDER, PLEDGER Y LIN (1974); RAVIV (1978); MACH Y SKILLINGS (1980); LAM Y LONGNECKER (1983); STELAND (1997). EN GENERAL, TAMBIÉN EMPLEAN ESTADÍGRAFOS BASADOS EN RANGOS, POR LO QUE FUNDAMENTALMENTE DETECTAN DIFERENCIAS EN POSICIÓN. FREITAQ, CZADO Y MUNK (2006) PRESENTAN UN MÉTODO PARA CONTRASTAR **similitud**, MÁS QUE IGUALDAD.

CRAMÉR Y VON MISES

EL CRITERIO DE CRAMÉR Y VON MISES SIRVE, EN PRINCIPIO, PARA CONTRASTAR LA BONDAD DE AJUSTE DE UNA FUNCIÓN DE DISTRIBUCIÓN EMPÍRICA (FDE) \hat{F}_n A UNA FUNCIÓN DE DISTRIBUCIÓN TEÓRICA F_0 :

$$W^2 = \int (\hat{F}_n(t) - F_0(t))^2 dF_0(t)$$

LO QUE KIEFER (1959) GENERALIZÓ PARA k MUESTRAS INDEPENDIENTES ASÍ:

$$W_k^2 = \sum_{i=1}^k n_i \int (\hat{F}_{n_i}(t) - \hat{F}_n(t))^2 d\hat{F}_n(t)$$

donde $n = \sum_{i=1}^k n_i$ y $\hat{F}_n(t)$ ES LA FDE PARA LA MUESTRA CONJUNTA.

MÉTODOS

SEA $\mathbf{X} = \{X_1, \dots, X_k\}$ UNA MUESTRA k -DIMENSIONAL DE TAMAÑO n , CON $X_i = \{x_{i1}, \dots, x_{in}\}$ PARA $i \in \{1, \dots, k\}$. SEA EL ESTADÍGRAFO

$$W_k^2(n) = \sum_{i=1}^k n \int \{\hat{F}_{n,i}(X_i, t) - \hat{F}_{n,\bullet}(\mathbf{X}, t)\}^2 d\hat{F}_{n,\bullet}(\mathbf{X}, t)$$

QUE ES LA ADAPTACIÓN DE W_k^2 PARA MUESTRAS APAREADAS, DONDE $\hat{F}_{n,i}(X_i, t)$ ($1 \leq i \leq k$) ES LA FDE DE LA MUESTRA i -ÉSIMA Y $\hat{F}_{n,\bullet}(\mathbf{X}, t) = k^{-1} \sum_{i=1}^k \hat{F}_{n,i}(X_i, t)$.

Distribución asintótica

EL DESARROLLO DE LA DISTRIBUCIÓN ASINTÓTICA SE BASA EN EL PRINCIPIO DE INVARIANCIA DE DONSKEER Y EN TEORÍA CLÁSICA SOBRE PROCESOS GAUSIANOS. EN CONCRETO, USAMOS LA DESCOMPOSICIÓN DE KARHUNEN Y LOÈVE PARA GARANTIZAR LA EXISTENCIA DE LOS COEFICIENTES (AUTOVALORES) Y VARIABLES NECESARIOS. EL RESULTADO PRINCIPAL ES QUE, SI H_0 ES CIERTA, ENTONCES SE DA LA CONVERGENCIA DÉBIL

$$W_k^2(n) \xrightarrow[n]{\mathcal{L}} \sum_{i=1}^k \sum_{l \in \mathbb{N}} \lambda_{i,l} M_{i,l}^2$$

donde $\{\mathbf{M}_l = (M_{1,l}, \dots, M_{k,l})\}_{l \in \mathbb{N}}$ ES UNA SECUENCIA DE VARIABLES ALEATORIAS k -DIMENSIONALES GAUSIANAS, CUYAS MARGINALES SIGUEN UNA DISTRIBUCIÓN $\mathcal{N}(0, 1)$, Y $\{\{\lambda_{i,l}\}_{l=1}^k\}_{l \in \mathbb{N}}$ SON CONSTANTES NO NEGATIVAS QUE CUMPLEN $\sum_{l \in \mathbb{N}} \lambda_{i,l}^2 < \infty$ PARA $1 \leq i \leq k$.

LA EXPRESIÓN COMPLETA DE LA DISTRIBUCIÓN ES ENREVESADA. ADEMÁS, EL USO PRÁCTICO DE LA FÓRMULA REQUIERE LA ESTIMACIÓN DE COVARIANZAS A PARTIR DE LA MUESTRA Y LA SUSTITUCIÓN DE LA COLECCIÓN DE AUTOVALORES POR UNA APROXIMACIÓN AL AUTOVALOR MAYOR.

REMUESTREO

PARA OBTENER UNA ESTIMACIÓN DE LA DISTRIBUCIÓN DEL ESTADÍGRAFO BAJO H_0 SIN RECURRIR A LA DISTRIBUCIÓN ASINTÓTICA, SURGE CON NATURALIDAD LA IDEA DE USAR UNA ALEATORIZACIÓN MEDIANTE PERMUTAS. TAL PROCEDIMIENTO REQUIERE PARA SU JUSTIFICACIÓN LA **INTERCAMBIABILIDAD** ENTRE LOS k COMPONENTES DE \mathbf{X} , ES DECIR, DADOS $i \neq j$ Y $k \neq l$, LAS DISTRIBUCIONES DE (X_i, X_j) Y DE (X_k, X_l) DEBEN COINCIDIR (**hipótesis de esfericidad**). NO ES GRAVE SI $k = 2$, PERO SÍ CON $k > 2$.

PRETENDEMOS SOSLAYAR LA NECESIDAD DE INTERCAMBIABILIDAD MEDIANTE EL USO DE ALGUNA VARIANTE DEL MUESTREO AUTOSUFICIENTE O BÚSTRAP. BAJO H_0 , DONDE CADA F_i SERÍA IGUAL A UNA F COMÚN, EL ESTADÍGRAFO SE PUEDE EXPRESAR COMO:

$$W_k^2(n) \stackrel{H_0}{=} \sum_{i=1}^k n \int \{\hat{F}_{n,i}(X_i, t) - F(t)\}^2 d\hat{F}_{n,\bullet}(\mathbf{X}, t) + \\ - nk \int \{\hat{F}_{n,\bullet}(\mathbf{X}, t) - F(t)\}^2 d\hat{F}_{n,\bullet}(\mathbf{X}, t)$$

Considérese ahora que $\mathbf{X}^* = \{X_1^*, \dots, X_k^*\}$ ES UNA MUESTRA ALEATORIA DE LA FDE MULTIVARIANTE $\hat{F}_n(\mathbf{X}, t)$. PARA $i \in 1, \dots, k$, SE DEFINE $\hat{F}_{n,i}^*(X_i^*, t)$ COMO LA FDE REFERIDA A X_i^* Y $\hat{F}_{n,\bullet}^*(\mathbf{X}^*, t) = k^{-1} \sum_{i=1}^k \hat{F}_{n,i}^*(X_i^*, t)$. SEA EL ESTADÍGRAFO

$$W_k^{2,*}(n) = \sum_{i=1}^k n \int \{\hat{F}_{n,i}^*(X_i^*, t) - \hat{F}_{n,i}(X_i, t)\}^2 d\hat{F}_{n,\bullet}^*(\mathbf{X}^*, t) + \\ - nk \int \{\hat{F}_{n,\bullet}^*(\mathbf{X}^*, t) - \hat{F}_{n,\bullet}(\mathbf{X}, t)\}^2 d\hat{F}_{n,\bullet}^*(\mathbf{X}^*, t)$$

PROPONEMOS EL SIGUIENTE ESQUEMA DE REMUESTREO:

1. DE LA MUESTRA ORIGINAL \mathbf{X} , CALCÚLESE $W_k^2(n)$.
2. DE LA FDE MULTIVARIANTE, $\hat{F}_n(\mathbf{X}, t)$, EXTRAÍGANSE B MUESTRAS INDEPENDIENTES DE TAMAÑO n : $\mathbf{X}^{*,b} = \{X_1^*, \dots, X_k^*\}$, $1 \leq b \leq B$.
3. PARA $b \in \{1, \dots, B\}$ CALCÚLESE $W_k^{2,*b}(n)$, A PARTIR DE $\mathbf{X}^{*,b}$.
4. LA DISTRIBUCIÓN DE $W_k^2(n)$ SE APROXIMA MEDIANTE $\{W_{k,*,1}^2(n), \dots, W_{k,*,B}^2(n)\}$, ES DECIR, EL P -VALOR ESTIMADO ES

$$\hat{p} = \frac{1 + \sum_{b=1}^B \chi\{W_{k,*,b}^2(n) \geq W_k^2(n)\}}{1 + B}$$

EN UN PLANTEAMIENTO **bústrap clásico**, SE REMUESTREARÍA A PARTIR DE UNA MUESTRA QUE REPRESENTASE LA HIPÓTESIS NULA. AQUÍ, SIN EMBARGO, USAMOS LA HIPÓTESIS NULA PARA OBTENER UNA NUEVA EXPRESIÓN DEL ESTADÍGRAFO. POR LA DIFERENCIA DE ENFOQUE, NOS REFERIREMOS A ESTE MÉTODO COMO **NORBÚSTRAP**.

RESULTADOS

ESTUDIAMOS LA POTENCIA, CON $\alpha = 0,05$, MEDIANTE 10 000 MUESTRAS GENERADAS A PARTIR DE VARIAS DISTRIBUCIONES. SE COMPARA EL ESTADÍGRAFO DE FRIEDMAN CON EL DE CRAMÉR Y VON MISES; PARA ÉSTE SE CONSIDERAN LA APROXIMACIÓN ASINTÓTICA (C_A) Y LA APROXIMACIÓN NORBÚSTRAP (C_B) CON $B = 499$.

SEA $\mathbf{Z} = (Z_1, Z_2, Z_3) \rightsquigarrow \mathcal{N}_3(\mathbf{0}, \Sigma)$, DONDE $\mathbf{0} = (0, 0, 0)$ Y DONDE $\Sigma = [\sigma_{i,j}]$ ES TAL QUE $\sigma_{i,j} = 1$ SI $i = j$ Y $\sigma_{1,2} = \sigma_{1,3} = 1/4$ Y $\sigma_{2,3} = b$.

SE EXTRAEN MUESTRAS DE TAMAÑO $n = 50$ DE LA VARIABLE TRIDIMENSIONAL $\mathbf{X} = (X_1, X_2, X_3)$ SEGÚN LOS MODELOS:

- M1.** $X_1 \equiv Z_1, X_2 \equiv Z_2, X_3 \equiv 1/4 * Z_3 + 3/4 * 3Z_3$.
- M2.** $X_1 \equiv Z_1, X_2 \equiv Z_2, X_3 \equiv 1/4 * Z_3 + 3/4 * (Z_3 + 1)$.
- M3.** $X_1 \equiv Z_1, X_2 \equiv Z_2, X_3 \equiv 1/4 * Z_3 + 3/4 * (\sqrt{3}Z_3 + 1)$.

donde la notación $\beta * \xi + \gamma * \zeta$ INDICA EXTRAER DE ξ CON PROBABILIDAD IGUAL A β Y EXTRAER DE ζ CON $\text{Pr} = \gamma$.

	b = 1/4			b = 3/4		
	C _B	C _A	Fr	C _B	C _A	Fr
M1	0'713	0'699	0'050	0'804	0'778	0'049
M2	0'980	0'980	0'938	1'000	1'000	1'000
M3	0'874	0'874	0'688	0'985	0'980	0'877

AGRADECIMIENTOS

ESTE TRABAJO HA SIDO FINANCIADO EN PARTE POR EL PROYECTO MTM2008-01519 DEL MINISTERIO DE CIENCIA E INNOVACIÓN.