

DISTRIBUCIÓN NORMAL MULTIVARIANTE

*Responsable: Mgr. Víctor Echegaray Munenaka
Miembro :Lic. Fernando Alvarez Carrillo*

RESUMEN

Los métodos estadísticos multivariados pueden ser de carácter exploratorio o inferencial. Los métodos inferenciales tienen como soporte a la distribución normal multivariante, la cual queda caracterizada por el vector de medias μ y la matriz de Varianzas y Covarianzas V .

La distribución T^2 de Hotelling es utilizada para probar hipótesis sobre el vector de medias y combinaciones lineales de ellos. La distribución de Wishart es usada para probar hipótesis referente a la matriz de Varianzas y Covarianzas V .

Algunos test especialmente importantes son el test de esfericidad y el test M de Box.

ABSTRACT

The multivariate statistical methods may be exploratory or inferential, the inferential methods take as a support multivariate normal distribution, characterized by multivariate vector μ mean and variance and covariance matrix V .

The T^2 distribution by Hotelling is used for testing hypotheses about average vector and linear combinations of them, the distribution of Wishart is used to test hypotheses relating to variance and covariance matrix V .

Some test specially important they are the test of esfericidad and the test M of Box.

I. INTRODUCCIÓN

La medición de varias características de una misma unidad experimental, ya sea en forma simultánea o con cierto intervalo de tiempo, genera una serie de datos que deben ser analizados con técnicas multivariantes.

En el caso multivariado, así como en el univariado, se disponen de ciertos estadísticos que permiten caracterizar una población y estudiar su comportamiento probabilístico, siendo la distribución normal la más importante de la teoría estadística.

El presente trabajo tratará específicamente sobre la distribución normal Multivariante, la cual es utilizada como modelo estadístico para explicar la estructura probabilística de las distribuciones multivariantes, pudiéndose realizar inferencias sobre los parámetros. Esto permitirá encontrar significación estadística, pues será posible asociar una medida de confiabilidad.

Nuestro objetivo es presentar un conjunto de conceptos necesarios para la caracterización de la distribución normal Multivariante, identificando sus parámetros, así como los criterios para la estimación y prueba de hipótesis referentes al vector de medias y la matriz de varianzas-covarianzas.

II. EL UNIVERSO MULTIVARIADO Y SU ANÁLISIS

El Análisis Multivariante es el conjunto de métodos estadísticos cuya finalidad es analizar simultáneamente conjuntos de datos multivariantes en el sentido de que hay varias variables medidas para cada individuo u objeto estudiado. Su razón de ser radica en un mejor entendimiento del fenómeno objeto de estudio, obteniendo información que los métodos estadísticos univariantes y bivariantes son incapaces de conseguir.

1. Objetivos del Análisis Multivariante

Pueden sintetizarse en dos:

Proporcionar métodos cuya finalidad es el estudio conjunto de datos multivariantes que el análisis estadístico uni y bidimensional es incapaz de conseguir.

Ayudar al analista o investigador a tomar decisiones óptimas en el contexto en el que se encuentre teniendo en cuenta la información disponible por el conjunto de datos analizado

III. TIPOS DE TÉCNICAS MULTIVARIANTES

Se pueden clasificar en tres grandes grupos:

1. **Métodos de dependencia:** Suponen que las variables analizadas están divididas en dos grupos: *las variables dependientes y las variables independientes*. El objetivo de los métodos de dependencia consiste en determinar si el conjunto de variables independientes afecta al conjunto de variables dependientes y de qué forma.
2. **Métodos de interdependencia:** Estos métodos no distinguen entre variables dependientes e independientes y su objetivo consiste en identificar qué variables están relacionadas, cómo lo están y por qué.
3. **Métodos estructurales:** Suponen que las variables están divididas en dos grupos: el de las variables dependientes y el de las independientes. El objetivo de estos métodos es analizar, no sólo como las variables independientes afectan a las variables dependientes, sino también cómo están relacionadas las variables de los dos grupos entre sí

IV. MATRIZ DE DATOS

Las técnicas estadísticas utilizan datos de conjuntos de varias variables medidas en múltiples individuos y estos datos se escriben en forma de tabla o matriz.

Sería impensable el desarrollo del Análisis Multivariante sin la ayuda de la Informática, pues dichos métodos procesan, en general, gran cantidad de datos.

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}$$

1. Vector de Promedios

Dada una matriz de datos al vector cuyos componentes son las medias muestrales de las variables individuales, se le llama vector de promedios.

$$\bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)$$

2. Matriz de Varianzas y Covarianzas

A partir de la matriz de datos, X, se calcula la matriz de varianzas y covarianzas, S

$$S = \begin{pmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{21} & S_{22} & \dots & S_{2p} \\ \dots & \dots & \dots & \dots \\ S_{p1} & S_{n2} & \dots & S_{np} \end{pmatrix}$$

Matriz de Correlación

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{n2} & \dots & r_{np} \end{pmatrix}$$

Varianza generalizada

Dada una matriz de varianzas y covarianzas, se denomina *varianza generalizada* al determinante de dicha matriz.

$$V = |S|$$

Variación Total

Dada una matriz de varianzas y covarianzas, se denomina *variación total* a la traza de la matriz S.

$$tr S = \sum S_{(jj)}$$

V. DISTRIBUCIÓN NORMAL MULTIVARIANTE

Se dice que un vector X de dimensión (p x 1) tiene distribución normal p-variada con vector de medias μ y matriz de varianzas-covarianzas V y se denotará:

$$X \sim N_p(\mu, V)$$

Si su función de densidad está dada por:

$$f(x) = \frac{1}{(2\pi)^{p/2} |V|^{1/2}} \exp\left\{-\frac{1}{2}(X-\mu)'V^{-1}(X-\mu)\right\}$$

$$x \in \mathbb{R}^p$$

VI. INFERENCIA PARA LA NORMAL MULTIVARIANTE

El objetivo es presentar una serie de procedimientos de inferencia para el vector de medias μ y la matriz de covarianzas V para la distribución normal multivariante.

1. Inferencia sobre la media:

Si V es conocida, entonces $n(\bar{x} - \mu)^T V^{-1} (\bar{x} - \mu)$ tiene una distribución χ^2 , con p grados de libertad. Podemos construir un elipsoide de confianza al $100(1 - \alpha)\%$ para μ , empleando la siguiente desigualdad:

$$n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu) \leq \chi_{p; 1-\alpha}^2$$

Si V es desconocida, sustituimos V por S_μ y utilizamos el resultado conocido como T^2 de Hotelling

$$T^2 = n(\bar{x} - \mu)^T S_\mu^{-1} (\bar{x} - \mu)$$

que nos permite calcular un elipsoide de confianza al $100(1 - \alpha)\%$ para μ .

Para el test $H_0: \mu = \mu_0$ pueden utilizarse los estadísticos

$$n(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu) \quad \text{ó}$$

$$n(\bar{x} - \mu)^T S_\mu^{-1} (\bar{x} - \mu)$$

dependiendo de si V se conoce o no. Para un nivel α el contraste de H_0 tiene por valores críticos $\chi_{p; 1-\alpha}^2$ y $(n-1)pF_{p, (n-p); 1-\alpha} / (n-p)$, respectivamente

2. Intervalos de confianza simultáneos para k combinaciones lineales específicas del vector de medias

Cuando el número k de combinaciones lineales es pequeño y conocidos los coeficientes antes de tomar la muestra es preferible utilizar los intervalos de confianza simultáneos dados por el método de Bonferroni.

Sean a_1, \dots, a_k los vectores de coeficientes, los intervalos de confianza son

$$a_i^T \mu : a_i^T \bar{x} \pm t_{1-(\alpha/2k)}; n-1 \times \sqrt{\frac{a_i^T S_{ii} a_i}{n}}$$

Test de hipótesis sobre relaciones lineales entre los componentes del vector de medias:

Este tipo de relaciones se puede contrastar en forma general construyendo un test de hipótesis $H_0: C\mu = 0$ frente a $H_1: C\mu \neq 0$, con C una matriz ($r \times p$) de rango $r <$

p . El estadístico del test será

$$T^2 = n(C\bar{x})^T (CS_\mu C^T)^{-1} (C\bar{x}) \sim T_{r; n-1}^2 \quad \text{bajo } H_0$$

Se rechaza H_0 si $T^2 > T_{r; n-1; 1-\alpha}^2$

Inferencia sobre la matriz de Covarianzas y de Correlación

Para la inferencia sobre V se utiliza el resultado conocido de que $X^T X$ tiene una distribución Wishart p dimensional $W_p(V, n)$ con n grados de libertad. La matriz nS tiene una distribución Wishart $W_p(V, n-1)$ con $(n-1)$ grados de libertad.

Test de Esfericidad

El Test estadístico utilizado para contrastar $H_0: V = \sigma^2 I$ viene dado por:

$$T = np \ln[A/G], \quad \text{con } A = \frac{1}{p} \sum l_i, \quad G = \left[\prod l_i \right]^{1/p}$$

Donde $l_1 = l_2 = \dots = l_p$ son los valores propios de S . Si H_0 es cierta, el estadístico T tiene una distribución χ^2 con $\frac{1}{2}(p-1)(p+2)$ grados de libertad, para tamaños muestrales grandes

Test para contrastar la igualdad entre Matrices de Covarianzas: test M de Box

Sean x_k $N_p(\mu_k, V_k)$, $k = 1, \dots, g$, donde g indica diferentes grupos. Para cada grupo se dispone de una muestra de tamaño n_k , $k = 1, \dots, g$. La hipótesis a contrastar es:

$$H_0: V_1 = \dots = V_g$$

y el estadístico que se utiliza es:

$$M = \frac{\prod_{k=1}^g |S_{u,k}|^{(n_k-1)/2}}{|S_{pl}|^{(n-g)/2}}$$

VII. CONCLUSIONES

1. La distribución normal multivariante desempeña un rol fundamental en la inferencia estadística multivariante. La distribución normal multivariante queda caracterizada por el vector de medias μ y el vector de varianzas y covarianzas V .

2. La distribución T^2 de Hotelling es utilizada para probar hipótesis referentes al vector de medias μ .

3. La distribución de Wishart es utilizada para probar hipótesis sobre la matriz de Varianzas y Covarianzas V .

4. La Varianza Generalizada y la Variación Total proporcionan información de la dispersión de los datos multivariantes.

VIII. BIBLIOGRAFÍA

Cramer, H: *Métodos matemáticos de estadística*. AGUILAR S.A. , Madrid, 1970.

Morrison. D. F: *Multivariate Statistical Methods*. Mc Graw Hills, 1990.

Mardia, K. , Kent, J. , Bibby, J. : *Multivariate Analysis*, Academic Press, 1982.

Hair, Anderson, Tatham, Black: *Análisis Multivariante*. 5^{ta} edc.

Giri, Narayan: *Multivariate Statistical*. Marcel Dekker, inc. New York, 1996.

Cuadra, C: *Métodos de Análisis Multivariante*, PBU, Barcelona, 1999.

Peña, Daniel: *Estadística, Modelos y métodos 1. Fundamentos*, Alianza Editorial, S.A., Madrid, 2000.

Peña, Daniel: *Análisis de datos multivariantes*, Mc Graw Hills Interamericana, 2002