Two-way analysis of variance for a concentrated von Mises-Fisher distribution

Figueiredo, Adelaide
Faculdade de Economia da Universidade do Porto and LIACC
Rua Dr. Roberto Frias
4200-464 Porto, Portugal
E-mail: adelaide@fep.up.pt

Abstract The von Mises-Fisher distribution is one of the most used distributions for vectorial data and it has many applications for spherical data. The one-way analysis of variance and a nested multiway layout technique for a concentrated von Mises-Fisher distribution have already been presented in the literature. In this paper, we extend the previous techniques to the two-way analysis of variance.

Introduction

The one-way analysis of variance technique for a concentrated von Mises-Fisher distribution was introduced by Watson (1956) and Watson and Williams (1956) and a nested multiway layout technique was proposed by Stephens (1982) (see also Stephens, 1992). In this paper we propose the two-way analysis of variance for a concentrated von Mises-Fisher distribution.

The von Mises-Fisher distribution defined on the unit sphere in R^p, S_{p-1} is usually denoted by $M_p(\mu, \xi)$ and it has probability density function given by

(1) $$f(\mathbf{x}|\mu, \xi) = c_p(\xi) \exp\{\xi\mu'\mathbf{x}\} \quad \mathbf{x} \in S_{p-1}, \quad \mu \in S_{p-1},\ \xi \geq 0\ ,$$

where $c_p(\xi)$ is the normalising constant given by

$$c_p(\xi) = \frac{\xi^{\frac{p-1}{2}}}{(2\pi)^{\frac{p}{2}} I_{\frac{p-1}{2}}(\xi)}$$

and I_ν denotes the modified Bessel function of the first kind and order ν defined by

$$I_\nu(k) = \frac{1}{2\pi}\int_0^{2\pi} \cos\nu\theta e^{k\cos\theta} d\theta.$$

The parameter μ is the mean direction and ξ is the concentration parameter around μ.
This distribution verifies the properties:

- It is rotationally symmetric about μ.
- If $\mathbf{x}$ comes from $M_p(\mu, \xi)$ and U is an orthogonal matrix, then $U\mathbf{x}$ comes from $M_p(U\mu, \xi)$.
- For $\mathbf{x} \in S_{p-1}$ from $M_p(\mu, \xi)$ then for large ξ

(2) $$2\xi(1 - x'\mu) \sim \chi^2_{p-1}.$$

Let $[\mathbf{x}_1|\mathbf{x}_2|...|\mathbf{x}_n]$ be a random sample of size n from the von Mises-Fisher distribution $M_p(\mu, \xi)$. Let $\overline{R}$ be the resultant length mean of the sample defined by $\overline{R} = \|\overline{\mathbf{x}}\| = (\overline{\mathbf{x}}'\overline{\mathbf{x}})^{\frac{1}{2}}$, where $\overline{\mathbf{x}}$ is the sample vector mean of $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ defined by $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$. Let R be the resultant length of the sample of size n, defined by $R = n\overline{R}$.

- The maximum likelihood estimator of μ is the sample mean direction, that is

$$\widehat{\mu} = \overline{\mathbf{x}}_0 = \frac{\overline{\mathbf{x}}}{\|\overline{\mathbf{x}}\|}.$$

- The maximum likelihood estimator of ξ is the solution of the following equation

$$A_p(\xi) = \|\overline{\mathbf{x}}\|,$$

where the function $A_p(\xi)$ is defined by

$$A_p(\xi) = \frac{c_p'(\xi)}{c_p(\xi)} = \frac{I_{\frac{p}{2}}(\xi)}{I_{\frac{p}{2}-1}(\xi)}.$$

(For more details about this distribution, see for instance, Fisher, Lewis and Embleton, 1987, Watson, 1983, Mardia and Jupp, 2000).

Two-way analysis of variance

We suppose that we have n observations classified according to a factor A with r treatments and a factor B with s treatments. When an observation is classified into treatment i of factor A and treatment j of factor B, this observation falls in cell (i,j) in row i and column j of a two-way table. We suppose that there is t observations in each cell of the table and let $n = rst$ be the total number of observations.

Let $\mathbf{x}_{ijk}$ be kth observation in cell (i,j), $i = 1,\ldots,r$, $j = 1,\ldots,s$, $k = 1,\ldots,t$. The observations in cell (i,j) are supposed to come from the subpopulation $M_p(\mu_{ij}, \xi_{ij})$, $i = 1,...,r$, $j = 1,...,s$. We suppose that the rs subpopulations are independent and we assume that $\xi_{11} = \xi_{12} = ... = \xi_{rs} = \xi$, where the common concentration ξ is unknown.

We want to test the null hypothesis: $H_0 : \mu_{11} = \mu_{12} = ... = \mu_{rs} = \mu$, against the alternative that at least one of the equalities is not satisfied.

Let

- R be the resultant length of all observations $\mathbf{x}_{ijk}$, $i = 1,...,r$, $j = 1,...,s$, $k = 1,...,t$.

- $R_{i.}$ be the resultant length of the observations in the row i, $\mathbf{x}_{ijk}$, $j = 1,...,s$, $k = 1,...,t$.

- $R_{.j}$ be the resultant length of the observations in the column j, $\mathbf{x}_{ijk}$, $i = 1,...,r$, $k = 1,...,t$.

- R_{ij} be the resultant length of the observations in the cell (i,j) of the two-way table, $\mathbf{x}_{ijk}$, $k = 1,...,t$.

The two-way analysis of variance is based on the following identity

$$\begin{aligned} 2\xi(n-R) &= 2\xi\left(\sum_{i=1}^{r} R_{i.} - R\right) + 2\xi\left(\sum_{j=1}^{s} R_{.j} - R\right) + \\ &\quad + 2\xi\left(\sum_{i=1}^{r}\sum_{j=1}^{s} R_{ij} - \sum_{i=1}^{r} R_{i.} - \sum_{j=1}^{s} R_{.j} + R\right) + 2\xi\left(n - \sum_{i=1}^{r}\sum_{j=1}^{s} R_{ij}\right), \end{aligned}$$

which is a decomposition of the total variance into between-rows variance, between-columns variance, interaction between the rows-columns and residual variance.

It can be proved that for large ξ

$$2\xi\left(\sum_{i=1}^{r} R_{i.} - R\right) \sim \chi^2_{(r-1)(p-1)}, \tag{3}$$

(4) $$2\xi \left(\sum_{j=1}^{s} R_{.j} - R \right) \sim \chi^2_{(s-1)(p-1)},$$

(5) $$2\xi \left(\sum_{i=1}^{r} \sum_{j=1}^{s} R_{ij} - \sum_{i=1}^{r} R_{i.} - \sum_{j=1}^{s} R_{.j} + R \right) \sim \chi^2_{(r-1)(s-1)(p-1)}.$$

(6) $$2\xi \left(n - \sum_{i=1}^{r} \sum_{j=1}^{s} R_{ij} \right) \sim \chi^2_{(n-rs)(p-1)}.$$

To test the null hypothesis $H_0^{'}$: There is no difference between rows, we use the statistic defined by

(7) $$F_1 = \frac{(n-rs)\left(\sum_{i=1}^{r} R_{i.} - R \right)}{(r-1)\left(n - \sum_{i=1}^{r} \sum_{j=1}^{s} R_{ij} \right)},$$

which has under $H_0^{'}$ and for large ξ, $F_{(r-1)(p-1),(n-rs)(p-1)}$ distribution. We reject $H_0^{'}$ for large values of F_1.

To test the null hypothesis $H_0^{''}$: There is no difference between columns, we use the following statistic

(8) $$F_2 = \frac{(n-rs)\left(\sum_{j=1}^{s} R_{.j} - R \right)}{(s-1)\left(n - \sum_{i=1}^{r} \sum_{j=1}^{s} R_{ij} \right)},$$

which has under $H_0^{''}$ and for large ξ, $F_{(s-1)(p-1),(n-rs)(p-1)}$ distribution. We reject $H_0^{''}$ for large values of F_2.

To test the null hypothesis $H_0^{'''}$: There is no interaction between the rows and columns, we use the following statistic

(9) $$F_3 = \frac{(n-rs)\left(\sum_{i=1}^{r} \sum_{j=1}^{s} R_{ij} - \sum_{i=1}^{r} R_{i.} - \sum_{j=1}^{s} R_{.j} + R \right)}{(r-1)(s-1)\left(n - \sum_{i=1}^{r} \sum_{j=1}^{s} R_{ij} \right)},$$

which has under $H_0^{'''}$ and for large ξ, $F_{(r-1)(s-1)(p-1),(n-rs)(p-1)}$ distribution. We reject $H_0^{'''}$ for large values of F_3.

REFERENCES (RÉFERENCES)

Fisher, N. I., Lewis, T. and B. J. J. Embleton (1987). *Statistical analysis of spherical data.* Cambridge University Press.

Mardia, K. V. and P. E. Jupp (2000). *Directional Statistics.* John Wiley and Sons, Chichester.

Stephens, M. A. (1982). Use of the von Mises distribution to analyse continuous proportions. *Biometrika* **69**, 1, 197-203.

Stephens, M. A. (1992). On Watson's ANOVA for directions. In Mardia K. V. (ed.)*Art of Statistical Science*, 75-85.

Watson, G. S. (1956). Analysis of dispersion on a sphere. *Mon. Not. R. Astr. Soc.:Geophys. Suppl.* **7**, 153-9.

Watson, G. S. and E. J. Williams (1956). On the construction of significance tests on the circle and the sphere. *Biometrika* **43**, 344-52.

Watson, G. S. (1983). *Statistics on spheres.* John Wiley and Sons, New York.