Frontiers in Educational Research
ISSN 2522-6398 Vol. 3, Issue 11: 147-157, DOI: 10.25236/FER.2020.031124

Application of Correspondence Analysis in Exploring the Statistical Characteristics of Uterine Fibroids and Age

Zhu Yanxin[1], Liu Qinghua[2], Zhu Ning[2,3*]

1. Changsha Maternal and Child Health Hospital, Changsha, Hunan, 410007, China
2. School of Mathematics and Computational Science, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China
3. Institute of Information Technology of Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China
**Corresponding Author*

ABSTRACT. *The paper uses statistical analysis methods to investigate the prevalence of 10 seeds of uterine fibroids from 21 to 75 years old in hospital A. The principal component analysis method is used to divide the 10 seeds of uterine fibroids disease into three types of comprehensive indicators. The corresponding analysis method is used to obtain the correspondence information between different age groups and the three types of comprehensive uterine fibroids diseases. The prevention and treatment should be targeted according to different uterine fibroids diseases at different ages. From the corresponding analysis chart and the difference results, it can be intuitively obtained that the high incidence of uterine fibroids disease is in the middle age (41-50 years old). The statistical analysis method in this paper provides theoretical support and relevant reference for related research in the medical field.*

KEYWORDS: *Uterine fibroids, Age distribution, Principal component analysis, Correspondce analysis, Difference test*

1. Introduction

Uterine fibroids are the most common benign tumors of female reproductive organs. They are mainly caused by the proliferation of immature uterine smooth muscle cells, which is the primary cause of hysterectomy in the clinic. Uterine fibroids are also common diseases and frequent diseases that affect women's reproductive health. It has a great adverse effect on the reproductive health of women, which puts a heavy economic burden on the patient. Therefore, it is of certain clinical significance to discuss the influencing factors and risk factors of uterine fibroids, which provides an important theoretical basis for effectively preventing and controlling the occurrence of uterine fibroids and improving the

ISSN 2522-6398 Vol. 3, Issue 11: 147-157, DOI: 10.25236/FER.2020.031124

detection of clinically asymptomatic uterine fibroids.

For the study of uterine fibroids, the commonly used methods are descriptive analysis, single factor analysis, multi-factor Logisic regression analysis[1] and so on. Lu Wei, Fan Bozhen (2008)[2] and others studied the etiology of uterine fibroids; Ye J and Wang H (2015)[3] and others analyzed the MED12 mutation in patients with uterine fibroids; Xu Ya, Chen Sidong Et al.(2000)[4] used a 1:2 ratio method, Yang Huiyun, Zhu Lujuan et al.(2004)[5]used a 1:1 ratio case-control study method to investigate the risk factors of multiple women with uterine fibroids; at present, principal component analysis and correspondence analysis methods are becoming powerful tools for exploratory research, which are widely used in the fields of biology, computer science, geological research and market research. Besides, it also uses in the analysis of mineral elements in medicinal materials[6], genome characteristics Natural selection[7], exploration of the relationship between the age of residents and heart disease[8] and research on the impact of mine dam accidents on environmental response[9]. At present, doctors mainly qualitatively analyze the relationship between uterine fibroids and age by experience. In this paper, 673 uterine fibroids patients of all ages in hospital A were modeled and analyzed and the principal component correspondence analysis model was established. By drawing the corresponding analysis chart, the relationship between uterine fibroids and age was quantitatively analyzed. Significance tests were conducted on the differences of the diseases in different age groups, which help obtain the main types of uterine fibroids disease occurred in each age group.

2. The Source and Preprocessing of Data

The data in the paper comes from 3801 cases counted by Hospital A. The original data contains only three indicators of gender, age and test results. There are 2445 cases that are recommended to be reviewed in the test results, among which 673 are diagnosed as uterine fibroids or suspected uterine fibroids. This paper attempts to explore the use of principal component correspondence analysis to analyze uterine fibroids patients in the recommended review cases and analyze the relationship between each uterine fibroids and age distribution.

Because the original data is text data, it is not convenient for quantitative analysis. Therefore, we carry out the dimensionless processing on the data. First of all, because all genders are women, this index is negligible; then 673 patients with uterine fibroids or patients with suspected uterine fibroids were screened; finally, the frequency the patients with uterine fibroids or patients with suspected uterine fibroids at various ages were obtained. The 673 patients recommended for re-examination were aged from 21 to 85 years, with an average age of 41.9 years. According to statistics, there are 205 patients among 637 patients were diagnosed with uterine fibroids, 21 were suspected of uterine fibroids. The diagnosis rate was 10.2%; 76 patients were diagnosed with small uterine fibroids and 7 patients were suspected of being uterine small uterine fibroids. The diagnosis rate was 8.4%; 42 patients were diagnosed with multiple small uterine fibroids, 1 was diagnosed with multiple suspected small uterine fibroids and the diagnosis rate was 2.3%; 304

ISSN 2522-6398 Vol. 3, Issue 11: 147-157, DOI: 10.25236/FER.2020.031124

patients were diagnosed with multiple uterine fibroids. 8 patients were diagnosed with subserosal uterine fibroids. 6 patients were diagnosed with submucosal uterine fibroids and 3 patients were diagnosed with cervical uterine fibroids. Among of people who were diagnosed, the diagnosis rate reached 100%.

Since the patients are from 21 to 75 years old, the age span is relatively large. If each age gets a frequency to analyze, this work is too complicated. Therefore, we use the general principles mentioned in “Probability Theory and Mathematical Statistics” written by Mao Shisong[10]. For the sample with a smaller volume, the number of groups is usually between 5-20 groups. The final number of groups is 11 groups and the group distance is 5 years old. Therefore, it can be obtained the frequency distribution table of tumor diseases of each uterine muscle in each age group.

Table 1 1 the Frequency Distribution of Patients with Uterine Fibroids At Various Ages

diagnostic result	Uterine fibroids	Uterine leiomyoma	Subserosal fibroids	Submucosal fibroids	Cervical fibroids
[21,25]	1	0	0	1	0
[26,30]	10	4	0	0	0
[31,35]	9	12	0	0	0
[36,40]	25	19	1	0	0
[41,45]	44	17	2	0	1
[46,50]	41	10	2	4	2
[51,55]	36	4	3	1	0
[56,60]	22	3	0	0	0
[60,65]	13	4	0	0	0
[66,70]	2	2	0	0	0
[71,75]	2	1	0	0	0
Total	205	76	8	6	3

Continued Table 1.1 The frequency distribution of patients with uterine fibroids by age

diagnostic result	Suspected uterine fibroids	Multiple uterine fibroids	Multiple small fibroids	Suspected uterine leiomyoma	Suspected multiple small fibroids
[21,25]	1	0	0	0	0
[26,30]	3	0	2	0	0
[31,35]	3	5	2	1	0
[36,40]	6	29	6	3	0
[41,45]	2	88	14	1	0
[46,50]	1	92	14	2	0
[51,55]	2	61	3	0	1
[56,60]	3	19	1	0	0

ISSN 2522-6398 Vol. 3, Issue 11: 147-157, DOI: 10.25236/FER.2020.031124

[60,65]	0	7	0	0	0
[66,70]	0	2	0	0	0
[71,75]	0	1	0	0	0
Total	21	304	42	7	1

In this paper, we use the 10 seed uterine fibroids diseases in Table 1.1 as indicators $x_i, i=1,2,\cdots,10$, use 11 age groups as the dependent variables $y_j, j=1,2,\cdots,11$, and conduct a principal component correspondence analysis on the data of 10 major uterine fibroids diseases in patients of various age groups in hospital A. In this way, we can explore the age distribution of these diseases.

3. Empirical Analysis

3.1 Normality Test

First, the normality of uterine fibroids data is studied. We make histograms of uterine fibroids diseases and various age groups. The frequency distribution maps of uterine fibroids are shown in Figure 2.1.

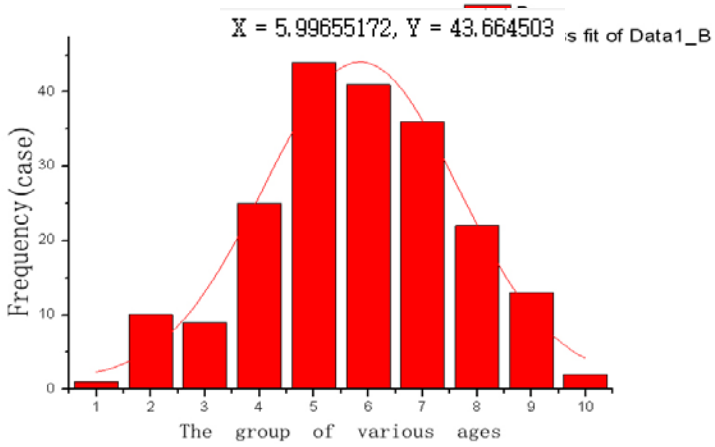

Fig.2 1 Histogram of Frequency of Uterine Fibroids in Various Ages

From Table 1.1, there are 205 cases diagnosed with uterine fibroids. From Figure 2.1, it shows that the frequency of uterine fibroids is relatively high at 36 to 60 years old. Combining Table 3.1, it can obtain that the frequency of uterine fibroids between 36 and 60 years old is 128, accounting for 81.95% of the total frequency of uterine fibroids. The probability of contracting uterine fibroids before the age of 35 is 9.76%, and the probability of contracting after 60 is 8.29%. From $R^2 = 0.96764$, it can obtain that the frequency of uterine fibroids in each age group follows a normal distribution. From the figure, we can also get a Gaussian fitting curve to fit

ISSN 2522-6398 Vol. 3, Issue 11: 147-157, DOI: 10.25236/FER.2020.031124

the frequency very well. Besides, it concentrates in the 5th, 6th and 7th groups. That is, the age is concentrated in the 41~50 years old. This shows that the high incidence of uterine fibroids is concentrated in middle-aged women aged 41 to 50 years.

3.2 Principal Component Analysis

To explore the statistical characteristics of uterine fibroids and age, there are 10 seeds of uterine fibroids as indicators, namely uterine fibroids (x_1), small uterine fibroids (x_2), subserosal fibroids (x_3), submucosal fibroids (x_4) , Cervical fibroids (x_5), suspected uterine fibroids (x_6), multiple uterine fibroids (x_7), multiple small uterine fibroids (x_8), suspected small uterine fibroids (x_9), suspected multiple small uterine fibroids (x_{10}). There are many survey indicators and some of the indicators have very small values. Therefore, the principal components of 10 indicators were analyzed to establish a comprehensive indicator of uterine fibroids disease.

Firstly, standardize the data to eliminate the dimensional relationship between variables, so that the data are comparable. Then, analyze the correlation of the standardized data. The results are shown in Figure 2.2.

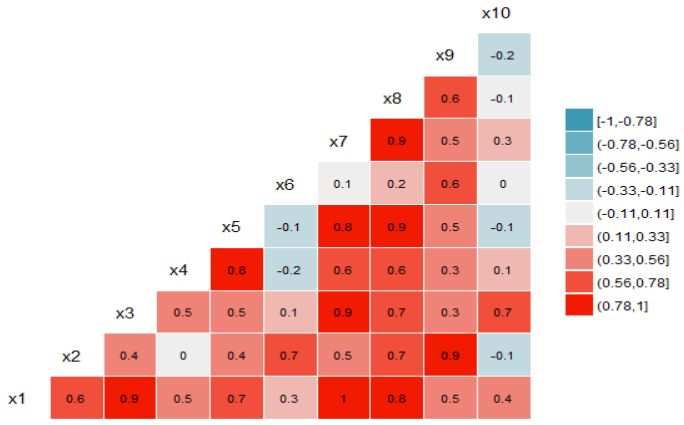

Fig.2 2 the Correlation Coefficient between Indicators.

The positive and negative signs in the figure indicate positive and negative correlations. The darker the color, the stronger the correlation between the two indicators. It can be seen from Figure 2.2 that most indicators show a strong positive correlation, some have a correlation coefficient as high as 0.9, and some indicators

ISSN 2522-6398 Vol. 3, Issue 11: 147-157, DOI: 10.25236/FER.2020.031124

have a small negative correlation. It shows that there is multicollinearity among these indexes. Therefore, on the premise of retaining the original index information, the principal component analysis is performed on the data to establish the unrelated principal component comprehensive index between the two indicators. In order to better explain each principal component, we perform orthogonal factor rotation on the principal component to construct a rotated principal component model. The eigenvalue and contribution rate between each impact indicator are calculated by SAS software. The results are shown in Table 2.1:

Table 2 1 Eigenvalues Of Correlation Coefficient Matrix

Main ingredient	Eigenvalues	Difference	Standard deviation	Contribution rate	Cumulative contribution rate
Comp1	3.91	1.53	1.98	0.39	0.39
Comp2	2.38	0.69	1.54	0.24	0.63
Comp3	1.69	0.08	1.30	0.17	0.80
Comp4	1.61	1.37 1.27	0.16	0.96	
Comp5	0.24	0.18	0.49	0.02	0.98
Comp6	0.11	0.05	0.33	0.01	0.99
Comp7	0.06	0.05	0.24	0.01	1.00
Comp8	0.01	0.01	0.10	0.00	1.00
Comp9	0.00	0.00	0.00	0.00	1.00
Comp10	0.00		0.00	0.00	1.00

From Table 2.1, the contribution rate of the first principal component (Comp1) reaches 55%, and the cumulative contribution rate of the first three principal components reaches 80%, indicating that the first three principal components contain the original 10 indicators 80% Information, so determine the number of principal components is 3. The load of uterine fibroids disease index on the main components and related information are shown in Table 2.2 as follows:

Table 2 2 the Load of Uterine Fibroids Disease Indicators on the Main Components

	Comp1	Comp2	Comp3 Common factor variance	Uniqueness	
x_1	0.65	0.44	0.54	0.92	0.08
x_2	0.29	0.92	0.02	0.93	0.07
x_3	0.56	0.24	0.78	0.98	0.02
x_4	0.85	-0.14	0.08	0.75	0.25
x_5	0.99	0.11	-0.05	0.99	0.01

ISSN 2522-6398 Vol. 3, Issue 11: 147-157, DOI: 10.25236/FER.2020.031124

x_6	-0.26	0.88	0.13	0.86	0.14
x_7	0.81	0.30	0.46	0.97	0.03
x_8	0.84	0.48	0.12	0.95	0.05
x_9	0.39	0.83	-0.12	0.85	0.15
x_{10}	-0.07	-0.16	0.96	0.96	0.04

The columns corresponding to Comp1, Comp2, and Comp3 are the loads of the three principal components on each observed variable, and reflect the correlation coefficient between the observed variable and the principal component. It can be seen from Table 2.2 that the first principal component is highly correlated to x_1, x_4, x_5, x_7, x_8, so Comp1 can be used as a comprehensive index of these five variables x_1, x_4, x_5, x_7, x_8. Similarly, Comp2 can be used as a comprehensive index, and Comp3 can be used as a comprehensive index. The variance of the common factor is the main component's explanation of the variance of each variable. Table 2.2 shows that a large part of the variance of the uterine fibroids disease index can be explained by these three principal components.

Table 2.3 shows the feature vectors corresponding to all feature roots, which are linearly independent vectors. The first column represents the score coefficient of the first principal component Comp1, the second column represents the score coefficient of the second principal component Comp2, and so on. The contribution rate of each principal component after Comp3 is very small, so it is not listed here. The results are shown in Table 2.3 as below.

Table 2 3 Feature Vector

	Comp1	Comp2	Comp3
x_1	0.08	0.08	0.20
x_2	-0.03	0.36	-0.05
x_3	0.04	0.00	0.35
x_4	0.29	-0.18	-0.08
x_5	0.32	-0.09	-0.18
x_6	-0.23	0.39	0.09
x_7	0.16	0.00	0.13
x_8	0.20	0.08	-0.07
x_9	0.03	0.29	-0.14
x_{10}	-0.15	-0.07	0.56

ISSN 2522-6398 Vol. 3, Issue 11: 147-157, DOI: 10.25236/FER.2020.031124

The expression of the three principal components can be obtained from Table 2.3.

$$\text{Comp1} = 0.08x_1 + 0.29x_4 + 0.32x_5 + 0.16x_7 + 0.20x_8 \quad ;$$

$$\text{Comp2} = 0.34x_2 + 0.39x_6 + 0.29x_9$$

$$\text{Comp3} = 0.35x_3 + 0.56x_{10}$$

3.3 Correspondence Analysis

Use the principal component index system established in Table 2.3 to observe the relationship between various uterine fibroids diseases and age, and use SAS software to analyze the principal component statistical data of patients with uterine fibroids. The results are shown in Table 2.4.

Table 2 4 Characteristic Inertia

Eigenvalues	Principal component inertia	Contribution rate	Cumulative contribution rate	Chi-square statistics	Degrees of freedom
0.595	0.354	0.94	0.94	40.74	20
0.146	0.021	0.06	1.00	2.47	

It can be seen from Table 2.4 that the number of dimensions is reduced to 2 dimensions. That is obtained by the way of the number of categories with the smaller number of categories in the two variables of row and column minus 1. The principal component inertia is equal to the square of the characteristic value of each factor, which is used to indicate the importance of each factor. The contribution rate of the first eigenvalue reaches 94%, which shows that the two-dimensional projection graph formed by the first factor axis already contains the information of 94% of the original vector. The statistics of the chi-square is 40.74+2.47=43.21, the degree of freedom is 20, the calculation of $P(\chi^2(20) = 43.21) = 1.92 \times 10^{-3} < 0.05$ can be obtained. That is, the original hypothesis is rejected. It can be considered that there is an inherent relationship between various uterine fibroids diseases and age distribution and the corresponding analysis can be performed. The results are shown in the figure. 2.3:

As can be seen from the corresponding analysis chart in Figure 2.3, Comp1 and E (from 41 to 45 years old), F (from 46 to 50 years old), H (from 56 to 60 years old), I (from 61 to 65 years old), K (from 71 to 75 years old) distributed in the first quadrant. It can be considered that Comp1 type uterine fibroids disease mainly occurs in these ages, and Comp1 is closest to E and F; similarly, the incidence of Comp2 type uterine fibroids disease is younger, mainly focused on the AD age group (21-40 years old); while the G age group (51-55 years old) women are susceptible to Comp3 uterine fibroids disease; the J age group (66-70 years old) deviates from other indicators and the sample, which shows that the probability of

ISSN 2522-6398 Vol. 3, Issue 11: 147-157, DOI: 10.25236/FER.2020.031124

uterine fibroids disease in this age group is not big, which is consistent with the data in Table 1.1. According to the results of the corresponding analysis, the corresponding analysis chart can clearly and truly reflect the concentration and age distribution of each uterine fibroids.

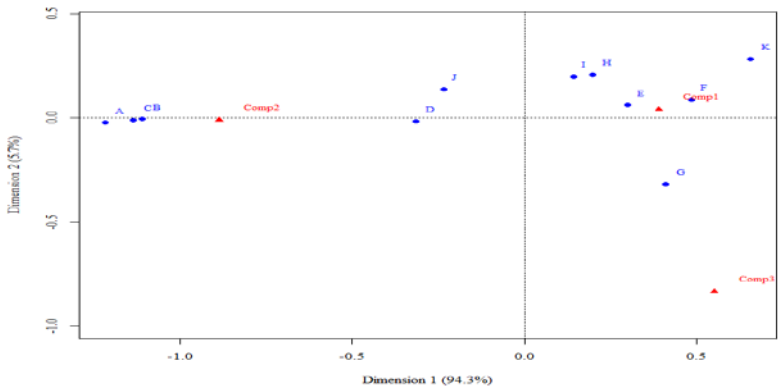

Fig.2 3 Correspondence Analysis between Each Diagnosed Uterine Fibroids and Age Group

(Among them, A-F are the 11 age groups in Table 1.1; Comp1-Comp3 are the three principal component indicators)

3.4 Difference Test

In order to investigate the differences of different diseases at various ages, a significance test was carried out on uterine fibroids disease. Since there are three main components of uterine fibroids disease samples, the Kruskal-Wallis test is used. The original hypothesis is that at the level of significance $\alpha = 0.05$. The parameters of the overall distribution position between samples are equal. The test results are shown in Table 2.5.

Table 2 5 Results Of the Difference Test for Each Uterine Fibroids

Category	Number of observations	Median	Average rank	Z statistic
Comp1	11	2.16	22.2	2.18
Comp2	11	2.19 20.4	1.43	
Comp3	11	0.00	8.4	-3.61

H = 13.21 DF = 2 P = 0.0012

ISSN 2522-6398 Vol. 3, Issue 11: 147-157, DOI: 10.25236/FER.2020.031124

Table 2.5 shows that at the significance level α , P value (P=0.0012)<0.05, it can be considered that the occurrence of the three types of diseases Comp1, Comp2 and Comp3 differs between different age groups. Therefore, different preventive measures should be taken for different uterine fibroids diseases at different ages. The larger the Z statistic, the more patients with the disease. From the Z statistic, it is known that Comp1 type uterine fibroids disease has the highest probability, followed by Comp2 type uterine fibroids disease and Comp3 type uterine fibroids disease.

4. Conclusion

In this paper, based on statistical thinking, the principal component analysis method and correspondence analysis method, the 673 cases of uterine fibroids patients in the hospital A were used as samples to establish the principal component correspondence analysis model. The model can quantitatively reflect the objective characteristics between each uterine fibroids and age distribution. Therefore, it can be closer to practical problems. The results show that there are differences in the prevalence of various uterine fibroids diseases at different ages, with the highest probability of Comp1 disease. Combined with Figure 2.3, E (41-45 years old) and F (46-50 years old) are closest to Comp1, so it can be concluded that the high incidence of uterine fibroids disease is 41-50 years old, which is consistent with the frequency distribution table in Table 1.1. This result has certain reference value for adjusting preventive measures of uterine fibroids disease and formulating corresponding strategies.

Acknowledgments

Guangxi Natural Science Foundation Project (No.2016GXNSFBA380102); Guilin University of Electronic Technology Graduate Education Innovation Program (2018YJCX58).

References

[1] Deng Lanyun(2014). Analysis of risk factors of uterine fibroids. China Public Health,, vol.20,no.6,pp:807-808.
[2] Lu Wei, Fan Bozhen, Li Huaifang, et al(2008). Analysis of the related factors of the incidence of uterine fibroids. Chinese Journal of Medical Guide, vol.62,no.9,pp:1320-1323.
[3] Ye J, Wang H, Chen Y B, et al(2015). MED12 mutation in patients with hysteromyoma. Oncology Letters, vol.9,no.6,pp:2771.
[4] Xu Ya, Chen Sidong, Zhu Chunyan, et al(2000). Case-control study on the 1:2 ratio of risk factors for uterine fibroids. Chinese Journal of Epidemiology, vol.21,no.5,pp:366-368.
[5] Yang Huiyun, Zhu Lujuan, Li Shiping, et al(2004). A case-control study on risk factors of uterine fibroids. Practical Preventive Medicine, vol.11,no.1,pp:5-6.

ISSN 2522-6398 Vol. 3, Issue 11: 147-157, DOI: 10.25236/FER.2020.031124

[6] Li Jinling, Zhao Zhi, Liu Hongchang, et al(2015). Research on the mineral element content of Gastrodia elata based on principal component analysis . China Journal of Chinese Materia Medica, vol.40,no.6,pp:1123-1128.
[7] Nicolas DF, Keurcien L, Guillaume L, et al(2016). Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data:. Molecular Biology & Evolution, vol.33,no.4,pp:1082-1093.
[8] Pang Hui, Yu Jianxing, Yu Yang, et al(2017). Application of correspondence analysis method in the study of mortality of different types of heart disease. Chinese Journal of Disease Control, vol.21,no.10,pp:1074-1076.
[9] Jin Z , Li Z , Li Q , et al(2015). Canonical correspondence analysis of soil heavy metal pollution, microflora and enzyme activities in the Pb-Zn mine tailing dam collapse area of Sidi village, SW China. Environmental Earth Sciences, vol.73,no.1,pp:267-274.
[10] Mao Shisong, Cheng Yiming, Pu Xiaolong(2011). Probability Theory and Mathematical Statistics Course . Shanghai: Higher Education Press, pp:115.