Advances in Applied Mathematics 应用数学进展, 2021, 10(6), 2155-2165
Published Online June 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.106225

Hans汉斯

四因素方差分析模型构建及癌症风险评估

邵梦瑶，贺兴时，李玲玲

西安工程大学理学院，陕西 西安
Email: shaomengyao0830@126.com, xsh1002@126.com, linglinglimath@163.com

收稿日期：2021年5月21日；录用日期：2021年6月9日；发布日期：2021年6月25日

摘　要

在实际问题中，往往需要分析多重因素对实验结果的影响，方差分析是解决此类问题的一个重要工具。然而，对于方差分析的理论目前集中在单因素方差分析和两因素方差分析。为此，基于两因素方差分析模型给出带有交互效应的四因素方差分析模型的理论推导，并应用其来分析地域、性别、吸烟、离子辐射对癌症患病风险的影响。本文在两因素方差分析模型基础上给出了具有交互作用的四因素方差分析模型的理论推导，并将其应用到具体实例中。在应用多因素方差分析模型时，可以通过对数据做变换来达到正态性、方差齐性的要求。对癌症发病率数据的方差分析结果表明，吸烟时长和离子辐射剂量对癌症的患病风险具有显著性影响，地域和性别并没有显示对癌症风险具有显著性影响。

关键词

四因素方差分析，离子辐射，吸烟，癌症风险

Construction of Four Factor Analysis of Variance (*ANOVA*) Model and Cancer Risk Assessment

Mengyao Shao, Xingshi He, Lingling Li

School of Science, Xi'an Polytechnic University, Xi'an Shaanxi
Email: shaomengyao0830@126.com, xsh1002@126.com, linglinglimath@163.com

Received: May 21st, 2021; accepted: Jun. 9th, 2021; published: Jun. 25th, 2021

Abstract

In practical problems, it is often necessary to analyze the influence of multiple factors on the ex-

文章引用: 邵梦瑶, 贺兴时, 李玲玲. 四因素方差分析模型构建及癌症风险评估[J]. 应用数学进展, 2021, 10(6): 2155-2165. DOI: 10.12677/aam.2021.106225

perimental results. Analysis of variance is an important tool to solve the problems. However, the theory of variance analysis (*ANOVA*) focuses on one-way analysis of variance and two-way analysis of variance (*ANOVA*). Therefore, based on the two factor analysis of variance (*ANOVA*) model, this paper gives the theoretical derivation of the four factor analysis of variance (*ANOVA*) model with interaction effect, and applies it to analyze the influence of region, gender, smoking and ion radiation on cancer risk. Based on the two factor analysis of variance model, this paper gives the theoretical derivation of the interactive four factor analysis of variance model and applies it to specific examples. When applying a multi-factor analysis of variance model, the data can be transformed to meet the requirements of normality and homogeneity of variance. The results of variance analysis of cancer incidence data show that smoking duration and ionizing radiation dose have a significant impact on the risk of cancer, region and gender do not show a significant impact on cancer risk.

Keywords

Four Factor Analysis of Variance, Ion Radiation, Smoking, Cancer Risk

Copyright © 2021 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

方差分析是由英国统计学家 R. A. Fisher 在 20 世纪 20 年代提出的，用于对样本均值的显著性检验[1]-[6]。通过推断各观测变量总体均值在控制变量的不同水平下是否存在显著差异，进而分析控制变量对观测变量是否存在显著的影响[1]。

方差分析包括单因素方差分析与多因素方差分析。多因素方差分析中的控制变量在两个或两个以上，主要研究目的是分析多个控制变量的主效应和交互作用对实验结果是否产生显著的影响。王苗苗基于当下多元方差分析的研究现状给出有无交互作用的双因素方差分析模型[7]。戴金辉与韩存在给出有交互作用双因素方差分析模型理论的基础上，通过例题说明方差分析有无交互作用在实验中的影响，使得方差分析理论更加成熟[8]。在方差分析中，样本必须满足独立、正态、方差齐性要求。教材[2] [3] [4] [5] [6]只给出了单因素及两因素方差分析的基本理论，但在实际应用中，试验结果的影响因素往往不止一个。刘晓华在双因素方差分析的基础上给出了三因素方差分析模型的数学推导[9]。黄伯强、李启才将带有交互效应的双因素方差分析进行了线性回归模型重构，证明了方差分析因素显著性 F 检验与回归模型的显著性检验是等价的[10]。戴金辉与代金辉将带有交互效应的双因素方差分析模型应用到跳水运动的成绩管理，分析运动员的动作和裁判员对运动员的主观因素对跳水运动员成绩的影响[11]。

正态性检验是方差分析的一个重要条件，Khatun Nasrin 证明了样本的正态性是统计推断中一个关键的假设条件，结合图形和检验方法，可以提高对数据正态性的判断[12]。Michael，JR 给出了 Q-Q 图与 P-P 图判定正态性的原理[13]。Philip Pallmann 等人提出了针对多个样本统计模型方差齐性的验证方法，发现 *Levene* 检验是作为检测多个组之间比例差异的有效方法，并验证了该方法的有效性[14]。

众所周知，外部因素是影响癌症发病的主要诱因，如生活方式、饮食习惯、环境污染等，因此，分析多重因素对癌症患病风险的影响具有重要的社会价值[15]。本文将双因素方差分析模型推广到具有交互效应的四因素方差分析模型，并通过模型分析地域、性别、吸烟、离子辐射是否对癌症患病风险具有显著影响。

2. 四因素方差分析模型

假定影响实验结果的因素有四个，记为 A，B，C，D。其中因子 A 取 r 个水平，因子 B 取 s 个水平，因子 C 取 t 个水平，因子 D 取 u 个水平，分别记为 $A_1,A_2,\cdots,A_r$，$B_1,B_2,\cdots,B_s$，$C_1,C_2,\cdots,C_t$，$D_1,D_2,\cdots,D_u$，在水平组合 $\left(A_i,B_j,C_k,D_l\right)$ 下样本相互独立且服从正态分布 $y_{ijkl}\sim N\left(\mu_{ijkl},\sigma^2\right)$。

令 $\mu=\frac{1}{rstu}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{l=1}^{u}\mu_{ijkl}$ 表示一般均值；

$\mu_{i\cdots}=\frac{1}{stu}\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{l=1}^{u}\mu_{ijkl}$ 表示在 A 的第 i 个水平下均值；

$\mu_{\cdot j\cdot\cdot}=\frac{1}{rtu}\sum_{i=1}^{r}\sum_{k=1}^{t}\sum_{l=1}^{u}\mu_{ijkl}$ 表示在 B 的第 j 个水平下均值；

$\mu_{\cdot\cdot k\cdot}=\frac{1}{rsu}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{l=1}^{u}\mu_{ijkl}$ 表示在 C 的第 k 个水平下均值；

$\mu_{\cdots l}=\frac{1}{rst}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\mu_{ijkl}$ 表示在 D 的第 l 个水平下均值；

$\mu_{ij\cdot\cdot}=\frac{1}{tu}\sum_{k=1}^{t}\sum_{l=1}^{u}\mu_{ijkl}$ 表示在 A 的第 i 个水平和 B 的第 j 个水平组合下均值；

$\mu_{i\cdot k\cdot}=\frac{1}{su}\sum_{j=1}^{s}\sum_{l=1}^{u}\mu_{ijkl}$ 表示在 A 的第 i 个水平与 C 的第 k 个水平组合下均值；

$\mu_{i\cdot\cdot l}=\frac{1}{st}\sum_{j=1}^{s}\sum_{k=1}^{t}\mu_{ijkl}$ 表示在 A 的第 i 个水平与 D 的第 l 个水平组合下均值；

$\mu_{\cdot jk\cdot}=\frac{1}{ru}\sum_{i=1}^{r}\sum_{l=1}^{u}\mu_{ijkl}$ 表示在 B 的第 j 个水平与 C 的第 k 个水平组合下均值；

$\mu_{\cdot j\cdot l}=\frac{1}{rt}\sum_{i=1}^{r}\sum_{k=1}^{t}\mu_{ijkl}$ 表示在 B 的第 j 个水平与 D 的第 l 个水平组合下均值；

$\mu_{\cdot\cdot kl}=\frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}\mu_{ijkl}$ 表示在 C 的第 k 个水平与 D 的第 l 个水平组合下均值；

$\mu_{ijk\cdot}=\frac{1}{u}\sum_{l=1}^{u}\mu_{ijkl}$ 表示在 A 的第 i 个水平、B 的第 j 个水平、C 的第 k 个水平组合下均值；

$\mu_{ij\cdot l}=\frac{1}{t}\sum_{k=1}^{t}\mu_{ijkl}$ 表示在 A 的第 i 个水平、B 的第 j 个水平、D 的第 l 个水平组合下均值；

$\mu_{i\cdot kl}=\frac{1}{s}\sum_{j=1}^{s}\mu_{ijkl}$ 表示在 A 的第 i 个水平、C 的第 k 个水平、D 的第 l 个水平组合下均值；

$\mu_{\cdot jkl}=\frac{1}{r}\sum_{i=1}^{r}\mu_{ijkl}$ 表示在 B 的第 j 个水平、C 的第 k 个水平、D 的第 l 个水平组合下均值。

其中，在上述定义中，$i=1,2,\cdots,r$；$j=1,2,\cdots,s$；$k=1,2,\cdots,t$；$l=1,2,\cdots,u$。

则各因子主效应和交互效应定义为：

$\alpha_i=\mu_{i\cdots}-\mu$ 为因子 A 的主效应；$\beta_j=\mu_{\cdot j\cdot\cdot}-\mu$ 为因子 B 的主效应；

$\gamma_k=\mu_{\cdot\cdot k\cdot}-\mu$ 为因子 C 的主效应；$\theta_l=\mu_{\cdots l}-\mu$ 为因子 D 的主效应；

$\eta_{ij}=\mu_{ij\cdot\cdot}-\mu-\alpha_i-\beta_j$ 为因子 A 和 B 的交互效应；

$\eta_{ik}=\mu_{i\cdot k\cdot}-\mu-\alpha_i-\gamma_k$ 为因子 A 和 C 的交互效应；

$\eta_{il}=\mu_{i\cdot\cdot l}-\mu-\alpha_i-\theta_l$ 为因子 A 和 D 的交互效应；

$\eta_{jk}=\mu_{.jk.}-\mu-\beta_j-\gamma_k$ 为因子 B 和 C 的交互效应；

$\eta_{jl}=\mu_{.j\cdot l}-\mu-\beta_j-\theta_l$ 为因子 B 和 D 的交互效应；

$\eta_{kl}=\mu_{..kl}-\mu-\gamma_k-\theta_l$ 为因子 C 和 D 的交互效应；

$\eta_{ijk}=\mu_{ijk.}-\left(\mu_{ij..}-\mu_{i...}\right)-\left(\mu_{.jk.}-\mu_{.j..}\right)-\left(\mu_{i\cdot k.}-\mu_{..k.}\right)-\mu$ 为因子 A，B，C 的交互效应；

$\eta_{ijl}=\mu_{ij\cdot l}-\left(\mu_{ij..}-\mu_{i...}\right)-\left(\mu_{i..l}-\mu_{...l}\right)-\left(\mu_{.j\cdot l}-\mu_{.j..}\right)-\mu$ 为因子 A，B，D 的交互效应；

$\eta_{ikl}=\mu_{i\cdot kl}-\left(\mu_{i\cdot k.}-\mu_{i...}\right)-\left(\mu_{i..l}-\mu_{...l}\right)-\left(\mu_{..kl}-\mu_{..k.}\right)-\mu$ 为因子 A，C，D 的交互效应；

$\eta_{jkl}=\mu_{.jkl}-\left(\mu_{.jk.}-\mu_{.j..}\right)-\left(\mu_{.j\cdot l}-\mu_{...l}\right)-\left(\mu_{..kl}-\mu_{..k.}\right)-\mu$ 为因子 B，C，D 的交互效应；

$$\eta_{ijkl}=\mu_{ijkl}+\mu-\left(\mu_{ijk.}-\mu_{ij..}+\mu_{i...}\right)-\left(\mu_{ij\cdot l}-\mu_{i..l}+\mu_{.j..}\right)-\left(\mu_{i\cdot kl}-\mu_{i\cdot k.}-\mu_{..kl}+\mu_{..k.}\right)-\left(\mu_{.jkl}-\mu_{.jk.}-\mu_{.j\cdot l}+\mu_{...l}\right)$$

为因子 A，B，C，D 的交互效应。

因此，四因素方差分析模型为：

$$\begin{cases}
y_{ijklw}=\mu+\alpha_i+\beta_j+\gamma_k+\theta_l+\eta_{ij}+\eta_{ik}+\eta_{il}+\eta_{jk}+\eta_{jl}+\eta_{kl}+\eta_{ijk}+\eta_{ijl}+\eta_{ikl}+\eta_{jkl}+\eta_{ijkl}+\varepsilon_{ijklw} \\
\sum_{i=1}^{r}\alpha_i=0,\ \sum_{j=1}^{s}\beta_j=0,\ \sum_{k=1}^{t}\gamma_k=0,\ \sum_{l=1}^{u}\theta_l=0 \\
\sum_{i=1}^{r}\eta_{ij}=\sum_{j=1}^{s}\eta_{ij}=0,\ \sum_{i=1}^{r}\eta_{ik}=\sum_{k=1}^{t}\eta_{ik}=0,\ \sum_{i=1}^{r}\eta_{il}=\sum_{l=1}^{u}\eta_{il}=0,\ \sum_{j=1}^{s}\eta_{jk}=\sum_{k=1}^{t}\eta_{jk}=0, \\
\sum_{j=1}^{s}\eta_{jl}=\sum_{l=1}^{u}\eta_{jl}=0,\ \sum_{k=1}^{t}\eta_{kl}=\sum_{l=1}^{u}\eta_{kl}=0 \\
\sum_{i=1}^{r}\eta_{ijk}=\sum_{j=1}^{s}\eta_{ijk}=\sum_{k=1}^{t}\eta_{ijk}=0,\ \sum_{i=1}^{r}\eta_{ijl}=\sum_{j=1}^{s}\eta_{ijl}=\sum_{l=1}^{u}\eta_{ijl}=0,\ \sum_{i=1}^{r}\eta_{ikl}=\sum_{k=1}^{t}\eta_{ikl}=\sum_{l=1}^{u}\eta_{ikl}, \\
\sum_{j=1}^{s}\eta_{jkl}=\sum_{k=1}^{t}\eta_{jkl}=\sum_{l=1}^{u}\eta_{jkl} \\
\sum_{i=1}^{r}\eta_{ijkl}=\sum_{j=1}^{s}\eta_{ijkl}=\sum_{k=1}^{t}\eta_{ijkl}=\sum_{l=1}^{u}\eta_{ijkl} \\
\varepsilon_{ijklw}\text{相互独立，均服从正态分布}N\left(0,\sigma^2\right) \\
i=1,2,\cdots,r,\ j=1,2,\cdots,s,k=1,2,\cdots,t,l=1,2,\cdots,u,w=1,2,\cdots,m
\end{cases}$$

在该模型下分析每个因子的主效应、任意两因子交互效应、三因子的交互效应以及四因子的交互效应对实验结果是否存在显著影响，需进行如下 15 个假设检验：

$H_{OA}:\alpha_i=0,i=1,2,\cdots,r$；

$H_{OB}:\beta_j=0,j=1,2,\cdots,s$；

$H_{OC}:\gamma_k=0,k=1,2,\cdots,t$；

$H_{OD}:\theta_l=0,l=1,2,\cdots,u$；

$H_{OAB}:\forall i,j,$均有$\eta_{ij}=0$；

$H_{OAC}:\forall i,k,$均有$\eta_{ik}=0$；

$H_{OAD}:\forall i,l,$均有$\eta_{il}=0$；

$H_{OBC}:\forall j,k,$均有$\eta_{jk}=0$；

$H_{OBD}:\forall j,l,$均有$\eta_{jl}=0$；

$H_{OCD}:\forall k,l,$均有$\eta_{kl}=0$；

$H_{OABC}:\forall i,j,k,$均有$\eta_{ijk}=0$；

$H_{OABD}:\forall i,j,l,$均有$\eta_{ijl}=0$；

$H_{OACD}:\forall i,k,l,$均有$\eta_{ikl}=0$；

$H_{OBCD}:\forall j,k,l,$均有$\eta_{jkl}=0$；

$H_{OABCD}:\forall i,j,k,l,$均有$\eta_{ijkl}=0$。

3. 模型的分析

基于方差分析中平方和分解的思想，四因素方差分析模型中各偏差平方和分解如下：

$$S_E=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{l=1}^{u}\sum_{w=1}^{m}\left(y_{ijklw}-\overline{y_{ijkl\bullet}}\right)^2,\quad S_A=stum\sum_{i=1}^{r}\left(\overline{y_{i\bullet\bullet\bullet}}-\overline{y}\right)^2,\quad S_B=rtum\sum_{j=1}^{s}\left(\overline{y_{\bullet j\bullet\bullet}}-\overline{y}\right)^2,$$

$$S_C=rsum\sum_{k=1}^{t}\left(\overline{y_{\bullet\bullet k\bullet}}-\overline{y}\right)^2,\quad S_D=rstm\sum_{l=1}^{u}\left(\overline{y_{\bullet\bullet\bullet l}}-\overline{y}\right)^2,$$

$$S_{AB}=tum\sum_{i=1}^{r}\sum_{j=1}^{s}\left(\overline{y_{ij\bullet\bullet}}-\overline{y_{i\bullet\bullet\bullet}}-\overline{y_{\bullet j\bullet\bullet}}+\overline{y}\right)^2,\quad S_{AC}=sum\sum_{i=1}^{r}\sum_{k=1}^{t}\left(\overline{y_{i\bullet k\bullet}}-\overline{y_{i\bullet\bullet\bullet}}-\overline{y_{\bullet\bullet k\bullet}}+\overline{y}\right)^2$$

$$S_{AD}=stm\sum_{i=1}^{r}\sum_{l=1}^{u}\left(\overline{y_{i\bullet\bullet l}}-\overline{y_{i\bullet\bullet\bullet}}-\overline{y_{\bullet\bullet\bullet l}}+\overline{y}\right)^2,\quad S_{BC}=rum\sum_{j=1}^{s}\sum_{k=1}^{t}\left(\overline{y_{\bullet jk\bullet}}-\overline{y_{\bullet j\bullet\bullet}}-\overline{y_{\bullet\bullet k\bullet}}+\overline{y}\right)^2$$

$$S_{BD}=rtm\sum_{j=1}^{s}\sum_{l=1}^{u}\left(\overline{y_{\bullet j\bullet l}}-\overline{y_{\bullet j\bullet\bullet}}-\overline{y_{\bullet\bullet\bullet l}}+\overline{y}\right)^2,\quad S_{CD}=rsm\sum_{k=1}^{t}\sum_{l=1}^{u}\left(\overline{y_{\bullet\bullet kl}}-\overline{y_{\bullet\bullet k\bullet}}-\overline{y_{\bullet\bullet\bullet l}}+\overline{y}\right)^2$$

$$S_{ABC}=um\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\left(\overline{y_{ijk\bullet}}-\overline{y_{ij\bullet\bullet}}-\overline{y_{i\bullet k\bullet}}-\overline{y_{\bullet jk\bullet}}+\overline{y_{i\bullet\bullet\bullet}}+\overline{y_{\bullet j\bullet\bullet}}+\overline{y_{\bullet\bullet k\bullet}}-\overline{y}\right)^2$$

$$S_{ABD}=tm\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{l=1}^{u}\left(\overline{y_{ij\bullet l}}-\overline{y_{ij\bullet\bullet}}-\overline{y_{i\bullet\bullet l}}-\overline{y_{\bullet j\bullet l}}+\overline{y_{i\bullet\bullet\bullet}}+\overline{y_{\bullet j\bullet\bullet}}+\overline{y_{\bullet\bullet\bullet l}}-\overline{y}\right)^2$$

$$S_{ACD}=sm\sum_{i=1}^{r}\sum_{k=1}^{t}\sum_{l=1}^{u}\left(\overline{y_{i\bullet kl}}-\overline{y_{i\bullet k\bullet}}-\overline{y_{i\bullet\bullet l}}-\overline{y_{\bullet\bullet kl}}+\overline{y_{i\bullet\bullet\bullet}}+\overline{y_{\bullet\bullet k\bullet}}+\overline{y_{\bullet\bullet\bullet l}}-\overline{y}\right)^2$$

$$S_{BCD}=rm\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{l=1}^{u}\left(\overline{y_{\bullet jkl}}-\overline{y_{\bullet jk\bullet}}-\overline{y_{\bullet j\bullet l}}-\overline{y_{\bullet\bullet kl}}+\overline{y_{\bullet j\bullet\bullet}}+\overline{y_{\bullet\bullet k\bullet}}+\overline{y_{\bullet\bullet\bullet l}}-\overline{y}\right)^2$$

$$S_{ABCD}=m\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{l=1}^{u}\left(\overline{y_{ijkl\bullet}}-\overline{y_{ijk\bullet\bullet}}-\overline{y_{ij\bullet l\bullet}}-\overline{y_{i\bullet kl\bullet}}-\overline{y_{\bullet jkl\bullet}}+\overline{y_{ij\bullet\bullet\bullet}}+\overline{y_{i\bullet k\bullet\bullet}}+\overline{y_{i\bullet\bullet l\bullet}}\right.$$
$$\left.+\overline{y_{\bullet jk\bullet\bullet}}+\overline{y_{\bullet j\bullet l\bullet}}+\overline{y_{\bullet\bullet kl\bullet}}-\overline{y_{i\bullet\bullet\bullet\bullet}}-\overline{y_{\bullet j\bullet\bullet\bullet}}-\overline{y_{\bullet\bullet k\bullet\bullet}}-\overline{y_{\bullet\bullet\bullet l\bullet}}+\overline{y}\right)^2$$

其中，

$$\overline{y}=\frac{1}{rstum}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{l=1}^{u}\sum_{w=1}^{m}y_{ijklw}=\mu+\overline{\varepsilon},$$

$$\overline{y_{ijkl\bullet}}=\frac{1}{m}\sum_{w=1}^{m}y_{ijklw}=\mu+\alpha_i+\beta_j+\gamma_k+\theta_l+\eta_{ij}+\eta_{ik}+\eta_{il}+\eta_{jk}+\eta_{jl}+\eta_{kl}+\eta_{ijk}+\eta_{ijl}+\eta_{ikl}+\eta_{jkl}+\eta_{ijkl}+\overline{\varepsilon_{ijkl\bullet}},$$

$$\overline{y_{i\cdots}}=\frac{1}{stum}\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{l=1}^{u}\sum_{w=1}^{m}y_{ijklw}=\mu+\alpha_i+\overline{\varepsilon_{i\cdots}},$$

$$\overline{y_{\bullet j\cdots}}=\frac{1}{rtum}\sum_{i=1}^{r}\sum_{k=1}^{t}\sum_{l=1}^{u}\sum_{w=1}^{m}y_{ijklw}=\mu+\beta_j+\overline{\varepsilon_{\bullet j\cdots}},$$

$$\overline{y_{\bullet\bullet k\bullet}}=\frac{1}{rsum}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{l=1}^{u}\sum_{w=1}^{m}y_{ijklw}=\mu+\gamma_k+\overline{\varepsilon_{\bullet\bullet k\bullet}},$$

$$\overline{y_{\cdots l\bullet}}=\frac{1}{rstm}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{w=1}^{m}y_{ijklw}=\mu+\theta_l+\overline{\varepsilon_{\cdots l\bullet}},$$

$$\overline{y_{ij\cdots}}=\frac{1}{tum}\sum_{k=1}^{t}\sum_{l=1}^{u}\sum_{w=1}^{m}y_{ijklw}=\mu+\alpha_i+\beta_j+\eta_{ij}+\overline{\varepsilon_{ij\cdots}},$$

$$\overline{y_{i\bullet k\bullet}}=\frac{1}{sum}\sum_{j=1}^{s}\sum_{l=1}^{u}\sum_{w=1}^{m}y_{ijklw}=\mu+\alpha_i+\gamma_k+\eta_{ik}+\overline{\varepsilon_{i\bullet k\bullet}},$$

$$\overline{y_{i\bullet\bullet l\bullet}}=\frac{1}{stm}\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{w=1}^{m}y_{ijklw}=\mu+\alpha_i+\theta_l+\eta_{il}+\overline{\varepsilon_{i\bullet\bullet l\bullet}},$$

$$\overline{y_{\bullet jk\bullet}}=\frac{1}{rum}\sum_{i=1}^{r}\sum_{l=1}^{u}\sum_{w=1}^{m}y_{ijklw}=\mu+\beta_j+\gamma_k+\eta_{jk}+\overline{\varepsilon_{\bullet jk\bullet}},$$

$$\overline{y_{\bullet j\bullet l\bullet}}=\frac{1}{rtm}\sum_{i=1}^{r}\sum_{k=1}^{t}\sum_{w=1}^{m}y_{ijklw}=\mu+\beta_j+\theta_l+\eta_{jl}+\overline{\varepsilon_{\bullet j\bullet l\bullet}},$$

$$\overline{y_{\bullet\bullet kl\bullet}}=\frac{1}{rsm}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{w=1}^{m}y_{ijklw}=\mu+\gamma_k+\theta_l+\eta_{kl}+\overline{\varepsilon_{\bullet\bullet kl\bullet}},$$

$$\overline{y_{ijk\bullet\bullet}}=\frac{1}{rstm}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{w=1}^{m}y_{ijklw}=\mu+\alpha_i+\beta_j+\gamma_k+\eta_{ij}+\eta_{ik}+\eta_{jk}+\eta_{ijk}+\overline{\varepsilon_{ijk\bullet\bullet}},$$

$$\overline{y_{ij\bullet l\bullet}}=\frac{1}{rsum}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{l=1}^{u}\sum_{w=1}^{m}y_{ijklw}=\mu+\alpha_i+\beta_j+\theta_l+\eta_{ij}+\eta_{il}+\eta_{jl}+\eta_{ijl}+\overline{\varepsilon_{ij\bullet l\bullet}},$$

$$\overline{y_{i\bullet kl\bullet}}=\frac{1}{rtum}\sum_{i=1}^{r}\sum_{k=1}^{t}\sum_{l=1}^{u}\sum_{w=1}^{m}y_{ijklw}=\mu+\alpha_i+\gamma_k+\theta_l+\eta_{ik}+\eta_{il}+\eta_{kl}+\eta_{ikl}+\overline{\varepsilon_{i\bullet kl\bullet}},$$

$$\overline{y_{\bullet jkl\bullet}}=\frac{1}{stum}\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{l=1}^{u}\sum_{w=1}^{m}y_{ijklw}=\mu+\beta_j+\gamma_k+\theta_l+\eta_{jk}+\eta_{jl}+\eta_{kl}+\eta_{jkl}+\overline{\varepsilon_{\bullet jkl\bullet}}$$

总的平方和分解如下：

$$S_T=S_E+S_{ABCD}+S_{ABC}+S_{ABD}+S_{ACD}+S_{BCD}+S_{AB}+S_{AC}+S_{AD}+S_{BC}+S_{BD}+S_{CD}+S_A+S_B+S_C+S_D$$

方差分析模型的检验分析如表 1 所示。

Table 1. Analysis of variance of four variables with interaction effect
表 1. 具有交互效应的四元方差分析表

原假设	自由度	F 统计量	临界值
H_{OA}	$r-1$	$F_A=\frac{S_A/(r-1)}{S_E/rstu(m-1)}$	$F_\alpha(r-1,rstu(m-1))$
H_{OB}	$s-1$	$F_B=\frac{S_B/(s-1)}{S_E/rstu(m-1)}$	$F_\alpha(s-1,rstu(m-1))$
H_{OC}	$t-1$	$F_C=\frac{S_C/(t-1)}{S_E/rstu(m-1)}$	$F_\alpha(t-1,rstu(m-1))$
H_{OD}	$u-1$	$F_D=\frac{S_D/(u-1)}{S_E/rstu(m-1)}$	$F_\alpha(u-1,rstu(m-1))$
H_{OAB}	$(r-1)(s-1)$	$F_{AB}=\frac{S_{AB}/(r-1)(s-1)}{S_E/rstu(m-1)}$	$F_\alpha((r-1)(s-1),rstu(m-1))$
H_{OAC}	$(r-1)(t-1)$	$F_{AC}=\frac{S_{AC}/(r-1)(t-1)}{S_E/rstu(m-1)}$	$F_\alpha((r-1)(t-1),rstu(m-1))$
H_{OAD}	$(r-1)(u-1)$	$F_{AD}=\frac{S_{AD}/(r-1)(u-1)}{S_E/rstu(m-1)}$	$F_\alpha((r-1)(u-1),rstu(m-1))$
H_{OBC}	$(s-1)(t-1)$	$F_{BC}=\frac{S_{BC}/(s-1)(t-1)}{S_E/rstu(m-1)}$	$F_\alpha((s-1)(t-1),rstu(m-1))$
H_{OBD}	$(s-1)(u-1)$	$F_{BD}=\frac{S_{BD}/(s-1)(u-1)}{S_E/rstu(m-1)}$	$F_\alpha((s-1)(u-1),rstu(m-1))$
H_{OCD}	$(t-1)(u-1)$	$F_{CD}=\frac{S_{CD}/(t-1)(u-1)}{S_E/rstu(m-1)}$	$F_\alpha((t-1)(u-1),rstu(m-1))$
H_{OABC}	$(r-1)(s-1)(t-1)$	$F_{ABC}=\frac{S_{ABC}/(r-1)(s-1)(t-1)}{S_E/rstu(m-1)}$	$F_\alpha((r-1)(s-1)(t-1),rstu(m-1))$
H_{OABD}	$(r-1)(s-1)(u-1)$	$F_{ABD}=\frac{S_{ABD}/(r-1)(s-1)(u-1)}{S_E/rstu(m-1)}$	$F_\alpha((r-1)(s-1)(u-1),rstu(m-1))$
H_{OACD}	$(r-1)(t-1)(u-1)$	$F_{ACD}=\frac{S_{ACD}/(r-1)(t-1)(u-1)}{S_E/rstu(m-1)}$	$F_\alpha((r-1)(t-1)(u-1),rstu(m-1))$
H_{OBCD}	$(s-1)(t-1)(u-1)$	$F_{BCD}=\frac{S_{BCD}/(s-1)(t-1)(u-1)}{S_E/rstu(m-1)}$	$F_\alpha((s-1)(t-1)(u-1),rstu(m-1))$
H_{OABCD}	$(r-1)(s-1)(t-1)(u-1)$	$F_{ABCD}=\frac{S_{ABCD}/(r-1)(s-1)(t-1)(u-1)}{S_E/rstu(m-1)}$	$F_\alpha((r-1)(s-1)(t-1)(u-1),rstu(m-1))$

当 $F \geq F_{\alpha}$ 时，拒绝原假设，说明因素对实验结果有显著影响。

4. 实证分析

选取 1958~2009 年日本广岛和长崎原子弹爆炸 0~80 岁患者癌症发病率数据为研究对象，分析患者所居城市(city)、性别(sex)、离子辐射剂量(dose)、吸烟时长(smoking)对癌症发病率(incidence)的影响，具体数据见表 2。

Table 2. Incidence rate of all solid cancers under different level combinations
表 2. 不同水平组合下的所有实体癌症的发病率数据

city	*sex*	*dose*	*smoking*	*incidence*	*city*	*sex*	*dose*	*smoking*	*incidence*
1	1	1	1	361.597	2	1	1	1	306.494
1	1	1	2	662.125	2	1	1	2	129.825
1	1	1	3	382.384	2	1	1	3	381.762
1	1	2	1	134.432	2	1	2	1	114.137
1	1	2	2	218.142	2	1	2	2	179.364
1	1	2	3	231.967	2	1	2	3	207.580
1	2	1	1	240.446	2	2	1	1	137.382
1	2	1	2	1456.028	2	2	1	2	618.969
1	2	1	3	190.193	2	2	1	3	0.000
1	2	2	1	112.042	2	2	2	1	96.603
1	2	2	2	174.703	2	2	2	2	145.028
1	2	2	3	212.863	2	2	2	3	151.766

城市分为广岛和长崎，1 代表广岛，2 代表长崎；性别 1 代表男性，2 代表女性；离子辐射剂量按照是否大于 4 Gy 的标准进行分类，1 代表剂量大于等于 4 Gy，2 代表剂量小于 4 Gy；吸烟时长按照吸烟年限进行分类，将吸烟时长分为，1 (从不吸烟)，2 (吸烟 0~20 年)以及 3 (大于 20 年)分为三个水平。

对表 2 数据进行四因素方差分析，验证样本的正态性[16] [17] [18]、独立性以及方差齐性[19] [20] [21]，结果表明原始数据因变量 incidence 不服从正态分布和方差齐性。因此，取 1/incidence 作为新的因变量进行方差分析，利用 SPSS 22.0 对变换后的数据进行正态性与方差齐性检验，其正态性检验结果如表 3、图 1 和图 2 所示，方差齐性检验结果如表 4 所示。

Table 3. Normality test
表 3. 正态性检验表

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	统计	自由度	*p* 值	统计	自由度	*p* 值
1/incidence	0.077	24	0.200	0.984	24	0.960

由表 3 可知，$p > 0.05$，故这组数据满足正态分布，且 *P-P* 图与 *Q-Q* 图表明变换后的数据符合正态分布。

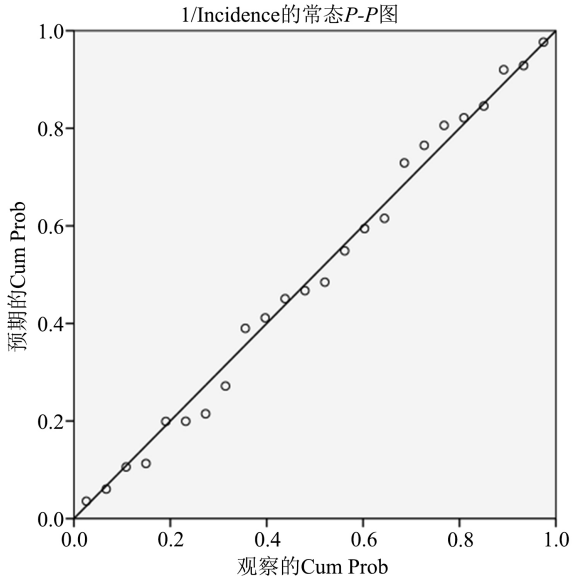

Figure 1. *P-P* diagram of normality test
图 1. 正态性检验 *P-P* 图

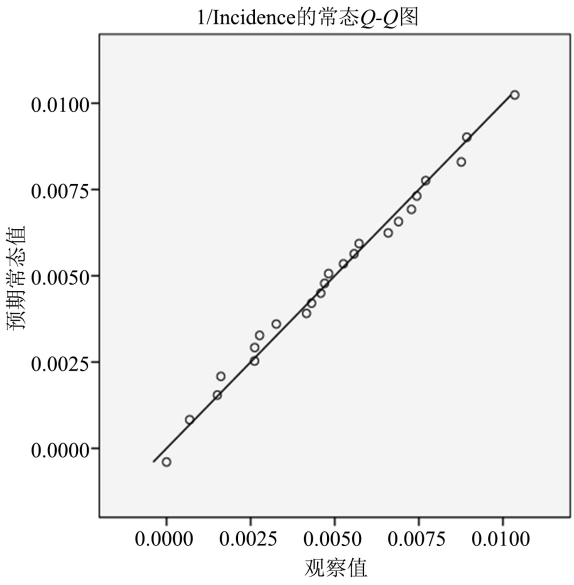

Figure 2. *Q-Q* diagram of normality test
图 2. 正态性检验 *Q-Q* 图

Table 4. Test table for homogeneity of variance
表 4. 方差齐性检验表

	Levene 统计资料	自由度 1	自由度 2	*p* 值
city	1.460	1	22	0.240
sex	0.910	1	22	0.350
dose	0.276	1	22	0.604
smoking	1.139	2	21	0.339

由表 4 可知，在显著水平取 0.05 时，城市、性别、离子辐射剂量、吸烟时长四个因素 p 值均大于 0.05，认为数据满足方差齐性检验。因此，变换后的数据满足方差分析条件。

对变换后的数据进行方差分析，通过 R 编程[22] [23] [24]得到结果如表 5 所示。

Table 5. Analysis of variance
表 5. 方差分析表

因子效应	自由度	平方和	均方	F 值	p 值
city	1	6.820e−06	6.820e−06	1.600	0.2414
sex	1	1.610e−06	1.610e−06	0.378	0.5558
dose	1	6.402e−05	6.402e−05	15.022	0.0047
smoking	1	3.035e−05	3.035e−05	7.121	0.0284
city: *sex*	1	1.620e−06	1.620e−06	0.380	0.5547
city: *dose*	1	1.400e−07	1.400e−07	0.033	0.8613
city: *smoking*	1	5.320e−06	5.320e−06	1.247	0.2965
sex: *dose*	1	3.510e−06	3.510e−06	0.823	0.3909
sex: *smoking*	1	2.480e−06	2.480e−06	0.583	0.4671
dose: *smoking*	1	4.090e−06	4.090e−06	0.959	0.3560
city: *sex*: *dose*	1	3.820e−06	3.820e−06	0.896	0.3716
city: *sex*: *smoking*	1	2.730e−06	2.730e−06	0.640	0.4469
city: *dose*: *smoking*	1	4.540e−06	4.540e−06	1.064	0.3324
sex: *dose*: *smoking*	1	1.250e−06	1.250e−06	0.293	0.6032
city: *sex*: *dose*: *smoking*	1	5.250e−06	5.250e−06	1.232	0.2992

由表 5 可知，在显著水平取 0.05 时，离子辐射剂量和吸烟时长对癌症发病风险具有显著性影响，城市和性别对癌症患病风险并没有显著性的影响，并且四个因素不存在显著性的交互效应，因此，研究癌症风险应该关注离子辐射剂量和吸烟时长等其他因素与癌症发病机制的关系。

5. 结论

本文在两因素方差分析模型基础上给出了具有交互作用的四因素方差分析模型的理论推导，并将其应用到具体实例中。在应用多因素方差分析模型时，可以通过对数据做变换来达到正态性、方差齐性的要求。

对癌症发病率数据的方差分析结果表明，吸烟时长和离子辐射剂量对癌症的患病风险具有显著性影响，地域和性别并没有显示对癌症风险具有显著性影响，因此，研究癌症的发病风险应该关注离子辐射和吸烟等其他因素在癌症发病机理中所起的作用，这将需要更具体的生物数学模型来验证，这也将是我们下一步的工作。

基金项目

陕西省教育厅专项科研计划项目，项目名称：多重因素对肺癌发病的影响与数据分析，项目编号：19JK0359。

参考文献

[1] 吴坚. 应用概率统计[M]. 第 2 版. 北京: 高等教育出版社, 2007: 262.

[2] 魏宗舒. 概率论与数理统计教程[M]. 北京: 高等教育出版社, 2001: 372-391.

[3] 张忠群. 概率论与数理统计[M]. 贵阳: 贵州大学出版社, 2008: 203.

[4] Hardle, W.K. and Simar, L. (2014) Applied Multivariate Statistical Analysis. 4th Edition, Springer-Verlag, Berlin.

[5] 何晓群. 多元统计分析[M]. 第四版. 北京: 中国人民大学出版社, 2015.

[6] 王学民. 应用多元统计分析[M]. 第五版. 上海: 上海财经大学出版社, 2017.

[7] 王苗苗. 双因素方差分析模型的构建及应用[J]. 统计与决策, 2015(18): 72-75.

[8] 戴金辉, 韩存. 双因素方差分析方法的比较[J]. 统计与决策, 2018, 34(4): 30-33.

[9] 刘晓华. 多元方差分析模型的构建与应用[J]. 统计与决策, 2019, 35(1): 75-78.

[10] 黄伯强, 李启才. 带交互作用的双因素方差分析的线性回归建模[J]. 统计与决策, 2021, 37(1): 10-15.

[11] 戴金辉, 代金辉. 方差分析在跳水运动成绩管理中的应用[J]. 统计与决策, 2016(22): 80-82.

[12] Khatun, N. (2021) Applications of Normality Test in Statistical Analysis. *Open Journal of Statistics*, **11**, 113-122. https://doi.org/10.4236/ojs.2021.111006

[13] Michael, J.R. (1983) The Stabilized Probability Plot. *Biometrika*, **70**, 11-17. https://doi.org/10.1093/biomet/70.1.11

[14] Pallmann, P., Hothorn, L.A. and Djira, G.D. (2014) A Levene-Type Test of Homogeneity of Variances against Ordered Alternatives. *Computational Statistics*, **29**, 1593-1608. https://doi.org/10.1007/s00180-014-0508-z

[15] Grant, E.J., Brenner, A., Sugiyama, H., *et al.* (2017) Solid Cancer Incidence among the Life Span Study of Atomic Bomb Survivors: 1958-2009. *Radiation Research*, **187**, 513-537. https://doi.org/10.1667/RR14492.1

[16] Sinz, F., Gerwinn, S. and Bethge, M. (2008) Characterization of the *p*-Generalized Normal Distribution. *Journal of Multivariate Analysis*, **100**, 817-820. https://doi.org/10.1016/j.jmva.2008.07.006

[17] Kolkiewicz, A., Rice, G. and Xie, Y. (2021) Projection Pursuit Based Tests of Normality with Functional Data. *Journal of Statistical Planning and Inference*, **211**, 326-339. https://doi.org/10.1016/j.jspi.2020.07.001

[18] Lilliefors, H.W. (2012) On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, **62**, 399-402. https://doi.org/10.1080/01621459.1967.10482916

[19] Conover, W.J., Johnson, M.E. and Johnson, M.M. (2012) A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data. *Technometrics*, **23**, 351-361. https://doi.org/10.1080/00401706.1981.10487680

[20] Esmailzadeh, N. (2019) A Comparison of Five Bootstrap and Non-Bootstrap Levene-Type Tests of Homogeneity of Variances. *Iranian Journal of Science and Technology*, *Transactions A*: *Science*, **43**, 979-989. https://doi.org/10.1007/s40995-018-0485-0

[21] Sharma, D. and Golam Kibria, B.M. (2013) On Some Test Statistics for Testing Homogeneity of Variances: A Comparative Study. *Journal of Statistical Computation and Simulation*, **83**, 1944-1963.

[22] 薛毅. 统计建模与 R 软件[M]. 北京: 清华大学出版社, 2007.

[23] Ran, Y. and Yuan, X. (2020) Analysis of the Influencing Factors of the Multi-Linear Regression Model Based on R Language on the Total Cost of Domestic Tourism. *Frontiers in Economics and Management*, **1**, 60-65.

[24] Kumar, M., Sonker, P.Kr., Saroj, A., Jain, A., Bhattacharjee, A. and Saroj, R.Kr. (2020) Parametric Survival Analysis Using R: Illustration with Lung Cancer Data. *Cancer Reports*, **3**, e1210. https://doi.org/10.1002/cnr2.1210