

使用R进行独立样本双因素方差分析

Alex / 2023-10-30 / free_learner@163.com / learning-archive.org

本文介绍在R环境下进行独立样本双因素方差分析、简单效应分析以及事后检验的基本方法。

一、背景

如果因变量是一个连续变量，自变量包括两个分类变量，每个分类变量具有两个或以上的水平，想要检验的问题是每个变量不同水平的均值是否相同以及两个变量之间是否存在交互作用，那么可以使用双因素方差分析模型（two-way ANOVA）。如果样本之间是相互独立的，那么这就是独立样本双因素方差分析。

二、平衡设计和非平衡设计

平衡设计（balanced design）指的是不同水平组合下的样本量是相同的，而非平衡设计（unbalanced design）指的是不同水平组合下的样本量不是完全相同的。对于非平衡设计的数据，一般需要使用Type-III ANOVA。关于不同类型的ANOVA，[这里](#)有一个很好的介绍。对于平衡设计的数据，不同类型的ANOVA结果是一致的。

三、样例数据

我这里使用R内置的 `ToothGrowth` 数据集，因变量是豚鼠的牙齿生长相关的测量，第一个自变量是喂食豚鼠维生素C的方式（OJ/VC），第二个自变量是维生素C的剂量（每天0.5/1/2毫克）。要检验的问题，不同的喂食维生素C的方式是否对牙齿生长具有不同的效果？不同的维生素C的剂量是否对牙齿生长具有不同的效果？以及喂食方式和剂量是否有交互效应？

```
> data("ToothGrowth")
> ToothGrowth$dose <- as.factor(ToothGrowth$dose)
> str(ToothGrowth)
'data.frame':   60 obs. of  3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

考虑到我日常接触的数据基本上都是非平衡设计，所以这里我去掉了两个样本，使得不同水平下的样本量不同：

```

> ToothGrowth <- ToothGrowth[-c(1,42),]
> table(ToothGrowth$supp, ToothGrowth$dose)

      0.5  1  2
OJ   10  9 10
VC   9 10 10

```

四、ANOVA检验

```

> library(car)
> lm_mod <- lm(len ~ supp + dose + supp:dose, data = ToothGrowth,
+             contrasts = list(supp="contr.sum", dose="contr.sum"))
> Anova(lm_mod, type = 'III')
Anova Table (Type III tests)

Response: len
      Sum Sq Df  F value    Pr(>F)
(Intercept) 20606.4  1 1539.9381 < 2.2e-16 ***
supp         181.0  1   13.5288 0.0005578 ***
dose        2286.9  2   85.4507 < 2.2e-16 ***
supp:dose    98.9  2    3.6957 0.0315703 *
Residuals   695.8 52
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

这里使用car包里的 `Anova` 进行Type-III ANOVA。在进行Type-III ANOVA时，要求因子变量具有正交的contrast。从上面的结果可以看到，`supp` 和 `dose` 交互效应显著（ $p=0.032$ ），即喂食方式对牙齿生长的影响与剂量大小有关（或者剂量对牙齿生长的影响与喂食方式有关）。通常，在交互效应（interaction effect）显著的情况下，就不应该去解读主效应（main effect）了。

五、简单效应分析

如果交互作用显著，一般会进行简单效应分析（simple effect analysis）：

```

> library(emmeans)
> EMM <- emmeans(lm_mod, ~ supp * dose)
> pairs(EMM, simple = "supp")
dose = 0.5:
  contrast estimate   SE df t.ratio p.value
OJ - VC      4.83 1.68 52   2.874 0.0059

dose = 1:
  contrast estimate   SE df t.ratio p.value
OJ - VC      5.86 1.68 52   3.489 0.0010

dose = 2:
  contrast estimate   SE df t.ratio p.value
OJ - VC     -0.08 1.64 52  -0.049 0.9612

```

这里使用emmeans包进行简单效应分析。从上面的结果可以看到，在剂量为0.5和1的时候，两种喂食方式对牙齿生长的影响具有显著差异，在剂量为2的时候，喂食方式差异不显著。如果我们想查看每种喂食方式下，不同剂量的差异，可以使用 `pairs(EMM, simple = "dose")`。

六、事后检验

如果交互效应不显著，但是主效应显著（且因子水平等于或者大于3），需要进行事后检验（post-hoc tests）。在当前的数据里，交互效应是显著的（因此不应该进行事后检验），这里只是为了展示一下进行事后检验的方法：

```

> EMM <- emmeans(lm_mod, ~ dose)
NOTE: Results may be misleading due to involvement in interactions
> pairs(EMM)
  contrast      estimate   SE df t.ratio p.value
dose0.5 - dose1   -8.89 1.19 52  -7.477 <.0001
dose0.5 - dose2  -15.29 1.17 52 -13.034 <.0001
dose1 - dose2     -6.40 1.17 52  -5.456 <.0001

Results are averaged over the levels of: supp
P value adjustment: tukey method for comparing a family of 3 estimates

```

同样使用emmeans包进行事后检验。默认情况下是使用Tukey's test进行多重比较校正，通过 `adjust` 参数可以设置其他校正方法。