使用R进行独立样本单因素方差分析

Alex / 2023-10-23 / free_learner@163.com / learning-archive.org

更新于2023-10-30，增加了一种事后检验的方法。

本文介绍在R环境下进行独立样本单因素方差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，自变量是一个等于或大于3个水平的分类变量，想要检验的问题是不同组别之间的均值是否相同，那么可以使用单因素方差分析模型（one-way ANOVA）。如果不同组别的样本是相互独立的，那么这就是独立样本单因素方差分析。

二、样例数据

我这里使用R内置的`PlantGrowth`数据集，因变量是农作物的产量，自变量是三种不同的种植方法，要检验的问题是，三种不同的种植方法得到的农作物产量的均值是否显著不同？

```
> data("PlantGrowth")
> str(PlantGrowth)
'data.frame':    30 obs. of  2 variables:
 $ weight: num  4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14 ...
 $ group : Factor w/ 3 levels "ctrl","trt1",..: 1 1 1 1 1 1 1 1 1 1 ...
```

三、ANOVA检验

```
> aov_mod <- aov(PlantGrowth$weight ~ PlantGrowth$group)
> summary(aov_mod)
                 Df Sum Sq Mean Sq F value Pr(>F)
PlantGrowth$group  2  3.766  1.8832   4.846 0.0159 *
Residuals         27 10.492  0.3886
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

这里使用base R里的`aov`函数拟合ANOVA模型，从结果可以看到组别变量对应的p值小于0.05，说明至少有一组的均值不同于另外两组。

四、事后检验

如果ANOVA检验显著，那么需要进行事后检验（post-hoc tests）来进一步检验任意两组之间的均值是否显著不同。如果ANOVA检验不显著，一般就没有必要进行事后检验。进行事后检验，需要考虑多重比较校正的问题。我自己一般使用两种方法来控制假阳性率，一种是使用Tukey's Test，一种是使用FDR。下面是具体的代码和结果：

```
> TukeyHSD(aov_mod)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = PlantGrowth$weight ~ PlantGrowth$group)

$`PlantGrowth$group`
            diff        lwr       upr     p adj
trt1-ctrl -0.371 -1.0622161 0.3202161 0.3908711
trt2-ctrl  0.494 -0.1972161 1.1852161 0.1979960
trt2-trt1  0.865  0.1737839 1.5562161 0.0120064
```

```
> pairwise.t.test(PlantGrowth$weight, PlantGrowth$group, p.adjust.method='fdr')

    Pairwise comparisons using t tests with pooled SD

data:  PlantGrowth$weight and PlantGrowth$group

     ctrl  trt1
trt1 0.194 -
trt2 0.132 0.013
```

从上面的结果可以看到，两种方法得到的结论是相同的，即`trt1`组显著小于`trt2`组。`TukeyHSD`和`pairwise.t.test`都是base R的函数。

20231030更新

也可以使用emmeans包进行事后检验，方法如下：

```
> library(emmeans)
> EMM <- emmeans(aov_mod, ~group)
> contrast(EMM, 'pairwise', adjust = 'fdr')
 contrast    estimate    SE df t.ratio p.value
 ctrl - trt1    0.371 0.279 27   1.331  0.1944
 ctrl - trt2   -0.494 0.279 27  -1.772  0.1315
 trt1 - trt2   -0.865 0.279 27  -3.103  0.0134

P value adjustment: fdr method for 3 tests
```

默认情况下是使用Tukey's test，通过`adjust`参数可以选择其他校正方法。