第十章 方差分析

10.1 方差分析的基本概念

方差分析 （Analysis of Variance, ANOVA）是一种常用的统计方法，用于比较两个或多个组间的均值是否存在显著差异。其基本思想是将总变差分解为组内变差和组间变差，通过比较组内变差与组间变差的比值，来判断均值的差异是否显著。

方差分析的基本公式：

总平方和（SST）= 组间平方和（SSB）+ 组内平方和（SSW）

其中，

- SST 表示所有数据离均差平方和，即总离差平方和。
- SSB 表示各组均值与总体均值差的平方和，反映不同组之间差异的大小。
- SSW 表示各组内个体与各自组均值差的平方和，反映同一组内个体变异程度的大小。

进行方差分析前，应保证模型要满足以下基本设定：

1. 正态性分布假定：每个分组内的观测值满足正态分布。
2. 方差齐性假定：每个组的方差相等。
3. 独立性假定：每个组内的观测值是相互独立的。
4. 随机抽样假定：从总体中随机选样并将其分为若干组。

如果模型不满足这些基本设定，将会对方差分析的正确性和准确性造成影响，可能会导致结果产生偏差或错误的结论。

10.2 单因素方差分析法

单因素方差分析法是一种常见的统计方法，用于比较不同组之间均值的差异性，以确定某个因素是否对观测变量产生了显著影响。

因为我们的样本都是服从正态分布的并且样本总体都是均值相等的，我们就可以得出

$$\frac{SS_E}{\sigma^2} \sim \chi^2_{(n-r)}, \quad \frac{SS_A}{\sigma^2} \sim \chi^2_{(r-1)}$$

然后再进而推出

$$F = \frac{MS_A}{MS_E} = \frac{SS_A/(r-1)}{SS_E/(n-r)} \sim F(r-1, n-r)$$

我们下面就来探究一下上述方法的分析步骤：

我们先来看个案例

为了保护人们健康，科学家想探究工人在工地工作过程中，周围的粉尘环境是否会对工人们的肺造成影响。这里我们总共挑选了 18 个小白鼠，按照数量平均分到了 A、B、C 三组，分别放在三个不同粉尘含量的地方。然后我们于数周之后再次去测量小白鼠们的全肺湿重。

我们首先应该提出假设。

原假设：粉尘环境不会影响小白鼠的全肺湿重

备择假设：原假设是错误的

然后我们来进行计算：

$$\sum_j^{n_1} X_{ij} = 22.9, \quad \sum_j^{n_2} X_{ij} = 25.4, \quad \sum_j^{n_3} X_{ij} = 28.4$$

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} X_{ij}^2 = 333.39, \quad C = \frac{76.7^2}{18} = 326.8272$$

$$SS_T = 333.39 - 326.8272 = 6.5628$$

$$SS_A = \frac{22.9^2}{6} + \frac{25.4^2}{6} + \frac{28.4^2}{6} - 326.8272 = 2.5278$$

$$SS_E = SS_T - SS_A = 6.5628 - 2.5278 = 4.0350$$

我们再通过计算出来的结果来绘制表格，得出下表

变异来源	SS	df	MS	F 值	P 值
组间	2.528	2	1.264	4.698	<0.05
组内	4.035	15	0.269		
总计	6.563	17			

我们再用 rstudio 来计算在 5%下的边界值和 P

```
> qf(0.05,2,15,lower.tail=FALSE)
[1] 3.68232
> pf(4.698,2,15,lower.tail = FALSE)
[1] 0.02604922
```

我们再引入一个新函数，aov 函数，这个函数是专门用来于方差分析的。用来给数据进行计算和检验。我们再用 R 语言来验证该模型的结果，代码如下

```
> X<-c(4.2,3.3,3.7,4.3,4.1,3.3,4.5,4.4,3.5,4.2,4.6,4.2,5.6,3.6,4.5,5.1,4.9,4.7)
> A<-factor(rep(1:3,each=6))
> my.data<-data.frame(X,A)> my.aov<-aov(X~A,data = my.data)
> summary(my.aov)
 Df Sum Sq Mean Sq F value Pr(>F)
A            2  2.528   1.264   4.698  0.026 *
Residuals   15  4.035   0.269
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

得出的结论否定了我们的原假设。