第九章 方差分析与回归分析

方差分析及回归分析是数理统计中有着应用的内容，本章对它们的最基本的部分作一介绍.

9.1 单因素试验的方差分析

在第八章中，我们讨论过两个独立样本的比较问题,两个独立样本常产生于两个独立的试验组.在实际工作中,有些试验涉及多个试验组,这时我们需要同时比较多个试验组的观测结果.例如,在医学试验中,可能需要同时比较两个以上的不同药物的试验结果，或同时比较不同药物和不同性别的试验结果.再比如,在工业生产中,可能需要比较不同配料成分、反应温度的试验结果.这类问题常归结为对多个总体的均值的比较,处理这类问题通常采用所谓的方差分析方法.要注意的是：方差分析不是分析总体的方差，而是通过分析数据的变异性来对多个总体的均值作比较.

影响试验结果（称为试验指标）的因素总是很多的.如果我们专注于某个特定因素（用A表示该因素）对试验指标的影响，那我们在设计试验时，让其他因素尽可能保持在相同的状态,而让因素A取定几个不同状态，这些不同的状态称为因素A的水平,然后在每个水平下进行若干次试验或观察.这样的试验我们称为单因素试验.我们先看下面例子.

例9.1.1 设有三台机器用来生产规格相同的铝合金薄板.抽取，测量薄板的厚度精确至千分之一厘米,得结果如表9-1所示

表9-1 铅合金板的厚度

|  |
| --- |
| 机器I 机器II 机器III |
| 0.236 0.257 0.258  0.238 0.253 0.264  0.248 0.255 0.259  0.245 0.254 0.267  0.243 0.261 0.262 |

在这个问题中,试验指标是薄板的厚度.因素是机器,3台机器就是这个因素的三个水平.我们假定除机器这一因素外，材料的规格、操作人员的水平等其他条件都相同或近似相同,这是单因素试验.我们的目的是为了考察各台机器生产的薄板的厚度有无显著差异,即考察机器这一因素对厚度有无显著影响.

例9.1.2 在饲料养鸡增肥的研究中,某研究所提出三种饲料配方: 是以鱼粉为主的饲料, 是以槐米粉为主的饲料, 是以苜蓿粉的饲料.为比较三种饲料的效果,选24只相似的雏鸡随机地分为三组,每组各喂一种饲料,60天后观察它们的重要.试验结果如表9-2所示:

表9-2 鸡饲料试验数据 单位:g

|  |  |
| --- | --- |
| 饲料 | 鸡重 |
|  |  |

本例中,饲料是因素,其有三个不同的水平即三种不同的配方.我们的目的之一是比较三种饲料配方下鸡的平均重量是否有差异?进一步，如果认为平均重量有显著差异,那么哪种饲料配方的效果最好?

在以上两个例子中,我们可以看到同一个水平下的试验结果不尽相同,这种组内的内在变异性我们归结于随机误差造成的.而不同的水平下的试验结果也有差异,我们要考查的就是不同水平下的试验结果之间的差异是系统性的因素(即因素的不同水平)造成的差异,还是可以仅仅归结为由随机误差造成的差异.为此，我们需要做一些基本假定以建立统计模型,并把所研究的问题归结为一个统计推断问题.下面介绍的方差分析方法就是基于所设定的统计模型下,为解决以上问题而开发出的一种统计技术,目的就是要检测出不同组之间是否存在系统性差异（即要回答因素的效应是否显著）以及差异程度如何？

9.1.1单因素试验方差分析的统计模型

在单因素试验中,用表示所考察的因素,设其有个水平,记为,在水平下安排次试验,试验结果记为.我们把看成是来自某个总体的样本,这样有个样本,由于各个水平下的试验是独立进行的，因此这个样本相互独立.设个总体的均值分别为.我们首先要检验这个总体的均值是否有显著差异?即检验假设

 对 不全相等 (9.1.1)

如经过检验拒绝原假设,我们就说因素A的效应是显著的.

为了解决以上问题,我们需要对各个水平对应的总体的分布要作进一步假定.通常假定各个水平下的样本来自正态总体.记，那么，故看成是随机误差.我们的统计模型如下:

,，

诸～. (9.1.2)

为了更好地描述数据,常在方差分析中引入总平均和水平效应的概念.记,,其中,称为总平均,称为水平的主效应,简称为的效应.有了以上记号,上面的统计模型表示为

,,

，

诸～. (9.1.3)

这里称为总平均,称为因素A的第个水平的效应.假设检验问题(9.1.1)改写为

 对 不全为零 （9.1.4）

9.1.2 平方和分解

我们从平方和分解着手,导出假设检验问题(9.1.4)的检验统计量.

通常在单因素方差分析中把试验数据汇总为如下表格:

表9-3 试验数据

|  |
| --- |
| 因素水平 试验数据 和 平均 |
|  |
|  |

表中最后两列的和及平均的含义如下:

,,,

,.

引入总偏差平方和

,

总偏差平方和度量了全部试验数据的变异性.

可将总偏差平方和分解为两部分.



,

注意上式中的交叉项

，

记

,,

于是我们将分解为

.

容易看出,是各组的内在变异性的总和，这是仅由随机误差引起的. 称好误差平方和或组内偏差平方和. 度量了试验之中的组与组之间的变异性, 除了反映随机误差外,还反映了效应间的差异,称为效应平方和或组间偏差平方和.下面进一步讨论和的统计特性.

，

是来自总体的样本，从而

，

并且

，

所以

，

又相互独立，由分布的可加性有

.

下面计算的期望.

,

,

,

所以









结合,得

.

进一步还可以证明与相互独立，且当时

.

9.1.3 检验方法

由平方和和的统计意义可以看出：在原假设成立时即各组的总体均值无差异时, 效应平方和可归结为仅由随机误差引起的.反之若各总体均值之间有差异时，还包含效应间的差异引起的试验数据的变异性.因此可以笼统地说:在偏大时有理由拒绝原假设,但这种“偏大”应该是相对的，这是因为的大小不仅与数据个数有关,还与背景“噪声” (即随机误差的方差)有关.由前面对和统计特性的讨论,可以看到

,

,

可见在原假设成立时

,

因此用相对于的大小来检验假设是合理的,即检验统计量取为

.

记,,称为的自由度. 称为的自由度.总偏差平方和的自由度为.

检验统计量常写为

,

其中,.

检验问题拒绝域具有形式.为了确定拒绝域的临界值,或计算检验的值,需要知道检验统计量的零分布(指原假设成立时检验统计量的分布).我们需要下面定理.

定理 在单因素方差分析模型(9.1.3)下,有

(1) ;

(2) 与相互独立;

(3)在原假设成立时, .

由此定理知,在原假设成立时, 检验统计量服从自由度为的分布.

有了此结论后,我们立即可得上面检验问题的拒绝域为

，

或计算出检验的值.

通常将上述计算过程和计算结果汇总成如方差分析表.

表9-5 方差分析表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方差来源 | 平方和 | 自由度 | 均方 | F比 |
| 因素 |  |  |  |  |
| 误差 |  |  |  |  |
| 总和 |  |  |  |  |

在实际中,我们可以按以下公式计算各偏差平方和.

,

,



例9.1.3 设例9.1.1符合模型(9.1.3),检验假设():

, 不全相等

解 ,

,



的自由度依次为,,,得方差分析表如下

表9-6 方差分析表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方差来源 | 平方和 | 自由度 | 均方 | F比 |
| 因素 |  |  |  |  |
| 误差 |  |  |  |  |
| 总和 |  |  |  |  |

由于,故在显著水平下拒绝原假设,认为各台机生产的薄板有显著差异.

例9.1.4 在方差分析中 ,对数据作线性变换不会影响方差分析的结果. 对于例9.1.2的数据, 我们将原始数据同时减去1000,得到下表:

表9-7 鸡饲料试验数据 单位:g

|  |  |
| --- | --- |
| 饲料 | 数据( 原始数据-1000) |
|  |  |
| 和 |  |

各偏差平方和为

,

,



的自由度依次为,,,得方差分析表如下

把各平方和及自由度填入方差分析表,并计算各均方及比,结果见表9-8

表9-8 方差分析表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方差来源 | 平方和 | 自由度 | 均方 | F比 |
| 因素 |  |  |  |  |
| 误差 |  |  |  |  |
| 总和 |  |  |  |  |

若取,则,由于,故认为因素饲料的效应是显著的,即三种不同饲料对鸡的增肥效果是有显著差异的.

9.1.4 参数估计

在检验结果为显著时,我们还需要估计各总体的均值,各水平效应,以及误差方差.这些参数的估计是简单的.

,,,

.

可以证明它们是各自参数的无偏估计,并且,是各自参数的最大的然估计.

当拒绝原假设时,还需给出两总体均值之差的区间估计.

由,,及正态分布的性质知

,

又与独立,且,从而

,

据此可得到我们需要的枢轴量



利用此枢轴量易得均值差的置信水平为的置信区间为

.

例9.1.5 求例9.1.3中未知参数,,c点估计,以及均值差的置信水平为的置信区间.

解: ,

,,,

,,.

由,及

,

可得均值差的0.95的置信区间分别为

:,

:,

:.