§8.5 总体分布拟合的检验

总体分布拟合检验是为检验观察到的一批数据是否与某种理论分布相符合.例如,我们考察某一产品的质量指标而打算用正态模型,可能事先有一些理论和经验上的根据,但这究竟是否可行?还需要通过样本进行检验.

拟合优度检验问题的一般提法如下：设为来自总体的样本，是一已知的分布函数（可能含有有限个参数）.要利用样本去检验假设:

～. (8.5.1)

导出这种检验的思想大致如下:设法提出一个反映实际数据与理论分布偏差的量.如果偏大,则认为观察数据与理论分布有较大偏离,从而否定原假设. 按显著性检验的范式,我们需要确定一个临界值,当,就拒绝原假设,否则不拒绝原假设.临界值满足,其中是预先选定的检验水平.然而这种“非此即彼”的做法不甚合理.一般说来,理论和实际没有截然的符合或不符合.原则上讲, 实际数据与理论分布完全符合的情况几乎没有.更恰当的提法是实际数据与理论分布符合的程度如何?或者说理论分布拟合实际数据的优良程度如何?因此对于检验问题(8.5.1),不是简单地以“是”或“否”来回答.最好的方式是用一个量化的指标来表示拟合的优良程度(简称为“拟合优度”),通常用一个介于0和1之间的数字作为拟合优度的刻画.一般采取如下做法:在有了具体的样本值后计算出偏差程度的量的具体值,将它记为,再计算概率

,

越大(即越接近于1),则表示理论分布与样本拟合得越好,原假设越可信.反之, 越接近于0,则反映理论分布与样本拟合得越不好,原假设越不可信.可见可以刻画用理论分布拟合观察数据的拟合优良程度.如果非要做出“非此即彼”的回答,就是事先指定一个(比如)如果,则否定原假设,否则不否定原假设.

由于可以有种种不同的方式定义,因而就有种种不同的拟合优度检验方法,其中最著名的一种检验方法是K.Pearson在1900年提出的拟合优度检验.

8.5.1 理论分布完全已知的情况

1.总体为只取有限个值的情形

设总体只取有限个值.考虑假设检验问题

 至少有一个，使得,

(8.5.2)

其中诸均己知,且.

设为来自该总体的样本，表示样本中取值为的频数.的MLE为,因此较大且成立, 诸与的偏离程度不应太大，换言之我们可用反映诸与的偏离程度的量来刻画理论分布拟合观察数据的优良程度.基于以上解释可用统计量



刻画偏离程度. 取何值呢? K.Pearson在1900年证明了若取,则在原假设成立时,有极限分布.因此检验统计量取为

,

称为的理论频数,又由于为原假设成立时的期望值,因此也称期望频数.而为实际观察到的频数, 因此也称观察,频数所以上面统计量常简记为

,

该检验问题的水平为拒绝域为：.

正如引言所述,对检验问题(8.5.2)只是简单地给出“是”或“否”的回答太牵强.常需求出一个拟合优度.做法如下:

由样本值算出检验统计量的观察值,该观察值记为,然后计算概率

,

由于在成立时, 有近似分布,因此我们可用分布计算上述概率的近似值.

例8.5.1 一家工厂分早、中、晚三班,每班8小时.近期发生了一些事故,早班6次,中班3次,晚班6次.据此怀疑事故发生率与班次有关.我们用这些数据来检验事故发生率与班次是否有关.

我们要检验的原假设为

事件发生率与班次无关.

如用分别表示早、中、晚班,表示事故发生的班次,则该检验问题的原假设可表述为

,

检验统计量观察值为



如取显著水平,拒绝域的临界值为.按分布可计算出.因此数据并未提供否定的证据.

如果以上数据改为早班30次,中班15次,晚班30次,再作这个问题的检验,可算得

,那么在显著水平下,可以拒绝,即可以认为事件发生率与班次有关.

两组数据的早、中、晚班发生的事故次数之比都是,而结论却不一样,这是为什么? 这是随机性的影响在起作用.对于前者,由于观察数太小,随机性影响就大.而对于后者, 观察数不小了, 随机性影响就小,各种事件发生的频率会相对稳定.在这么小的观察数之下,对目前这个结果,只宜解释为:一方面数据未能提供事故率与班次有关的支持,一方面也认表面上的差异也不宜完全忽视,值得进一步观察.

例(P200)

2. 理论分布为任一确定分布的情形

考虑假设检验问题:

服从分布,

其中为完全已知的分布.这种情况下的做法是

第一步 先将总体的取值范围分成个区间:,这里.即将的取值分成个区间.

第二步 计算在成立时概率

.

第三步 记为样本落入区间的频数.检验统计量为

,

同样地有在原假设成立时,有极限分布.往后的做法与前面相同.

8.5.2 理论分布含有未知参数的情形

考虑假设检验问题:

服从分布

这里为未知参数.该问题的检验方法如下

第一步 先将总体的取值范围分成个区间:,这里.即将的取值分成个区间.

第二步 记.

依赖于未知参数,必须用某种方法先估计出,进而得到的估计值.通常用最大似然估计法先估计,似然函数为

,

其中为样本落入区间的频数.由此似然函数得最大似然估计,从而得的最大似然估计值为.

第三步 计算检验统计量

,

在原假设成立时,有极限分布.往后的做法与前面相同.

注 利用似然函数,求最大似然估计并不容易.一种常用的做法是:对分布,直接利用样本求出参数的MLE.然后利用此MLE计算检验统计量.用这种方法计算的检验统计量的极限分布与有差距,但这种差距并不太,因此实用中人们常常这样做.

例(P202,204)

例 设表示人的性别(用1,2分别表示男性和女性);表示是否色盲(1,2分别表示正常和色盲).令,这时参数有3个,,其中,且.按遗传学模型性别和是否色盲的四种组合情况出现的机率分别为,,若调查了个人,桉性别和是否色盲的四种组合出现的人数分别记为,检验观察数据与遗传模型是否相符?即检验假设



解: 服从多项分布，其密度函数为



的MLE为，在原假设成立时,的MLE记为(写不出解析表达式),

检验统计量为

，

其中.拒绝域为.

容易看出这种检验的结果会依赖于分类(或者说分组),分类的不同有可能得出不同的结论.这便是在连续分布场合拟合优度检验的不足之处.因此对于正态分布这种非常重要的分布的检验就得寻找专门的检验方法.然而其它分布尚缺少专门的检验方法,因而也不得不用拟合优度检验.

补充

正态性检验：偏度、峰度检验及W检验

用于判断总体分布是否为正态分布的检验称为正态性检验.正态模型是最常用的模型，人们在对连续数据作统计分析时，往往先会考虑能否用正态分布拟合数据? 统计学家为此统计学家已有寻找了多种专门的检验方法，至今已有几十种正态性检验方法.比如简单而又直观的检验方法有正态概率纸方法. 统计学家们对正态性检验的种种方法进行了比较,以为总的来说以峰度、偏度检验方法,Wilk-Shapiro的W检验和Dagustino的D检验“夏皮洛-威尔克方法”较为有效.它们和第二类错误的概率较小.

1.峰度、偏度检验

设为来自总体的样本,我们检验假设

～,

总体的偏度和峰度的矩估计分别为

,,

表示样本的阶中心矩.正态分布的偏度为,峰度为,因此在原假设成立时样本偏度与0不应偏离太大, 样本峰度与3不应偏离太大.因此我们用与0的偏离程度以及与3的偏离程度来反映观察样本与正态分布的偏离程度.因此可以选取与作为检验统计量.

为了给出拒绝域,需要知道与的零分布.但它们精确的零分布不易得到,我们使用渐近分布.有以下结论:若总体服从正态分布,当充分大时,近似地有

～,

～,

记,,,

,,

那么当原假设为真且充分大时,近似地有～,～.拒绝域应具有形式

,或,

对于给定的显著水平,由以下两式确定

,,

利用近似的正态分布可得,,于是得拒绝域

,或.

例(P207)

2.W检验,D检验

W检验是夏皮诺(Shapiro)和威尔克(Wilk)于1965年提出的,这个检验当时可以利用.

设为样本,次序统计量为.W统计量定义为



其中系数在样本容量给定后,有表可查.对于假设

总体服从正态分布

其拒绝域为,而其中可查表得到.

可将W统计量简化为



W检验的导出比较复杂,此处略.

在样本容量时,的值已偏制成表,利用表中的可以比较方便地计算检验统计量.而在样本容量时,的计算越来越困难.为此提出了D检验,该检验不需要使用系数.下面介绍在样本容量时,检验正态性的D检验法.

.

经过修改，可以成为的一个线性无偏估计.取



作为正态性假设的检验统计量,由此导出的检验方法称为D检验,其中,.

在正态性假设成立时, 统计量的分布仅与样本容量有关,有渐近正态分布,并且

,



因此的近似标准化随机变量



的渐近分布为标准正态分布.但是统计量趋于标准正态分布的速度很慢,故Dagustino用随机模拟方法获得的分布分位数表.

在正态性假设成立时, 的值集中在零附近, 在正态性假设不成立时, 的值不是过小就是偏大,所以在给定显著水平后,D检验的拒绝域为

或

其中和可查表得到.