5.3 拟合优度检验

拟合优度检验是为检验观察到的一批数据是否与某种理论分布符合.例如,我们考察某一产品的质量指标而打算采用正态分布模型,或考察一种元件的寿命而打算采用指数分布模型,可能事先有一些理论或经验上的根据.但这究竟是否可行?有时就需要通过样本进行检验.例如,抽取若干个产品测定其质量指标,得 X_1, …, X_n.然后依据它们以决定"总体分布是正态分布"这样的假设能否被接受.又如,有人制造了一个骰子,他声称是均匀的,即出现各面的概率都是 1/6,是否如此?单审视骰子外形恐还不足以下判断,于是把骰子投掷若干次,记下其出现 1 点,2 点,…,6 点的次数,去检验这结果与"各面概率都是 1/6"的说法能否符合.

拟合优度检验在应用上很重要.除直接用于分布拟合外,列联表(见下文 5.3.3 段)也是一项重要应用.另外,这个问题在数理统计学发展史上占有一定的地位.其历史情况是这样的:统计分析方法在 19 世纪时多用于分析生物数据,那时曾流行一种看法认为正态分布普遍地适合于这类数据.到上世纪末,K. 皮尔逊对此提出问题,他指出有些数据有显著的偏态,不适于用正态模型.他于是

---

* 关于统计推断与统计决策的异同的论述,可参看前面所引陈希孺与倪国熙合著的书 pp. 222—225.

提出了一个包罗甚广的，日后以他的名字命名的分布族，其中包含正态分布，但也有很多偏态的.皮尔逊认为：第一步工作是根据数据去从这一大族分布中挑选一个最能反映所得数据性态的分布*.第二步就是要检验所得数据与这个分布的拟合如何，这一步就是似合优度检验.他为此引进了著名的“卡方检验法”(以后写为 χ^2 检验法).20 年代，R. A. 费歇尔对 χ^2 检验法作出了重要贡献，他纠正了皮尔逊工作中的一个关键性的错误.

5.3.1 理论分布完全已知且只取有限个值的情况

设有一总体 X.设从某种理论，或单纯作为一种假定，认为 X 的分布为

$$H_0: P(X = a_i) = p_i, i = 1, \cdots, k \qquad (3.1)$$

其中 $a_i, p_i, i=1, \cdots, k$，都为已知，且 $a_1, \cdots, a_k$ 两两不同，$p_i > 0$，$i=1, \cdots, k$.

现在从该总体中抽样 n 次，或者说，对 X 进行 n 次观察，得样本 $X_1, \cdots, X_n$.要根据它们去检验(3.1)的原假设 H_0 是否成立.至于为什么这种检验称为拟合优度检验，将在下文解释.

先设想 n 足够大.则按大数定理，若以 ν 记 $X_1, \cdots, X_n$ 中等于 a_i 的个数，应有 $\nu_i/n \approx p_i$，即 $\nu_i \approx np_i$.我们把 np_i 称为 a_i 这个“类”的理论值，而把 ν_i 称为其经验值或观察值.如下表所示

类　别	a_1	a_2	$\cdots$	a_i	$\cdots$	a_k
理论值	np_1	np_2	$\cdots$	np_i	$\cdots$	np_k
经验值	ν_1	ν_2	$\cdots$	ν_i	$\cdots$	ν_k

显然，表中最后两行差异愈小，则 H_0 愈像是对的，我们也就愈乐于接受它.现在要找出一个适当的量来反映这种差异.皮尔逊采用

* 我们在讲点估计时提到的“矩估计法”，就是皮尔逊为这个目的而创立的.有趣的是，目前在数理统计学中，矩法的 Popularity 反倒超过了皮尔逊分布族，这恐怕是皮尔逊始料所不及的.

的量是

$$\begin{aligned} Z &= \sum(\text{理论值} - \text{经验值})^2 / \text{理论值} \\ &= \sum_{i=1}^{k} (np_i - \nu_i)^2 / np_i \end{aligned} \tag{3.2}$$

这个量每项的分子部分好解释，分母用 np_i 则难于从直观上说清楚了，见下文.

这个统计量称为皮尔逊的拟合优度 χ^2 统计量，下文简称 χ^2 统计量. 名称的得来是因为下面这个重要定理，它是皮尔逊在1900年证明的：

定理 3.1 如果原假设 H_0 成立，则在样本大小 $n \to \infty$ 时，Z 的分布趋向于自由度 $k-1$ 的 χ^2 分布，即 χ^2_{k-1}.

这个定理从理论上说明了在 Z 的定义中，分母取为 np_i 的道理：若用别的值，就得不到这么简单的极限分布.

这个定理的严格证明超出了本课程范围之外. 为使读者相信其正确性，我们对 $k=2$ 这个简单情况仔细考察一下，在这一情况，有

$$np_2 = n(1-p_1), \nu_2 = n - \nu_1$$

于是

$$\begin{aligned} Z &= (np_1 - \nu_1)^2/np_1 + (n - np_1 - n + \nu_1)^2/n(1-p_1) \\ &= (\nu_1 - np_1)^2/np_1(1-p_1) \\ &= [(\nu_1 - np_1)/\sqrt{np_1(1-p_1)}]^2 \end{aligned}$$

据中心极限定理（第三章定理 4.3），当 $n \to \infty$ 时，$(\nu_1 - np_1)/\sqrt{np_1(1-p_1)}$ 的分布收敛于标准正态 $N(0,1)$. 于是 Z 的分布收敛于标准正态变量之平方的分布，按定义，即 $\chi^2_1 = \chi^2_{k-1}$，因此处 $k=2$.

用这个定理就可以对 H_0 作检验. 显然，应当在 $Z > C$ 时否定 H_0，$Z \leqslant C$ 时接受 H_0. C 的选取根据给定的水平 α，若近似地认为 Z 的分布就是 χ^2_{k-1}，则显然应取 C 为 $\chi^2_{k-1}(\alpha)$. 于是得到检验：

$$\varphi: \text{当 } Z \leqslant \chi^2_{k-1}(\alpha) \text{ 时接受 } H_0, \text{不然就否定 } H_0 \tag{3.3}$$

这是一个“非此即彼”的解决方式，在实用上，有时采取一种更有弹性的看法，它能提供更多的信息，且解释了“拟合优度”这个名词.

假定据一组具体数据算出的 Z 值为 Z_0. 我们提出这样的问题：在 H_0 成立之下，出现像 Z_0 这么大的差异或更大的差异的可能性有多大？按定理 3.1，这概率，暂记为 $p(Z_0)$，近似地为

$$p(Z_0) = P(Z \geqslant Z_0) \mid H_0) \approx 1 - K_{k-1}(Z_0)$$

其中 $K_{k-1}(X)$ 为自由度 $k-1$ 的 χ^2 分布函数. 显然，这个概率愈大，就说明即使在 H_0 成立时，出现 Z_0 这么大的差异就愈不稀奇，因而就愈使人们相信 H_0 的正确性. 以此之故，把 $p(Z_0)$ 解释为数据对理论分布(3.1)的“拟合优度”. 拟合优度愈大，就表示数据与理论之间的符合愈好，该理论分布也就获得更充足的实验或观察支持. 检验(3.3)不过是树立了一个门槛 α：当拟合优度 $p(Z_0)$ 低于 α 时，即放弃 H_0.* 自然，若取 $\alpha = 0.05$，则当 $p(Z_0) = 0.06$ 或 $p(Z_0) = 0.94$ 时，都接受 H_0. 但后者数据对理论分布的支持显然比前者大得多：前者虽勉强过关，但已接近崩溃的边缘.

例 3.1 考虑前面提的检验骰子均匀的问题，它相当于 $a_i = i, p_i = 1/6, i = 1, \cdots, 6$（$a_i$ 的具体值不重要，它只是代表一个类而已），设作了 $n = 6 \times 10^{10}$ 次投掷，得出各点出现的次数为（理论值：$np = 10^{10}$）

$$\nu_1 = 10^{10} - 10^6, \nu_2 = 10^{10} + 1.5 \times 10^6, \nu_3 = 10^{10} - 2 \times 10^6$$
$$\nu_4 = 10^{10} + 4 \times 10^6, \nu_5 = 10^{10} - 3 \times 10^6, \nu_6 = 10^{10} + 10^6/2 \tag{3.4}$$

算出这组数据的拟合优度统计量 Z 之值为

---

* 这种看法不仅适合于此处，也适合于前面所讲过的那些检验问题. 举例而言，设 X 是抽自正态总体 $N(\theta, 1)$ 的样本，要检验 $H_0: \theta \leqslant 0$，设 x_0 是 X 的具体值. 可以把 $P(X \geqslant x_0 \mid \theta = 0) = 1 - \Phi(x_0)$ 作为 x_0 这个数值的拟合优度. 如果 $x_0 > \nu_\alpha$，则拟合优度低于 α 而否定 H_0. 如果 $\alpha = 0.05$，则 $x_0 = 2$ 和 $x_0 = 100$ 都要否定 H_0，但后者提供的否定 H_0 的证据，显然比前者有力得多.

$$Z_0 = (10^{12} + 2.25 \times 10^{12} + 4 \times 10^{12} + 16 \times 10^{12}$$
$$+ 9 \times 10^{12} + 10^{12}/4)/10^{10} = 3250$$

此处 $k=6, k-1=5$. 查 χ^2 分布表, $K_5(3250)=0.9999\cdots$, 故拟合优度 $p(Z_0)$ 几乎是 0. 这说明, 实验数据极不支持"骰子均匀"这个假设.

这个结果值得玩味. 如拿数据(3.4)对 p_i 作估计, 则估出 p_i 的值都在 $1/6 \pm 10^{-4}$ 数量级之内. 从实用的观点看这恐怕可认为是足够均匀了. 这种差异, 即使存在, 也许并无实用意义. 可是, 由于试验次数极大, 我们达到了"明察秋毫"的地步, 把这么小的差异也检测出来了. 本例说明: 假设检验的结果的含义必须结合其他方面的考虑(样本大小, 估计值等), 才能得到更合理的解释. 统计上的显著性并不等于实用上的重要性, 这一点在前面已提醒过了.

下面举一个反方向的例子.

例 3.2 一家工厂分早、中、晚三班, 每班 8 小时, 近期发生了一些事故, 计早班 6 次, 中班 3 次, 晚班 6 次. 据此怀疑事故发生率与班次有关, 比方说, 中班事故率小些, 要用这些数据来检验一下.

我们把

$$H_0: \text{事故发生率与班次无关} \tag{3.5}$$

作为原假设. 如分别以 1,2,3 作为早、中、晚班的代号, 这个假设相当于(3.1)中的 $a_i = i$, $p_i = 1/3$, $i = 1,2,3$. 理论值为 $np_i = 15 \times 1/3 = 5$. 算出 Z 之值为

$$Z_0 = [(5-6)^2 + (5-3)^2 + (5-6)^2]/5 = 1.2$$

$k-1=3-1=2$. 查 χ_2 分布表, 得拟合优度

$$p(Z_0) = 1 - K_2(1.2) = 1 - 0.451 = 0.549$$

故数据未提供否定 H_0 的证据. 更清楚地说, 即使事故与班次完全无关, 在每一百家工厂中, 你平均会观察到 55 家, 其各班次事故数表面上的差异甚至比这里观察到的还大. 因此, 表面上 6:3:6 的差异其实并不稀奇.

没有统计思想的人易倾向于低估随机性的影响. 在此例中, 由

于观察数 $n=15$ 太小，随机性的影响就大了. 读者可计算一下：若观察的总事故达到 75 而仍维持上述比例(即早班 30 次，中班 15 次，晚班 30 次)，则 $p(Z_0)$ 降至 0.05 以下，因而有较充分的理由认为三个班次有差异了. 在 15 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，一方面也认为表面上的差异究竟不宜完全忽视，值得进一步观察.

5.3.2 理论分布只含有限个值但不完全已知的情况

先举两个例子.

例 3.3 回到“符号检验”中讨论过的那个问题. 被调查者对甲、乙两牌号何者为优的回答可能有三种：1. 甲优. 2. 乙优. 3. 认为一样或不回答. 所谓“甲乙两牌号一样”这时应理解为，这三种情况的概率依次为 $p_1=\theta$，$p_2=\theta$，$p_3=1-2\theta$，对某个 $\theta\geqslant 0$，$\theta\leqslant 1/2$. 在这里，理论分布只是部分已知(有上述形式，特别是，$p_1=p_2$)，但其中包含未知参数 θ，并不完全知道.

例 3.4 想要考察特定一群人的收入与其花在文化上的支出有无关系的问题. 把收入分成高、中、低三档，文化上的支出分为多、少两档. 则每个人可归入六个类别中之一. 分别以 $X=1,2,\cdots,6$ 记(高，多)，(高，少)，…，(低，少)这 6 类. 如果这二者独立，则应有，例如

$$P(\text{高},\text{多})=P(\text{高})P(\text{多})$$

分别以 p_1,p_2,p_3 记 $P(\text{高}),P(\text{中}),P(\text{低})$. 这三个数就是收入为高、中、低档者在全体人口中的比率，$p_1+p_2+p_3=1$. 类似地以 q_1 和 q_2 分别记 $P(\text{多}),P(\text{少})$，有 $q_1+q_2=1$. 这样，若独立性成立，则 X 的理论分布为

$$\begin{cases} P(X=1)=p_1q_1, P(X=2)=p_1q_2, P(X=3)=p_2q_1 \\ P(X=4)=p_2q_2, P(X=5)=p_3q_1, P(X=6)=p_3q_2 \end{cases} \tag{3.6}$$

这里，我们知道理论分布有(3.6)这种特殊形状，但并不完全知道，因为其中包含未知参数 p_1,p_2 和 q_1，其数目为 3.

这个例子代表了一类重要应用,将在下一段专门讨论.

现在我们可以提出一般的形式,设总体 X 只取有限个值 a_1, $\cdots, a_k$,其概率为

$$P(X = a_i) = p_i(\theta_1, \cdots, \theta_r), i = 1, \cdots, k \qquad (3.7)$$

其中 $\theta_1, \cdots, \theta_r$ 为未知参数,可在一定范围内变化.如在例 3.4 中,三个参数 p_1, p_2, q_1 的变化范围为

$$p_1 \geqslant 0, p_2 \geqslant 0, p_1 + p_2 \leqslant 1, 0 \leqslant q_1 \leqslant 1$$

参数个数 $r \leqslant k-2$.

设对 X 进行了 n 次观察,仍如前,以 ν_i 记 X 取 a_i 的次数.所要检验的假设是

$$H_0: (3.7) \text{ 对 } (\theta_1, \cdots, \theta_r) \text{ 的某一组值 } (\theta_1^0, \cdots, \theta_r^0) \text{ 成立} \qquad (3.8)$$

检验这个假设的步骤与前面相似,只多了一个参数估计问题:

1° 利用数据对参数 $\theta_1, \cdots, \theta_r$ 之值作一估计.采用极大似然估计法,即使(略去了与 $\theta_1, \cdots, \theta_r$ 无关的因子 $n! / (\nu_1! \cdots \nu_k!)$)

$$L = p_1^{\nu_1}(\theta_1, \cdots, \theta) \cdot p_2^{\nu_2}(\theta_1, \cdots, \theta_r) \cdots p_k^{\nu_k}(\theta_1, \cdots, \theta_r))$$

达到最大.取 $\log L$,对 θ_i 求偏导数并命之为 0,得

$$\sum_{j=1}^{k} \frac{\nu_j}{p_j(\theta_1, \cdots, \theta_r)} \frac{\partial p_j(\theta_1, \cdots, \theta_r)}{\partial \theta_i} = 0, i = 1, \cdots, r \qquad (3.9)$$

此方程组的解记为 $\hat{\theta}_1, \cdots, \hat{\theta}_r$.

2° 就以 $(\hat{\theta}_1, \cdots, \hat{\theta}_r)$ 作为 $(\theta_1, \cdots, \theta_r)$ 的真值.算出

$$p_i = p_i(\hat{\theta}_1, \cdots, \hat{\theta}_r), i = 1, \cdots, k$$

然后按公式(3.2)算出统计量 Z 之值.有如下的定理:

定理 3.2 在一定的条件下,若原假设(3.8)成立,则当样本大小 $n \to \infty$ 时,Z 的分布趋向于自由度 $k-1-r$ 的 χ^2 分布,即 χ^2_{k-1-r}.

这个定理是费歇尔在 1924 年证明的,其确切条件很复杂,不在此细述了.与皮尔逊定理 3.1 相比,差别在于自由度由 $k-1$ 下降为 $k-1-r$.即:所减少的自由度正好等于要估计的参数个数.

在这以前,皮尔逊曾认为这自由度仍为 $k-1$.

3° 据定理 3.2,若以 Z_0 记统计量 Z 的具体值,算出 Z_0 的拟合优度 $p(Z_0)=1-K_{k-1-r}(Z_0)$. 如给定检验水平 α,则当 $P(Z_0)<\alpha$ 时(即 $Z_0>\chi^2_{k-1-r}(\alpha)$时),否定 H_0.

在这几步中,最麻烦的往往是解方程组(3.9). 要计算 $p(Z_0)$,得有较细的 χ^2 分布表.

现在回到例 3.3. 调查了 n 个人,以 ν_1,ν_2 和 ν_3 分别记回答"甲优","乙优"和"认为一样或不回答"的人数. 例 3.3 已指出 $p_1(\theta)=p_2(\theta)=\theta, p_3(\theta)=1-2\theta$. 由此得出(3.9)为

$$(\nu_1+\nu_2)/\theta-2\nu_3/(1-2\theta)=0$$

其解为 $\hat{\theta}=(\nu_1+\nu_2)/2n$. 于是算出各类的理论值:

$$np_1(\hat{\theta})=(\nu_1+\nu_2)/2, np_2(\hat{\theta})=(\nu_1+\nu_2)/2, np_3(\hat{\theta})=\nu_3$$

因此

$$\begin{aligned} Z&=\left(\nu_1-\frac{\nu_1+\nu_2}{2}\right)^2\Big/\frac{\nu_1+\nu_2}{2}+\left(\nu_2-\frac{\nu_1+\nu_2}{2}\right)^2\Big/\frac{\nu_1+\nu_2}{2}\\ &=\frac{(\nu_1-\nu_2)^2}{\nu_1+\nu_2} \end{aligned} \tag{3.10}$$

此处 $k=3, r=1$,自由度为 $k-1-r=1$.

不难看出,(3.10)与用以下方法算出的 Z 一致:只考虑有效回答数 $N=\nu_1+\nu_2$. 把它作为一个 $k=2$. $p_1=p_2=1/2$ 的假设去检验. 事实上,按这个处理法,Z 值为

$$(\nu_1-N/2)^2/(N/2)+(\nu_2-N/2)^2/(N/2)=(\nu_1-\nu_2)^2/N$$

即(3.10). 按定理 3.1,当原假设 $p_1=p_2=1/2$ 成立时,其极限分布应为 $\chi^2_{2-1}=\chi^2_1$,与由定理 3.2 得出的一致,这个特例说明了:在有需要由数据估计的参数时,自由度确有所降低.

5.3.3 对列联表的应用

列联表是一种按两个属性作双向分类的表. 例如,一群人按男女(属性 A)和有否色盲(属性 B)分类,目的是考察性别对色盲有

无影响.属性也可以是在数量划分之下形成的.如在例 3.4 中属性 A——收入可按月 3000 元以上(高)、月 1000—3000 元(中)、月 1000 元以下(低)分为三档,如数据量大,档次还可以多分一些.

表 3.1 显示一个 $a \times b$ 双向列联表.属性 A 有 a 个水平 $1,2,\cdots,a$,B 有 b 个水平 $1,2,\cdots,b$.随机观察了 n 个个体,其中属性 A 处在水平 i,而属性 B 处在水平 j 的个体数,为表中之 n_{ij}.又

表 3.1　$a \times b$ 列联表

B \ A	1	2	$\cdots$	i	$\cdots$	a	和
1	n_{11}	n_{21}	$\cdots$	n_{i1}	$\cdots$	n_{a1}	$n_{\cdot 1}$
2	n_{12}	n_{22}	$\cdots$	n_{i2}	$\cdots$	n_{a2}	$n_{\cdot 2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	n_{1j}	n_{2j}	$\cdots$	n_{ij}	$\cdots$	n_{aj}	$n_{\cdot j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
b	n_{1b}	n_{2b}	$\cdots$	n_{ib}	$\cdots$	n_{ab}	$n_{\cdot b}$
和	$n_{1\cdot}$	$n_{2\cdot}$	$\cdots$	$n_{i\cdot}$	$\cdots$	$n_{a\cdot}$	n

$$n_{i\cdot} = \sum_{j=1}^{b} n_{ij}, n_{\cdot j} = \sum_{i=1}^{a} n_{ij} \tag{3.11}$$

分别是属性 A 处在水平 i 的个体数和属性 B 处在水平 j 的个体数.记

$$p_{ij} = P(\text{属性 } A, B \text{ 分别处在水平 } i, j)$$

问题是要检验 A,B 两属性独立的假设 H_0.如 H_0 为真,应有

$$p_{ij} = u_i v_j, i = 1,\cdots,a; j = 1,\cdots,b \tag{3.12}$$

其中

$$u_i = P(\text{属性 } A \text{ 有水平 } i), v_j = P(\text{属性 } B \text{ 有水平 } j)$$

因此,H_0 之成立,等价于存在 $\{u_i\}, \{v_j\}$,满足

$$u_i > 0, \sum_{i=1}^{a} u_i = 1; v_j > 0, \sum_{j=1}^{b} v_j = 1 \tag{3.13}$$

使(3.12)式成立.

在这个模型中,u_i, v_j 等充当了参数 $\theta_1,\cdots,\theta_r$ 的作用.总的独

立参数个数为

$$r = (a-1)+(b-1) = a+b-2$$

为估计 u_i, v_j，写出似然函数

$$L = \prod_{i=1}^{a}\prod_{j=1}^{b}(u_i v_j)^{n_{ij}} = \prod_{i=1}^{a} u_i^{n_{i\cdot}} \cdot \prod_{j=1}^{b} v_j^{n_{\cdot j}}$$

取对数

$$\log L = \sum_{i=1}^{a} n_{i\cdot} \cdot \log u_i + \sum_{j=1}^{b} n_{\cdot j} \log v_j$$

注意独立参数为 $u_1, \cdots, u_{a-1}$ 和 $v_1, \cdots, v_{b-1}$，而 $u_a = 1 - u_1 - \cdots - u_{a-1}$，$v_b = 1 - v_1 - \cdots - v_{b-1}$，故 $\dfrac{\partial u_a}{\partial u_i} = -1, i = 1, \cdots, a-1$，$\dfrac{\partial v_b}{\partial v_j} = -1, j = 1, \cdots, b-1$. 由此得方程

$$0 = \frac{\partial \log L}{\partial u_i} = \frac{n_{i\cdot}}{u_i} - \frac{n_{a\cdot}}{u_a}, i = 1, \cdots, a-1$$

$$0 = \frac{\partial \log L}{\partial v_j} = \frac{n_{\cdot j}}{v_j} - \frac{n_{\cdot b}}{v_b}, j = 1, \cdots, b-1$$

由这方程组，并利用(3.13)以及

$$\sum_{i=1}^{a} n_{i\cdot} = \sum_{j=1}^{b} n_{\cdot j} = n$$

即得解为

$$\hat{u}_i = n_{i\cdot}/n, i = 1, \cdots, a; \quad \hat{v}_j = n_{\cdot j}/n, j = 1, \cdots, b \tag{3.14}$$

其实，估计量(3.14)不是别的，正是用频率估计概率. 例如，$n_{i\cdot}$ 是在 n 个个体中，属性 A 取水平 i 的个体数，故 $n_{i\cdot}/n$ 正好是频率*.

由估计量(3.14)得 $\hat{p}_{ij} = \hat{u}_i \hat{v}_j = n_{i\cdot} n_{\cdot j}/n^2$，因而得到第 (i, j)

---

* $\hat{u}_i, \hat{v}_j$ 不允许为 0. 实际上，若某个 $n_{i\cdot} = 0$(因而 $\hat{u}_i = 0$,)则表 3.1 的第 i 列全为 0. 这时 A 的水平 i 应当划去. 这当然不是说 A 不能有水平 i，只是在样本中未出现，无法讨论，只能看作没有.

格的理论值为 $n\hat{p}_{ij}=n_{i\cdot}n_{\cdot j}/n$，因此统计量 Z 为

$$Z=\sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij}-n_{i\cdot}n_{\cdot j}/n)^2/(n_{i\cdot}n_{\cdot j}/n)$$

$$=\sum_{i=1}^{a}\sum_{j=1}^{b}(nn_{ij}-n_{i\cdot}n_{\cdot j})^2/(nn_{i\cdot}n_{\cdot j}) \tag{3.15}$$

自由度为 $k-1-r=ab-1-(a+b-2)=(a-1)(b-1)$.

在 $a=b=2$ 这个特例，表 3.1 有时也称为“四格表”. 简单的代数计算证明，这时有

$$Z=n(n_{11}n_{22}-n_{12}n_{21})^2/(n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}) \tag{3.16}$$

自由度为 1.

例 3.5 考虑例 3.4. 设随机从某特定一大群人中，调查了 201 名，结果如下表. 其中 A 表收入，1,2,3 分别表示低、中、高；B 表文化支出，1,2 分别表示“少”和“多”.

B \ A	1	2	3	和
1	63	37	60	160
2	16	17	8	41
和	79	54	68	201

须分别就每个格子计算和(3.15)中的项. 例如，第一个格子为

$$(201\times 63-79\times 160)^2/(201\times 160\times 79)=0.0002.$$

其他 5 个格子之值依次算出为 0.8333, 0.6367, 0.0008, 3.2521, 2.4847. 这 6 个数的和，即统计量 Z 的值 Z_0，为 7.2078，自由度为 $(3-1)(2-1)=2$. 查 χ^2 分布表，得拟合优度 $p(Z_0)=0.0207$. 此值很低，说明“收入与文化支出无关联”的假设极不可能成立. 考察所得数据，收入高者文化支出偏低.

例 3.6 有三个工厂生产同一种产品. 产品分 1,2,3 三个等级. 为考察各工厂产品质量水平是否一致，从这三个工厂中分别随机地抽出产品 109 件，100 件和 91 件，每件鉴定其质量等级，结果如下表.

等级 \ 工厂	1	2	3	和
1	58	38	32	138
2	28	44	45	117
3	23	18	14	55
和	109	100	91	300

“各工厂产品质量一致”这个假设,可看作是“工厂”和“质量等级”这两个属性独立的假设.用公式 (3.15),算出统计量 Z 之值 $Z_0=13.59$.自由度为$(3-1)(3-1)=4$.查 χ^2 分布表,得拟合优度为$p(Z_0)=1-K_4(13.59)<0.01$,故结果高度显著,即有明显证据说明各工厂产品质量并不一致.从表上数据看,1 厂质量明显优于另两厂,而 2,3 厂的差别似不大.

本例与例 3.5 比有一点不同.在例 3.5 中,每一个体抽出后,才去确定其两属性的水平,故表中边缘的数据,即 79,54,68,及 160,41,都是随机观察结果.本例则不然.三厂各自抽样数 109,100,91 等,在抽样前已定下,并非随机,每一个体在被抽出时,其 A 属性的水平事先已定(从第一厂抽的产品,事先就知其 A 属性的水平必为 1).虽有这个差别,但理论上可以证明:定理 3.2 仍然适用.

像例 3.6 这种检验问题常称为“齐一性检验”.因为,本例更自然的看法是把三个工厂的产品看成三个分别的总体,每总体依质量等级各有其分布,共有三个分布.检验的假设是“这三个分布一致”(或齐一).而像例 3.5 那种检验问题则称为“独立性检验”,其目的是判定两个属性有无关联存在.

5.3.4 总体分布为一般分布的情形

这包括总体分布为离散型,但能取无限多个值例如波哇松分布的情形,以及总体分布为连续型,例如正态分布的情形.设 X_1,$\cdots$,X_n 为自某总体中抽出的样本,要检验原假设

$$H_0:\text{总体分布为 } F(x) \tag{3.17}$$

其中 $F(x)$可以是完全已知，也可以带有未知参数，这时 $F(x)$成为 $F(x;\theta_1,\cdots,\theta_r)$. 其中$(\theta_1,\cdots,\theta_r)$可以在一定的范围内取值，而(3.17)则改为

$$H_0:\text{对其一组值}(\theta_1^0,\cdots,\theta_r^0),\text{总体分布为 } F(x;\theta_1^0,\cdots,\theta_r^0) \tag{3.18}$$

检验这一假设的办法是，通过区间分划把它转化为已讨论过的情况. 为确定计，设 F 是连续型的. 把$(-\infty,\infty)$分割为一些区间

$$-\infty = a_0 < a_1 < a_2 < \cdots < a_{k-1} < a_k = \infty$$

一共 k 个区间：$I_1=(a_0,a_1],\cdots,I_i=(a_{i-1},a_i],\cdots,I_k=(a_{k-1},a_k)$. 如果总体的分布为 $F(x;\theta_1,\cdots,\theta_r)$，则区间 I_i 有概率

$$p_i(\theta_1,\cdots,\theta_r)=F(a_i;\theta_1,\cdots,\theta_r) - F(a_{i-1};\theta_1,\cdots,\theta_r), i=1,\cdots,k \tag{3.19}$$

以 ν_i 记样本 $X_1,\cdots,X_n$ 中落在区间 I_i 内的个数，$i=1,\cdots,k$. 通过这个办法，我们就回到了在 5.3.2 段中已讨论过的情况，连记号 $p_i(\theta_1,\cdots,\theta_r)$，$\nu_i$ 也一样. 以下的步骤就与那里讲的完全一样，基于定理 3.2，拟合优度统计量 Z 的极限分布为 χ^2_{k-1-r}，故分区间的数目 k 不能小于 $r+2$.

当然，通过分区间，我们实际上是用另外一个假设 H_0' 代替了原来的假设(3.18). H_0' 是："对某一组值$(\theta_1^0,\cdots,\theta_r^0)$，总体在区间 I_i 内的概率为 $p_i(\theta_1^0,\cdots,\theta_r^0)$，$i=1,\cdots,k$". 若(3.18)成立，$H_0'$ 当然成立. 反之，由 H_0' 成立推不出(3.18)成立，因为 H_0' 丝毫没有限制总体在每个区间 I_i 内的分布如何. 所以如否定了 H_0'，则更有理由否定 H_0. 若接受 H_0'，则我们也接受 H_0——这方法就是如此规定的. 可以设想，若区间分得很细，则每个小区间 I_i 内的概率都不大，H_0' 与 H_0 之间也就更接近，但是，分区间数 k 取决于样本大小 n. 为了使定理 3.1 或 3.2 中的极限分布与 Z 的确切分布的差距缩小，就要求分区间数少些，以使每区间内样本数目(即 ν_i)大一些，这是两个互相矛盾的要求，在实际工作中，通常是根据样本值

的情况来划分区间*，以使每个区间内所含样本数不小于5，而区间数 k 又不要太大或太小. 一般在 $40\leqslant n\leqslant 100$ 时，区间数可取为6至8；当 $100\leqslant n\leqslant 200$ 时，可取为 $9,\cdots,12$. 当 $n>200$ 时可适当增加，一般以不超过15—20为宜. 这样划分时，有时不能照顾到各区间(除 I_1 和 I_k 外)之长相等.

对总体为离散的情况，设它能取的值按大小排列为 $a_1<a_2<\cdots$. 若样本 $X_1,\cdots,X_n$ 中有较多个(例如至少5个以上)取 a_i 为值，则 a_i 自成一组. 若不然，则把相邻的几个 a_i 并成一组，分组数目的考虑与上述相同.

这个检验中最难的一部分就是计算出 $\theta_1,\cdots,\theta_r$ 的估计值 $\hat{\theta}_1,\cdots,\hat{\theta}_r$. 这要通过解方程组(3.9)，其中 $p_i(\theta_1,\cdots,\theta_r)$ 由(3.19)给出. 这种方程确切解的计算很难. 例如，若要检验总体分布为正态 $N(\mu,\sigma^2)$，则 $r=2,\theta_1=\mu,\theta_2=\sigma$. 而

$$p_i(\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\int_{a_{i-1}}^{a_i}\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]\mathrm{d}x$$

要把这样的表达式代入(3.9)而求解是很难的，因此在应用上，常使用更易于计算的估计，如用

$$\hat{\mu}=\overline{X},\hat{\sigma}=S$$

其中 $\overline{X}$ 和 S 分别是样本均值和样本方差. 理论上知道，用这一估计代替由(3.9)决定的估计去计算统计量 Z，已使定理3.2的结论不成立了，但差距还不大，故应用上还是可以.

以下这两个数字例子取自H. 克拉美的《统计数学方法》第30章.

例3.7 有一取0,1,2…为值的离散变量，对其进行了2608次观察，结果如下表所示：

---

* 按理论的要求(为了使定理3.1,3.2的结论有效)，划分区间必须在未看到样本之前就做好，而不能依样本情况去划分. 但实际工作中难于遵守这一点，它引起的误差一般也很小，不必拘泥.

i	0	1	2	3	4	5	6	7	8	9	(10	11	12)
ν_i	57	203	383	525	532	408	273	139	45	27	(10	4	2)

要检验其分布为波哇松分布的假设.

先是分组.对 $i=0,1,\cdots 9$,ν_i 都比较大,可单独成组.10,11 和 12 合并为一组,故该组的 ν_i,应改为 $10+4+2=16$.

其次是用样本估计波哇松分布的参数 λ.要是用(3.9),则甚为麻烦.此处用其通常估计 $\overline{X}$:

$$\hat{\lambda} = \overline{X} = 3.870$$

然后据此算出各组理论值.除最后一组外,理论值是

$$n\mathrm{e}^{-\hat{\lambda}}\hat{\lambda}^i/i! = 2608\mathrm{e}^{-3.870}(3.870)^i/i!, i = 0,1,\cdots,9$$

例如,算出 $i=0$ 时为 54.399,$i=1$ 时为 210.523 等.最后一组的理论值为

$$2608\sum_{i=10}^{12}\mathrm{e}^{-3.870}(3.870)^i/i! = 17.075$$

最后按公式(3.2)算出统计量 Z 之值,结果为 $Z_0=12.885$.此处 $k=11$(共分 11 个组),$r=1$(有一个参数 λ 被估计),故自由度为 $11-1-1=9$.查 χ^2 分布表,得拟合优度为 $p(Z_0)=1-K_9(12.885)=0.17$.这拟合优度尚可,但不太好:即使总体真服从波哇松分布,也有 17% 的机会产生比本例数据更大的偏离.0.17 概率的事件当然不稀奇.但这概率毕竟偏小一些,使人不很放心.

例 3.8 瑞典斯德哥尔摩自 1841 年至 1940 年,百年期间 6 月份平均温度的记录,分组后如下表.要检验这温度的分布服从正态分布 $N(\mu,\sigma^2)$对某个(μ,σ^2).

区 间 (摄氏度)	观察数 ν_i	区 间 (摄氏度)	观察数 ν_i
—12.4	10	14.5—14.9	10
12.5—12.9	12	15.0—15.4	9
13.0—13.4	9	15.5—16.0	6
13.5—13.9	10	16.0—16.4	7
14.0—14.4	19	16.5—	8

克拉美给出的 μ 和 σ 估计值是 $\hat{\mu}=14.28$，$\hat{\sigma}=1.574$. 利用这组估计就可以算出各组的理论值. 例如，12.5—12.9 这一组是

$$100\frac{1}{\sqrt{2\pi}1.574}\int_{12.45}^{12.95}\exp\left[-\frac{(x-14.28)^2}{2\times(1.574)^2}\right]dx$$

$$=100\times0.0789=7.89$$

而—12.4 这一组为

$$100\frac{1}{\sqrt{2\pi}1.574}\int_{-\infty}^{12.45}\exp\left[-\frac{(x-14.28)^2}{2\times(1.574)^2}\right]dx$$

$$=100\times0.1289=12.89$$

等等，式中的积分可通过转化到标准正态分布函数去计算：

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{a}^{0}\exp\left[-\frac{1}{2\sigma^2}(X-\mu)^2\right]dx$$

$$=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)$$

查标准正态分布表即得.

注意以上计算中的积分限，它是取在相邻区间的相邻端点的中点，这符合四舍五入法则.

这样算出各组理论值后，用(3.2)算出 Z 值. 本例结果为 7.86. 自由度 $k-1-r=10-1-2=7$. 拟合优度为 $1-K_7(7.86)=0.85$. 拟合程度很高.

如果数据一开始就用分组形式给出(原始数据没有给，或最初记录时就只记下它在何区间)，则 $\hat{\mu}$ 和 $\hat{\sigma}$ 只能用这分组数据算. 可用公式

$$\hat{\mu}=\frac{1}{n}\sum m_i\nu_i,\ \hat{\sigma}^2=\frac{1}{n}\sum\nu_i(m_i-\hat{\mu})^2$$

其中 m 是第 i 个组区间之中点. 这时，最左最右两个区间也要界定，可取其长为其相邻区间之长.

最后，如果理论分布 F 不包含参数，则各区间理论值直接由 $n[F(a_i)-F(a_{i-1})]$ 算出，一切简单得多. 自由度是 $k-1$，k 为分区间数目.