§6.5 分布拟合检验

判断总体是否为某种分布（如正态分布）的检验问题，通称为分布的拟合优度检验，简称为分布拟合检验.

一、总体分布只取有限个值的情况

设总体X可以分成k类，记为$A_1, A_2, \cdots, A_k$，现对该总体作了n次观测，k个类出现的频数分别为：

$$n_1, \cdots, n_k, \text{且} \sum_{i=1}^{k} n_i = n.$$

检验如下假设：

$$H_0: P(A_i) = p_i, \Leftrightarrow H_1: \text{某些} P(A_i) \neq p_i, i = 1, \cdots, k.$$

其中诸$p_i \geq 0$且$\sum_{i=1}^{k} p_i = 1.$

情形1　诸p_i均已知

如果H_0成立，则对每一类A_i，其频率n_i/n与概率p_i应较接近，即观测频数n_i与理论频数np_i应相差不大. 据此，英国统计学家**K.Pearson**提出如下检验统计量：

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}.$$

并证明在H_0成立时对充分大的n，上述检验统计量近似服从自由度为$k-1$的χ^2分布. 拒绝域为：

$$W=\left\{\chi^2>\chi_\alpha^2(k-1)\right\}$$

例1 为募集社会福利基金，某地方政府发行福利彩票, 中彩者用摇大转盘的方法确定最后中奖金额大转盘均分为20份, 其中金额为5万、10万、20万、30万、50万、100万的分别占2份、4份、6份、4份、2份、2份. 假定大转盘是均匀的，则每一点朝下是等可能的，于是摇出各个奖项的概率如下：

额度	5万	10万	20万	30万	50万	100万
概率	0.1	0.2	0.3	0.2	0.1	0.1

现**20**人参加摇奖，摇得**5**万、**10**万、**20**万、**30**万、**50**万、**100**万的人数分别为**2**、**6**、**6**、**3**、**3**、**0**，由于没有一个人摇到**100**万, 于是有人怀疑大转盘是不均匀的，那么该怀疑是否成立？

解 这是一个典型的分布拟合优度检验, 总体共有**6**类，发生概率分别为**0.1**、**0.2**、**0.3**、**0.2**、**0.1**和**0.1**. 这里$k=6$，检验拒绝域为：

$$\left\{\chi^2 \geq \chi_\alpha^2(5)\right\},$$

若取$\alpha = 0.05$，查表知$\chi_{0.05}^2(5) = 11.07$. 由数据得到

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = 3.75 < \chi_{0.05}^2(5)$$

故接受原假设，没有理由认为转盘不均匀.

在分布拟合检验中使用p值也是方便的. 使用统计软件可以算出

$$p = P(\chi^2(5) \geq 3.75) = 0.5859.$$

这个p值就反映了数据与假设的分布拟合程度的高低，p值越大，拟合越好.

情形2　诸p_i不完全已知

若诸 $p_i, i=1,\cdots,k$由$r(r<k)$个未知参数$\theta_1,\cdots,\theta_r$确定. 即

$$p_i = p_i(\theta_1,\cdots,\theta_r), \quad i=1,\cdots,k.$$

首先给出$\theta_1,\cdots,\theta_r$的极大似然估计$\hat{\theta}_1,\cdots,\hat{\theta}_r$, 然后给出诸 p_i 的极大似然估计$\hat{p}_i = p_i(\hat{\theta}_1,\cdots,\hat{\theta}_r)$.

Fisher证明了 $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$在$H_0$成立时, 近似服从自由度为$k-r-1$的$\chi^2$分布，故检验拒绝域为

$$\left\{\chi^2 > \chi_\alpha^2(k-r-1)\right\}$$

例2 卢瑟福在2608个等时间间隔内观测一枚放射性物质放射的粒子数X,下表是观测结果的汇总,其中n_i表示2608次观测中放射粒子数为 i 的次数.

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

试利用该组数据检验该放射物质在单位时间内放射出的粒子数是否服从泊松分布.

解：本例中, 要检验总体是否服从泊松分布. 观测到0, 1, …, 11共12个不同取值，这相当于把总体分成12类. 这里有一个未知参数λ，采用极大似然估计，

$$\hat{\lambda}=\frac{1}{2608}(1\times 203+2\times 383+...+11\times 6)=3.87.$$

将$\hat{\lambda}$代入可以估计出诸 $\hat{p}_i$.见下表

若取$\alpha=0.05$，则$\chi_{\alpha}^{2}(k-r-1)=\chi_{0.05}^{2}(10)=18.307.$

本例中 $\chi^2=12.8967<18.307$，故接受原假设.

使用统计软件可以算出检验的p值是0.2295.

列表如下：

i	n_i	$\hat{p}_i$	$n\hat{p}_i$	$\left(n_i - n\hat{p}_i\right)^2 / n\hat{p}_i$
0	57	0.0209	54.5	0.1147
1	203	0.0807	210.5	0.2672
2	383	0.1562	407.4	1.4614
3	525	0.2015	525.5	0.0005
4	532	0.1950	508.6	1.0766
5	408	0.1509	393.5	0.5343
6	273	0.0973	253.8	1.4525
7	139	0.0538	140.3	0.0120
8	45	0.0260	67.8	7.6673
9	27	0.0112	29.2	0.1658
10	10	0.0043	11.2	0.1258
11	6	0.0022	5.7	0.0158
合计	2608	1.0000	2068	χ^2=12.8967

二、连续型分布的拟合检验

基本思想：为了检验随机变量X是否服从连续型分布$F_0(x)$，可将X的取值范围分割为若干个区间，并在每个区间上算出相应的理论概率p_{i_0}.

通过这样处理, 把连续型问题转化为离散型问题.

具体步骤：

(1)分组：将$F_0(x)$的自变量分成 k 组

$$(b_0, b_1], (b_1, b_2], \cdots, (b_{k-1}, b_k].$$

(每区间至少有5个样本, 否则并入邻区间)

(2)求各组上的理论概率p_{i_0}及理论频数np_{i_0}：

$$p_{i_0} = P\{b_{i-1} < X \le b_i\} = F_0(b_i) - F_0(b_{i-1}).$$

(3) 计算统计量 $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i_0})^2}{np_{i_0}}$;

(4) 判断: 若 $\chi^2 > \chi_\alpha^2(k-1)$, 则拒绝 H_0: $F(x) = F_0(x)$.

注：若 $F_0(x)$ 中有 r 个待估参数，则首先估计参数. 最后判断时，统计量的自由度降低 r .

例4 测量了100根人造纤维的长度(毫米)，所得的数据如下表：

长度	5.5~6.0	~6.5	~7.0	~7.5	~8.0	~8.5	~9.0	~9.5	~10	~10.5	~11
频数	2	7	6	17	17	14	16	10	7	3	1

问：能认为人造纤维的长度服从正态分布吗？

组号	b_i	u_i	$F_0(b_i)$	p_{i0}	$n\,p_{i0}$	n_i
1	6.0	-1.89	0.0294	0.0294	2.94	2
2	6.5	-1.44	0.0749	0.0454	4.54	7
3	7.0	-0.98	0.1635	0.0886	8.86	6
4	7.5	-0.53	0.2981	0.1346	13.46	17
5	8.0	-0.08	0.4681	0.1700	17.00	17
6	8.5	0.37	0.6443	0.1762	17.62	14
7	9.0	0.83	0.7967	0.1524	15.24	16
8	9.5	1.28	0.8997	0.1030	10.30	10
9	10	1.74	0.9591	0.0594	5.94	7
10	10.5	2.19	0.9857	0.0166	1.66	3
11	+∞	+∞	1	0.0143	1.43	1

（略）

三、列联表的独立性检验

列联表是将观测数据按两个或更多属性(定性变量)分类时所列出的频数表。例如，对随机抽取的1000人按性别（男或女）及色觉(正常或色盲)两个属性分类，得到如下二维列联表，又称2×2表或四格表.

性别	视觉	
	正常	色盲
男	535	65
女	382	18

一般,若总体中的个体可按两个属性A与B分类，A有r个类$A_1,\cdots,A_r$.B有s个类从总体中抽取大小为n的样本，设其中有n_{ij}个个体既属于A_i类又属于B_j类，称为频数, 将$r\times s$个n_{ij}排列为一个r行s列的二维列联表, 简称$r\times s$表.

$A\backslash B$	1	$\cdots$	j	$\cdots$	s	和
1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$n_{r\cdot}$
和	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot s}$	n

列联表分析的基本问题是：考察各属性之间有无关联，即判别两属性是否独立.

如在前例中，问题是：一个人是否色盲与其性别是否有关？在$r \times s$表中，若以$p_{i\cdot}$、$p_{\cdot j}$和p_{ij}分别表示总体中的个体仅属于A_i ,仅属于B_j和同时属于A_i?与B_j的概率，可得一个二维离散分布表，则“A、B两属性独立”的假设可以表述为

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}, \qquad i = 1, \cdots, r, \ \ j = 1, \cdots, s$$

这就变为上一小节中诸p_{ij}不完全已知时的分布拟合检验. 这里诸p_{ij}共有rs个参数，在原假设H_0成

立时，这rs个参数由$r+s$个参数$p_{1\cdot},\cdots,p_{r\cdot}$和$p_{\cdot 1},\cdots,p_{\cdot s}$决定. 在这后$r+s$个参数中存在两个约束条件：

表　二维离散分布表

$A\backslash B$	1	$\cdots$	j	$\cdots$	s	行和
1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1s}	$p_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{is}	$p_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	p_{r1}	$\cdots$	p_{rj}	$\cdots$	p_{rs}	$p_{r\cdot}$
列和	$p_{\cdot 1}$	$\cdots$	$p_{\cdot j}$	$\cdots$	$p_{\cdot s}$	1

$$\sum_{i=1}^{r} p_{i\cdot} = 1, \quad \sum_{j=1}^{s} p_{\cdot j} = 1.$$

所以，此时p_{ij}实际上由$r+s-2$个独立参数所确定.

总体共分为rs类，因此检验统计量为

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}.$$

H_0成立时，χ^2服从自由度为$rs-(r+s-2)-1$的χ^2分布. 其中诸$\hat{p}_{ij}$是在H_0成立下p_{ij}的极大似然估计,

$$\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}.$$

对给定的显著性水平α，检验的拒绝域为

$$W = \{\chi^2 > \chi_\alpha^2((r-1)(s-1))\}.$$

例：为研究儿童智力发展与营养的关系，某研究机构调查了**1436**名儿童, 得到如下表的数据, 试在显著性水平**0.05**下，判断智力发展与营养有无关系.

表　儿童智力与营养的调查数据

	智 商				合计
	<80	80~90	90~99	≥100	
营养良好	367	342	266	329	1304
营养不良	56	40	20	16	132
合计	423	382	286	345	1436

解：用A表示营养状况，它有两个水平：A_1表示营养良好，A_2表示营养不良；B表示儿童智商，它有四个水平，B_1, B_2, B_3, B_4分别表示表中四种情况，沿用前面的记号，首先建立假设H_0：营养状况与智商无关联，即A与B独立.

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}, \quad i = 1,2, \quad j = 1,2,3,4.$$

在H_0成立条件下，计算诸参数的极大似然估计值：

$$\hat{p}_{1\cdot} = 1304/1436 = 0.9081, \quad \hat{p}_{2\cdot} = 132/1436 = 0.0919,$$

$$\hat{p}_{\cdot 1} = 423/1436 = 0.2946, \quad \hat{p}_{\cdot 2} = 382/1436 = 0.2660,$$

$$\hat{p}_{\cdot 3} = 286/1436 = 0.1992, \quad \hat{p}_{\cdot 4} = 345/1436 = 0.2403,$$

进而可给出诸$n\hat{p}_{ij} = n\hat{p}_{i\cdot}\hat{p}_{\cdot j}$，列表如下

	<80	80~90	90~99	≥100	$\hat{p}_{i.}$
营养良好	384.1677	346.8724	259.7631	313.3588	0.9081
营养不良	38.8779	35.1036	26.2881	31.7120	0.0919
$p_{.j}$	0.2946	0.2660	0.1992	0.2403	

由表可以计算检验统计量的值

$$\chi^2 = \sum\sum (n_{ij} - n\hat{p}_{ij})^2 \Big/ n\hat{p}_{ij} = 19.2785.$$

取$\alpha = 0.05$，有$\chi^2_{0.05}(3) = 7.815$，.

由于$19.2785 > 7.815$，故拒绝H_0, 认为营养状况对智商有影响. p值为0.0002。