第五章　非参数假设检验

非参数假设检验

若假设$H_0: F(x) = F_0(x;\theta)$, 其中$F_0(x;\theta)$为一个指定的分布,$\theta$是参数向量.
(1)若$\theta = \theta_0$已知, 即总体分布完全确定,此时H_0称为**简单假设**.
(2)若θ部分或完全未知,即$F_0(x;\theta)$形式上确定,但含有未知参数,此时H_0称为**复合假设**.

当假定某一理论分布$F_0(x;\theta)$,实际数据$x_1, \cdots, x_n$与理论分布$F_0(x;\theta)$偏离的量用$\Delta(x_1, \cdots, x_n; F)$表示, 规定一个界限$\Delta_0$,若$\Delta$超过这个界限$\Delta_0$,则认为理论分布与数据$x_1, \cdots, x_n$不符,因而拒绝$H_0$,否则接受$H_0$. 这个$\Delta$就称为**“拟合优度”**(Goodness of Fit), 这种检验称为**“拟合优度检验”**

χ^2拟和优度检验

一.χ^2拟和优度检验

1.简单假设

设总体分布为$F(x)$.$(X_1, \cdots, X_n)$为取自总体的样本,提出假设$H_0 : F(x) = F_0(x; \theta_0)$, θ_0为已知参数。

(1)选取$r-1$个实数$-\infty < y_1 < y_2 < \cdots < y_{r-1} < +\infty$,它们将随机变量$X$的一切可能取值的集合分为$r$个区间,并用$n_i$表示样本观测值落入第$i$个区间$(y_{i-1}, y_i]$的**观测频数** (这里设$y_0 = -\infty$, $y_r = +\infty$). $\sum_{i=1}^{r} n_i = n$.

(2)在H_0为真下,则由给定的分布函数$F_0(x; \theta_0)$可以求出$p_i = F_0(y_i; \theta_0) - F_0(y_{i-1}; \theta_0)$, $i = 1, 2, \cdots, r$.

$\sum_{i=1}^{r} p_i = 1$,称np_i为样本落入第i个小区间的**理论频数**.

χ^2拟和优度检验

(3)考虑统计量$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$，它表示实际观测频数$n_i$与理论频数$np_i$的相对差异的总和. 由Pearson定理:$\chi^2 \sim \chi^2(r-1)$,因此当$n$充分大时,对给定的显著性水平$\alpha$,检验的拒绝域为$W = \{\chi^2 \geq \chi^2_{1-\alpha}(r-1)\}$.

χ^2拟和优度检验

2.复合假设

$H_0 : F(x) = F_0(x;\theta)$,$\theta$未知,为s维向量,$\theta = (\theta_1, \cdots, \theta_s)$

(1)由MLE法得$\hat{\theta}$ 代替未知参数,

(2)$\hat{p}_i = F_0(y_i;\hat{\theta}) - F_0(y_{i-1};\hat{\theta}), i = 1, 2, \cdots, r$

(3)$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} \sim \chi^2(r - s - 1)$

χ^2拟和优度检验

3.假设检验步骤

(1)将观测值(数据)分为r个互不相容的区间，算出n_i,每个区间至少有5个样本，区间长度可以不一样。

(2)在H_0为真下，用MLE估计法去估计分布中所含的未知参数。

(3)在H_0为真下，计算理论概率$p_i = F(y_i) - F(y_{i-1}), i = 1, \cdots, r$,并计算$np_i$.

(4)计算$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$

(5)对给定的显著性水平α,检验的拒绝域为$W = \{\chi^2 \geq \chi^2_{1-\alpha}(r-s-1)\}$，查表得临界值，s为未知参数的个数。

(6)若$\chi^2 > \chi^2_{1-\alpha}$,则拒绝$H_0$,否则接受$H_0$。

χ^2拟和优度检验

注：(1)χ^2拟和优度检验必需在大样本下进行。
(2)要求$np_i \geq 5$
(3)在简单假设检验中，分区间时最好各区间概率相同。

例1. 有一正二十面体的20个面上分别标以数字$0, 1, \cdots, 9$,每个数字在两个对称的面上标出，为检验其均匀性，共做800次投掷试验，数字$0, 1, \cdots, 9$朝上的次数为

数字	0	1	2	3	4	5	6	7	8	9
频数	74	92	83	79	80	73	77	75	76	91

问：该正20面体是否均匀？$(\alpha = 0.05)$

χ^2拟和优度检验

解: H_0:该正20面体均匀,
即$p_i = P(x = i) = \frac{1}{10}, i = 0, 1, \cdots, 9$,则$np_i = 80$

数字	n_i	np_i	$n_i - np_i$	$(n_i - np_i)^2$
0	74	80	-6	36
1	92	80	12	144
2	83	80	3	9
3	79	80	-1	1
4	80	80	0	0
5	73	80	-7	47
6	77	80	-3	9
7	75	80	-5	25
8	76	80	-4	16
9	91	80	11	121
$\sum$	800			410

$$\chi^2 = \sum_{i=1}^{10} \frac{(n_i - np_i)^2}{np_i} = \frac{1}{80} * 410 = 5.125,$$

对$\alpha = 0.05$,查表$\chi^2_{0.95}(10 - 1) = 16.9$,因为$\chi^2 < \chi^2_{0.95}(9)$,所以接受$H_0$,即认为该正20面体是均匀的。

χ^2拟和优度检验

例2. 遗传学中常常有考虑拟合检验的例子.例如某种动物身上的毛可分成三种类型:极卷,中等卷曲,正常,而毛的卷曲由二个遗传基因F, f所控制,(F, F)的后代身上的毛是极卷的,(F, f)的后代是中等卷曲,(f, f)则为正常,并且两个基因随机结合,因此极卷,中等卷曲,正常的比例应是$1:2:1$.现在进行了93次试验,所得下面结果.

极卷	中等卷曲	正常
23	50	20

设$p_1 = P\{\text{后代的毛是极卷的}\}$, $p_2 = P\{\text{后代的毛中等卷曲}\}$, $p_3 = P\{\text{后代的毛正常}\}$,

χ^2拟和优度检验

则假设检验为

$$H_0: \ p_1 = p_{10} = \frac{1}{4}, \ p_2 = p_{20} = \frac{1}{2}, \ p_3 = p_{30} = \frac{1}{4}$$
$$\leftrightarrow \quad H_1: \ \text{至少有一个} p_i \neq p_{i_0}.$$

$$\chi^2 = \frac{(23 - 93 \times \frac{1}{4})^2}{93 \times \frac{1}{4}} + \frac{(50 - 93 \times \frac{1}{2})^2}{93 \times \frac{1}{2}} + \frac{(20 - 93 \times \frac{1}{4})^2}{93 \times \frac{1}{4}} = 0.72,$$

对$\alpha = 0.05$, $\chi^2_{0.95}(2) = 5.991$ $\chi^2 = 0.72 < \chi^2_{0.95}(2)$,因此不能拒绝原假设,即毛的卷曲程度是由遗传基因$(F, F)$, (F, f)和(f, f)所控制的遗传学理论是站得住脚的.

χ^2拟和优度检验

例3.电话交换台在某一小时内接到用户的呼唤次数，按每分钟计

呼叫次数n_i	0	1	2	3	4	5	6	≥ 7
频数n_i	8	16	17	10	6	2	1	0

问：电话每分钟呼叫次数是否服从泊松分布？($\alpha = 0.05$)

解：由MLE得$\hat{\lambda} = \bar{X}$,由样本观测值得

$$\bar{x} = (0 * 8 + 1 * 16 + \cdots + 6 * 1)/60 = 2$$

$H_0 : \hat{p}_i = P(X = i) = \frac{2^i e^{-2}}{i!}, i = 0, 1, \cdots$
在H_0为真下，每分钟接到呼唤次数的理论频数

$$n\hat{p}_i = 60 * \frac{2^i e^{-2}}{i!}, i = 0, 1, \cdots$$

χ^2拟和优度检验

i	n_i	$n\hat{p}_i$	$n_i - n\hat{p}_i$	$(n_i - n\hat{p}_i)^2$	$(n_i - n\hat{p}_i)^2/n\hat{p}_i$
0	8	8.1204	-0.1204	0.0145	0.0018
1	16	16.2402	-0.2402	0.0577	0.0036
2	17	16.2402	0.7598	0.5773	0.0355
3	10	10.8264	-0.8264	0.6829	0.0631
4	6	5.4134			
5	2	2.1654	0.4932	0.2432	0.0286
6	1	0.7218			
7	0	0.2062			
$\sum$	60				0.1326

$\chi^2 = \sum_{i=1}^{5} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} = 0.133, r = 5, s = 1$,对$\alpha = 0.05$,查表$\chi^2_{0.95}(3) = 7.815$,因为$\chi^2 < \chi^2_{0.95}(3)$,所以接受$H_0$,即认为每分钟呼叫次数服从参数为2的泊松分布。

列联表的独立性检验

二.列联表的独立性检验

“对所考察的总体中每一个元素同时测定两个指标X, Y,要检验这两个指标是否有关.”

例:考虑对某种疾病的几种治疗方法与治疗结果之间的关系.
将n个病人按不同的治疗方法(第一个指标)分组,观察各组内病人的不同效果(第二个指标).设X可能取值为$1, 2, \cdots, p$, Y可能取值为$1, 2, \cdots, q$, 现在对(X, Y) 进行了n次独立观测而得$(x_1, y_1), \cdots, (x_n, y_n)$, 用$n_{ij}$表示样本观测值中“X取$i$, Y取j”的次数. 检验假设

$$H_0 : X\text{与}Y\text{独立}.$$

把数据排列成表的形式,这种表称为p*q**列联表**

列联表的独立性检验

X \ Y	1	2	$\cdots$	q	
1	n_{11}	n_{12}	$\cdots$	n_{1p}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2p}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
p	n_{p1}	n_{p2}	$\cdots$	n_{pq}	$n_{p\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot q}$	n

其中$n_{i\cdot}=\sum_{j=1}^{q} n_{ij}$, $n_{\cdot j}=\sum_{i=1}^{p} n_{ij}$, $n=\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}$.

检验X与Y是否相互独立$\Leftrightarrow$ $H_0: p_{ij}=p_{i\cdot}\cdot p_{\cdot j}$,对于所有$(i,j)$ 都成立;$H_1: p_{ij}\neq p_{i\cdot}\cdot p_{\cdot j}$,对于某个$(i,j)$成立

列联表的独立性检验

由MLE得

$$\begin{cases} \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}, & i = 1, \cdots, p; \\ \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}, & j = 1, \cdots, q. \end{cases}$$

选取统计量

$$\chi^2 = n \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - \hat{n}_{ij})^2}{n_{i\cdot} n_{\cdot j}} = n\left(\sum_{i=1}^{p} \sum_{j=1}^{q} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1\right) \sim \chi^2((p-1)(q-1))$$

其中$\hat{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$，
对给定的α，拒绝域$W = \{\chi^2 \geq \chi^2_{1-\alpha}((p-1)(q-1))\}$。

列联表的独立性检验

例5. 某校甲乙两班进行某种技能训练，测验成绩按优，良，及格及不及格四级给分，结果如下表，问成绩与班级有无关系？$\alpha = 0.05$

班级	优	良	及格	不及格	合计
甲	14	20	15	11	60
乙	18	10	20	12	60
合计	32	30	35	23	120

列联表的独立性检验

解:H_0:成绩与班级无关。
在H_0为真下，理论频数表如下，$\hat{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$

班级	优	良	及格	不及格	合计
甲	16	15	17.5	11.5	60
乙	16	15	17.5	11.5	60
合计	32	30	35	23	120

$$\chi^2 = n\sum_{i=1}^{2}\sum_{j=1}^{4}\frac{(n_{ij}-\hat{n}_{ij})^2}{n_{i\cdot} n_{\cdot j}}$$

$$= \frac{(14-16)^2}{16} + \frac{(20-15)^2}{15} \cdots \frac{(12-11.5)^2}{11.5} = 4.592$$

对$\alpha = 0.05$，$p = 2, q = 4$,查表
得$\chi^2_{0.95}(3) = 7.815$,则$w = \{\chi^2 > 7.815\}$,由$\chi^2 < \chi^2_{0.95}(3)$,所以接受$H_0$,即在显著性水平0.05下，认为成绩与班级无关。

列联表的独立性检验

注: (1)列联表检验独立性时，实际上是χ^2拟和优度检验中χ^2检验统计量极限定理的应用。

(2)也可用于连续型，将变量值分成若干个互不相容的区间。

(3)当$p = q = 2$时，得到2*2列联表，也叫四格表，用a, b, c, d表示观测值。

	1	2	$\sum$
1	a	b	a+b
2	c	d	c+d
$\sum$	a+c	b+d	n

有一简便的计算公式，

$$\chi^2 = \frac{n(ad - bc)^2}{(a + c)(b + d)(c + d)(a + b)} \sim \chi^2(1)$$

列联表的独立性检验

例6.调查339名50岁以上吸烟习惯与患慢性气管炎的关系，得下表，问吸烟与患慢性气管炎是否有关？$\alpha = 0.01$

	患	不患	$\sum$
吸烟	43	162	205
不吸烟	13	121	134
$\sum$	56	283	339

解:设H_0 :吸烟与患慢性气管炎无关，由四格表检验

$$\chi^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(c+d)(a+b)} = 7.469$$

对$\alpha = 0.01$，查表$\chi^2_{0.99}(1) = 6.635$,因为$\chi^2 > \chi^2_{0.99}(1)$,所以拒绝$H_0$,即认为吸烟与患慢性气管炎有关。