第六章非参数假设检验

§6.4 χ^2拟合优度检验

这是一种非参数的检验方法.

这一方法是在分类数据的检验问题中提出的,是由英国统计学家K.Pearson在1900年首次提出的.

假定一个总体的取值可以分成r类, 现从该总体获得了一组样本, 这是一批分类数据, 现在需要我们从这些分类数据出发, 去判断各类出现的概率是否与已知的概率相符.

Example

某工厂制造骰子, 声称它是均匀的. 为了检验骰子的均匀性, 抽取了部分骰子进行投掷. 如: 随机抽取了100颗, 各投掷一次, 这样就得到了容量为100的一组样本, 此样本就是一批分类数据.

该应用问题是检验H_0 : 骰子是均匀的$\leftrightarrow$ H_1 : 骰子是不均匀的
(有时写成H_0 : 骰子是均匀的)

用A_i表示"抛出的点数为i". 即需要检验

$$H_0 : \mathsf{P}(A_i) = \frac{1}{6}, \quad i = 1, 2, \cdots, 6.$$

Example

某大公司的人事部门希望了解公司职工的病假是否均匀分布在周一至周五,以便合理安排工作. 如今抽取了100病假职工, 其病假日分布如下:

工作日	周一	周二	周三	周四	周五
病假人数	17	27	10	28	18

试问: 该公司职工的病假是否均匀分布在一周五个工作日中($\alpha = 0.05$)?

用A_i表示"病假在周i". 则需要检验

$$H_0 : \mathsf{P}(A_i) = \frac{1}{5}, \quad i = 1, 2, \cdots, 5.$$

一、分类数据的χ^2检验

1. 总体可分成有限个类, 总体理论分布不含未知参数

设总体X可以分成r类: $A_1, A_2, \cdots, A_r$, 要检验的假设为

$$H_0 : \mathsf{P}(A_i) = p_i, \quad i = 1, 2, \cdots, r.$$

其中各个p_i已知, 且$p_i \geq 0$, $\sum_{i=1}^{r} p_i = 1$.
现对总体作了n次观察, 各类出现的频数分别为n_1, n_2, $\cdots$, n_r, $\sum_{i=1}^{r} n_i = n$.

如果H_0为真, 则各个概率p_i与频率n_i/n应该相差不大, 即, 各个观察频数(即,实际频数)n_i与理论频数np_i应相差不大.

K. Pearson提出: 检验统计量

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

拒绝域为

$$D = \{\chi^2 > c\}.$$

Theorem

在H_0为真时,

$$\chi^2 \xrightarrow{D} \chi^2(r-1), \quad as \quad n \to \infty.$$

那么对于给定的显著性水平α, 当n较大时, 可取$c = \chi^2_\alpha(r-1)$, 即

$$D = \{\chi^2 > \chi^2_\alpha(r-1)\}.$$

这一检验一般在$n \geq 50$及$np_i \geq 5, n_i \geq 5 (i = 1, 2, \cdots, r)$的条件下进行.

在病假的例子中. 注意到$\alpha = 0.05$, $n = 100$ (可以认为其充分大). 这时, $r = 5$, $\chi^2_{0.05}(4) = 9.488$. 因此拒绝域为

$$D = \{\chi^2 > 9.488\}.$$

列表如下

χ^2值计算表

工作日	周一	周二	周三	周四	周五	合计
n_i	17	27	10	28	18	100
np_i	20	20	20	20	20	100
$(n_i - np_i)^2/(np_i)$	0.45	2.45	5.00	3.20	0.20	11.30

$\chi^2 = 11.30 > 9.488$. 在$\alpha = 0.05$水平上拒绝原假设H_0, 认为该公司职工病假在五个工作日中不是均匀分布的.

有时χ^2的计算会用

$$\chi^2 = \sum_{i=1}^{r} \frac{n_i^2}{np_i} - n.$$

此检验的p-value为

$$P(\chi^2(r-1) \geq \chi^2\text{的值}).$$

此题中的p-value为

$$P(\chi^2(4) \geq 11.3) = 0.0234.$$

在$\alpha = 0.05$水平上, $0.05 > 0.0234$, 所以拒绝原假设H_0, 认为该公司职工病假在五个工作日中不是均匀分布的.

Example

某工厂制造一批骰子, 声称它是均匀的, 即在投掷中出现1点至6点的概率都应是1/6. 为检验骰子是否均匀, 要把骰子实地投掷若干次(如: 100次)或者从中随机选取100颗骰子, 各投掷一次(此类数据就是“分类数据”), 统计各点出现的频数如下表, 请检验H_0 : 骰子是均匀的(显著性水平为0.05).

点数	1	2	3	4	5	6
频数	15	17	16	18	16	18

解: 要检验的是H_0 :骰子是均匀的.

若记$A_i = \{$抛出的点数为 $i\}$, $i = 1, 2, 3, 4, 5, 6$, 则骰子的均匀性检验等价于检验问题

$$H_0 : \mathsf{P}(A_i) = \frac{1}{6}, \quad i = 1, 2, 3, 4, 5, 6.$$

取检验统计量为

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \quad (\text{或写为} = \sum_{i=1}^{r} \frac{n_i^2}{np_i} - n),$$

则检验的拒绝域形为

$$D = \{\chi^2 > c\}.$$

通过下表计算样本观测值

类别	1	2	3	4	5	6	总和n
实际频数n_i	15	17	16	18	16	18	100
理论频数np_i	50/3	50/3	50/3	50/3	50/3	50/3	100
$(n_i - np_i)^2/(np_i)$	1/6	1/150	2/75	8/75	2/75	8/75	0.44
或$n_i^2/(np_i)$	13.5	17.34	15.36	19.44	15.36	19.44	100.44

则

$$\chi^2 = \sum_{i=1}^{6} \frac{(n_i - np_i)^2}{np_i} = 0.44 \quad (\text{或} = \sum_{i=1}^{6} \frac{n_i^2}{np_i} - n = 100.44 - 100 = 0.44).$$

(解法一) 拒绝域为

$$D = \{\chi^2 > \chi^2_\alpha(r-1)\}.$$

现$\alpha = 0.05$, $r = 6$,查表得$\chi^2_{0.05}(5) = 11.070$, 则拒绝域为

$$D = \{\chi^2 > 11.070\}.$$

而统计量的值$\chi^2 = 0.44 < 11.070$, 故保留H_0, 认为骰子是均匀的.

(解法二) 检验的P_值为

$$\text{P\_值} = \mathsf{P}_{H_0}\{\chi^2 \geq 0.44\} = \mathsf{P}\{\chi^2(5) \geq 0.44\} = 0.9942.$$

现P_值$= 0.9942 > 0.05$, 故保留H_0, 认为骰子是均匀的.

2. 总体可分成有限个类, 总体理论分布含未知参数

设总体X可以分成r类:$A_1, A_2, \cdots, A_r$, 要检验的假设为

$$H_0 : \mathsf{P}(A_i) = p_i, \quad i = 1, 2, \cdots, r.$$

其中各个$p_i = p_i(\theta_1, \cdots, \theta_m)$依赖于$m$个未知(自由)参数, 且$p_i \geq 0$, $\sum_{i=1}^{r} p_i = 1$.

根据(一(1)), 自然的想法是先求出$\theta_1, \cdots, \theta_m$ 的极大似然估计$\widehat{\theta}_1, \cdots, \widehat{\theta}_m$, 从而得到诸$p_i$的估计$\widehat{p}_i = p_i(\widehat{\theta}_1, \cdots, \widehat{\theta}_m)$, 代入

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

得到

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

作为检验统计量.

K. Pearson仍认为在H_0为真时, χ^2的极限分布是$\chi^2(r-1)$. R. A. Fisher 纠正了这一错误, 证明了

$$\chi^2 \xrightarrow{D} \chi^2(r-m-1).$$

χ^2分布的自由度= (分类数) − (未知参数空间的维数)−1

Example

某种配偶的后代按体格的属性分为三类, 按某遗传学模型, 其相对频率之比应为

$$p^2 : 2p(1-p) : (1-p)^2,$$

其中$0 < p < 1$未知. 现进行了一次试验, 获得各类体格的个数分别为10, 53, 46. 试问: 这一样本是否在显著性水平$\alpha = 0.05$下与模型相符.

解 记配偶的后代属于第i类体格的属性为A_i, $i = 1, 2, 3$. 由题意知, 需检验

$$H_0 : p_1 = \mathsf{P}(A_1) = p^2,\ p_2 = \mathsf{P}(A_2) = 2p(1-p),\ p_3 = \mathsf{P}(A_3) = (1-p)^2.$$

由于原假设中含有一个未知参数$(m = 1)$, 需先来计算p的极大似然估计.

注意到似然函数

$$L(p) = L(p; \widetilde{x}) = (p^2)^{10} \cdot (2p(1-p))^{53} \cdot ((1-p)^2)^{46} = 2^{53} p^{73} (1-p)^{145},$$

$$\Longrightarrow \quad l(p) = \ln L(p) = 53 \ln 2 + 73 \ln p + 145 \ln(1-p)$$

由对数似然方程$\dfrac{dl(p)}{dp} = \dfrac{73}{p} - \dfrac{145}{1-p} = 0$, 并由微分法检验, 可得$p$的极大似然估计值为$\widehat{p}_{\mathrm{MLE}} = \dfrac{73}{218}$.

取检验统计量为

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - n\widehat{p}_i)^2}{n\widehat{p}_i} = \sum_{i=1}^{r} \frac{n_i^2}{n\widehat{p}_i} - n,$$

故拒绝域为

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

现$r=3$, $m=1$, $\alpha=0.05$, 查表得$\chi^2_{0.05}(1)=3.841$, 故$D=\{\chi^2>3.841\}$.
由下表计算χ^2的值

类别	1	2	3	总和
实测频数n_i	10	53	46	109
概率估计$\widehat{p}_i$	$\widehat{p}^2$=0.1121	$2\widehat{p}(1-\widehat{p})$=0.4455	$(1-\widehat{p})^2$=0.4424	1
理论频数$n\widehat{p}_i$	12.2225	48.555	48.2225	109
$(n_i-n\widehat{p}_i)^2/(n\widehat{p}_i)$	0.404	0.407	0.102	0.913

检验统计量的值为

$$\chi^2=0.913\le 3.841,$$

或考虑

$$\text{P\_值}=\mathsf{P}(\chi^2(r-m-1)\ge 0.913)=\mathsf{P}(\chi^2(1)\ge 0.913)=0.339>0.05,$$

故在显著性水平$\alpha=0.05$下, 保留原假设, 认为这一样本与模型相符.

二、分布拟合的χ^2检验

设$X_1, X_2, \cdots, X_n$是来自总体$F(x)$的样本, 需要检验的原假设为

$$H_0: \text{总体分布} F(x) = F_0(x),$$

其中$F_0(x)$称为理论分布. 它可以是一个完全已知的分布, 也可以是一个仅依赖于有限个未知实参数且具体数学形式已知的分布函数.

这个分布检验问题就是检验数据是否与理论分布相符合.

在样本容量较大时, 这个问题可以用分类数据的χ^2检验来解决.

1. 离散型分布

总体X在H_0下仅取有限个或者可列个值, 检验问题为

$$H_0 : X\text{服从分布}F_0(x;\theta).$$

若理论分布$F_0(x;\theta)$是离散型分布, 即仅取有限个或者可列个值$\{a_i\}$, 那么把若干个a_i值并成一类, 使得其可能取值$a_1, a_2, \cdots$ 被分成有限个类$B_1, \cdots, B_r$, 并使得样本的观察值$x_1, \cdots, x_n$落到每一个B_i内的个数n_i都不小于5. 记

$$\mathsf{P}(X \in B_i | H_0) = p_i = p_i(\theta).$$

那检验问题就转化成为检验

$$H_0 : \text{类}B_i\text{所占的比例为}p_i, i = 1, 2, \cdots, r.$$

Example

例: 在某交叉路口记录每15秒钟内通过的汽车数量(记为X), 共观察了25分钟(假设这100个15秒内通过的汽车数量是独立), 得100个数据如下:

通过的汽车数量	0	1	2	3	4	5	6	7	8	9	10	11
频数	1	5	15	17	26	11	9	8	3	2	2	1

在$\alpha = 0.05$水平上检验

$$H_0\text{: } X\text{服从泊松分布}P(\lambda)(\lambda > 0).$$

解: 在此题中, 虽然只观察到了$0,1,2,\cdots,11$这12个值, 但理论分布$P(\lambda)$的可能取值为非负整数值, 故可将其分为以下的12类:

$$A_i = \{15\text{秒钟内通过的汽车数量为}i\text{辆}\},\quad i=0,1,\cdots,10,$$

$$A_{11} = \{15\text{秒钟内通过的汽车数量不少于}11\text{辆}\}.$$

那么当H_0为真时,每一类出现的概率分别为

$$p_i = \frac{\lambda^i}{i!}e^{-\lambda},\quad i=0,1,\cdots,10,$$

$$p_{11} = \sum_{i=11}^{\infty}\frac{\lambda^i}{i!}e^{-\lambda}.$$

那么所要检验的原假设可以转化为

$$H_0 : \mathsf{P}(A_i) = p_i,\quad i=0,1,\cdots,11.$$

注意到原理论分布中含有未知参数λ, 我们先用λ的极大似然估计去估计它,

$$\widehat{\lambda} = \overline{x} = (0 \times 1 + 1 \times 5 + \cdots + 11 \times 1)/100 = 4.28.$$

那么

$$\widehat{p}_i = \frac{4.28^i}{i!} e^{-4.28}, \quad i = 0, 1, \cdots, 10, \quad \widehat{p}_{11} = \sum_{i=11}^{\infty} \frac{4.28^i}{i!} e^{-4.28}.$$

取检验统计量为

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - n\widehat{p}_i)^2}{n\widehat{p}_i},$$

则拒绝域为$D = \{\chi^2 > c\}$, 其中$\chi^2_\alpha(r - m - 1)$. 下由统计量的近似分布来确定临界值, 故要求$n_i \geq 5$(本来严格而言要求$np_i \geq 5$,为了方便起见常用$n_i \geq 5$替代). 所以将$\{0, 1\}$合并, $\{8, 9, 10, 11\}$合并.即, 将之分成了8类.

从而这里

$$\text{真正的分类数}r=8,\quad \text{未知参数个数}m=1.$$

在$\alpha=0.05$时,检验的临界值为

$$\chi^2_{0.05}(r-m-1)=\chi^2_{0.05}(6)=12.5916,$$

所以拒绝域为

$$D=\{\chi^2>12.5916\}.$$

而经计算各类概率

$$
\begin{aligned}
P(X \leq 1) =& e^{-\lambda} + \lambda e^{-\lambda}, \\
P(X = i) =& \frac{\lambda^i}{i!} e^{-\lambda}, \quad i = 2, 3, \cdots, 7, \\
P(X \geq 8) =& 1 - \sum_{i=0}^{7} \frac{\lambda^i}{i!} e^{-\lambda}
\end{aligned}
$$

的估计,得列表如下

χ^2检验的计算

i	n_i	$\widehat{p}_i$	$n\widehat{p}_i$	$\frac{(n_i-n\widehat{p}_i)^2}{n\widehat{p}_i}$
≤ 1	6	0.0731	7.3089	0.2344
2	15	0.1268	12.6788	0.4250
3	17	0.1809	18.0884	0.0655
4	26	0.1935	19.3546	2.2817
5	11	0.1657	16.5675	1.8710
6	9	0.1182	11.8182	0.6720
7	8	0.0723	7.2260	0.0829
≥ 8	8	0.0696	6.9577	0.1561
$\sum$	100			5.7886

由于$12.5916 > 5.7886$,

$$\Big(\text{p-value} = \mathsf{P}(\chi^2(r-m-1) \geq 5.7886) = 1 - \chi^2(5.7886, 6) = 0.4473\Big).$$

故在$\alpha = 0.05$下, 接受H_0, 认为通过交叉路口的汽车数量服从泊松分布.

Matlab 程序

```
>>n1=[1,5,15,17,26,11,9,8,3,2,2,1];
>>n=[6,15,17,26,11,9,8,8];
>>k=[0,1,2,3,4,5,6,7,8,9,10,11];
>> lambda=n1*k'/sum(n1)
lambda =
    4.2800
>> p1=poisscdf(1,lambda);p8=1-poisscdf(7,lambda);
>> p=[p1,poisspdf(2:7,lambda),p8]
p =
    0.0731    0.1268    0.1809    0.1935
    0.1657    0.1182    0.0723    0.0696
```

```
>> ntheory=sum(n1).*p
ntheory =
    7.3089   12.6788   18.0884   19.3546
    16.5675   11.8182    7.2260    6.9577
>> chisqure=(n-ntheory).^2./ntheory
chisqure =
    0.2344    0.4250    0.0655    2.2817
    1.8710    0.6720    0.0829    0.1561
```

```
>> sum(chisqure)
ans =
    5.7886
>> chi2inv(0.95,8-1-1)
ans =
   12.5916
>> pvalue=1-chi2cdf(sum(chisqure),8-1-1)
pvalue =
    0.4473
```

2. 连续型分布:

检验问题为

$$H_0 : X\text{服从分布}F_0(x),$$

其中$F_0(x)$中可以含m个未知参数, 若$m = 0$, 则F_0完全已知.

检验H_0的做法如下:

1. 把$F_0(x)$的取值范围分成r个区间, 不妨设

$$-\infty = a_0 < a_1 < a_2 < \cdots < a_{r-1} < a_r = \infty.$$

记$A_1 = (a_0, a_1)$, $A_2 = [a_1, a_2), \cdots$, $A_{r-1} = [a_{r-2}, a_{r-1})$, $A_r = [a_{r-1}, a_r)$.

2. 统计样本落入这r个区间的频数, 分别记为n_1, n_2, $\cdots$, n_r. 要求$n_i \geq 5$(本来严格而言要求$np_i \geq 5$,为了方便起见常用$n_i \geq 5$替代).

3. 记

$$p_i = \mathsf{P}(X \in A_i | H_0) = F_0(a_i) - F_0(a_{i-1}).$$

当$m \neq 0$时, 对m个参数给出其极大似然估计, 然后算得各p_i的估计$\widehat{p}_i$.

这样就把问题转化为分类数据的检验问题了, 然后按(一(1))或(一(2))中的步骤进行检验.

Example

例: 下面列出了84个伊特拉斯坎(Etruscan)人男子的头颅的最大宽度(mm):

141	148	132	138	154	142	150	146	155	158	150	140
147	148	144	150	149	145	149	158	143	141	144	144
126	140	144	142	141	140	145	135	147	146	141	136
140	146	142	137	148	154	137	139	143	140	131	143
141	149	148	135	148	152	143	144	141	143	147	146
150	132	142	142	143	153	149	146	149	138	142	149
142	137	134	144	146	147	140	142	140	137	152	145

试根据此批数据在$\alpha = 0.1$显著性水平下检验

H_0: 伊特拉斯坎(Etruscan)人男子的头颅的最大宽度X服从正态分布.

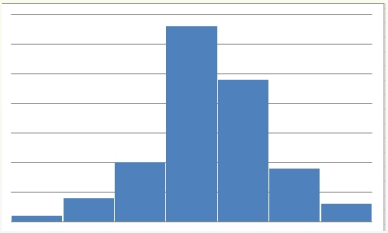
解: 为粗略了解数据的分布情况，先画出直方图. 步骤如下:

1.找出数据的最小值、最大值为126、158，取区间$[124.5, 159.5]$,它能覆盖$[126, 158]$;

2.将区间$[124.5, 159.5]$等分为7个小区间,小区间的长度$\Delta = \frac{159.5-124.5}{7} = 5$, Δ称为组距, 小区间的端点称为组限，建立下表:

组限	频数n_i	频率n_i/n	累计频率
124.5-129.5	1	0.0119	0.0119
129.5-134.5	4	0.0476	0.0595
134.5-139.5	10	0.1191	0.1786
139.5-144.5	33	0.3929	0.5715
144.5-149.5	24	0.2857	0.8527
149.5-154.5	9	0.1071	0.9524
154.5-159.5	3	0.0357	1

3.自左向右在各小区间上作以$n_i/(n\Delta)$为高的小矩形, 见下图, 即为直方图.



注: 直方图的小区间可以不等长, 但小区间的长度不能太大, 否则平均化作用突出, 淹没了密度的细节部分; 也不能太小, 否则受随机化影响太大, 产生极不规则的形状.

从本例的直方图看, 该图呈现中间高, 两头低, 较对称, 因此比较像似来自正态总体. 于是就提出假设检验

$$H_0\text{: } X\text{服从正态分布}N(\mu,\sigma^2)(-\infty<\mu<\infty,\sigma>0).$$

注意到原假设中含有两个未知参数, 先分别求出其极大似然估计量分别为

$$\widehat{\mu}=\overline{X},\quad \widehat{\sigma^2}=m_{n,2}.$$

结合样本, 可得其极大似然估计值分别为$\widehat{\mu}=143.8,\quad \widehat{\sigma^2}=6.0^2$. 由此可得$H_0$下, 总体的概率密度函数的估计为

$$\widehat{p(x)}=\frac{1}{6\sqrt{2\pi}}\exp\{-\frac{(x-143.8)^2}{2\times 6^2}\},\quad -\infty<x<\infty.$$

取检验统计量为

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - n\widehat{p}_i)^2}{n\widehat{p}_i},$$

则拒绝域为$D = \{\chi^2 > c\}$, 其中$c = \chi^2_\alpha(r-m-1)$, $m = 2$. 注意到

A_i	n_i	$\hat{p}_i$	$n\hat{p}_i$	$n_i^2/n\hat{p}_i$
A_1 x<129.5	1	0.0087	0.73	
A_2 129.5≤x<134.5	4	0.0519	4.36	4.91
A_3 134.5≤x<139.5	10	0.1752	14.72	6.79
A_4 139.5≤x<144.5	33	0.3120	26.21	41.55
A_5 144.5≤x<149.5	24	0.2811	23.61	24.40
A_6 149.5≤x<154.5	9	0.1336	11.22	10.02
A_7 154.5≤x<∞	3	0.0375	3.15	Σ=87.67

5.09

14.37

故$r = 5$, 又由于$\alpha = 0.1$, 故拒绝域为$D = \{\chi^2 > \chi^2_{0.1}(2)\} = \{\chi^2 > 4.605\}$. 现检验统计量的值为$87.67 - 84 = 3.67 < 4.605$, 故在显著性水平0.1 下保留$H_0$, 认为数据来自正态总体.

Example

例: 测得200件混泥土制件的抗压强度, 按区间分布如下:

抗压强区间$(a_{i-1}, a_i]$	频数
$(-\infty, 200]$	10
$(200, 210]$	26
$(210, 220]$	56
$(220, 230]$	64
$(230, 240]$	30
$[240, \infty)$	14
合计	200

试在$\alpha = 0.05$水平下检验: H_0 : 抗压强度服从正态分布$N(\mu, \sigma^2)$.

解: 先求正态分布$N(\mu,\sigma^2)$中μ和σ^2的MLE. 如果由原始的200个数据,我们可以根据这些数据求得μ和σ^2 的MLE:

$$\widehat{\mu}=\overline{x},\quad \widehat{\sigma^2}=m_{n,2}.$$

如果没有原始数据,我们用组中值(区间中点)去代替原始数据. 即,
记$x_1=195$, $x_2=205$, $x_3=215$, $x_4=225$, $x_5=235$, $x_6=245$, 于是

$$\widehat{\mu}=\overline{x}=\frac{1}{200}\sum_{i=1}^{6}n_ix_i=221,$$

$$\widehat{\sigma^2}=\frac{1}{200}\sum_{i=1}^{6}n_i(x_i-\overline{x})^2=152.$$

按

$$p_i = F_0(a_i) - F_0(a_{i-1}) = \Phi(\frac{a_i - \mu}{\sigma}) - \Phi(\frac{a_{i-1} - \mu}{\sigma}),$$

得

$$\widehat{p}_i = \Phi(\frac{a_i - 221}{\sqrt{152}}) - \Phi(\frac{a_{i-1} - 221}{\sqrt{152}}), \quad i = 1, 2, \cdots, r.$$

列表如下

χ^2值计算表

区间	n_i	$\frac{a_i-221}{\sqrt{152}}$	$\Phi(\cdot)$	$\widehat{p}_i$	$n\widehat{p}_i$	$\frac{(n_i-n\widehat{p}_i)^2}{n\widehat{p}_i}$
$(-\infty, 200]$	10	-1.7033	0.0443	.0443	8.8507	.1492
$(200, 210]$	26	-0.8922	0.1861	.1419	28.3769	.1991
$(210, 220]$	56	-0.0811	0.4677	.2815	56.3078	.0017
$(220, 230]$	64	0.7300	0.7673	.2996	59.9254	.2771
$(230, 240]$	30	1.5411	0.9384	.1711	34.2101	.5181
$[240, \infty)$	14	∞	1	.0616	12.3292	.2264
$\sum$	200			1	200	1.3716

取检验统计量为

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - n\widehat{p}_i)^2}{n\widehat{p}_i},$$

则拒绝域为$D = \{\chi^2 > \chi^2_\alpha(r-m-1)\}$, 现$r = 6, m = 2, \alpha = 0.05$,查表得$\chi^2_{0.05}(r-m-1) = \chi^2_{0.05}(6-2-1) = 7.8147$, 拒绝域为

$$D = \{\chi^2 > 7.8147\}.$$

所以在水平为$\alpha = 0.05$下, 接受H_0, 认为接受抗压强度的分布服从正态分布.

$$\text{p-value} = P\{\chi^2(r-m-1) > 1.3716\} = 0.7122.$$

Matlab 程序

```
>>a=[200,210,220,230,240];b=(a-221)/sqrt(152)
b =
   -1.7033   -0.8922   -0.0811    0.7300    1.5411
>> q=normcdf(b)
q =
    0.0443    0.1861    0.4677    0.7673    0.9384
>> q1=[0,q];q2=[q,1];p=q2-q1
p =
    0.0443    0.1419    0.2815    0.2996    0.1711    0.0616
>> n=[10,26,56,64,30,14]; ntheory=sum(n).*p
ntheory =
    8.8507   28.3769   56.3078   59.9254   34.2101   12.3292
```

```
>> chisqure=(n-ntheory).^2./ntheory
chisqure =
    0.1492    0.1991    0.0017    0.2771    0.5181    0.2264
>> sum(chisqure)
ans =
    1.3716
>> pvalue=1-chi2cdf(sum(chisqure),6-2-1)
pvalue =
    0.7122
>> chi2inv(0.95,6-2-1)
ans =
    7.8147
```

用Pearson的χ^2检验来检验

$$H_0 : F(x) = F_0(x)$$

时, 实际上只检验了

$$H_0' : F(a_i) - F(a_{i-1}) = F_0(a_i) - F_0(a_{i-1}) = p_i, i = 1, \cdots, r.$$

这样即使H_0'为真, H_0也不一定为真, 会导致犯第二类错误.

§Kolmogorov-Smirnov 检验

一、Kolmogorov 检验

设$X_1, X_2, \cdots, X_n$是取自总体$X \sim F(x)$, 样本经验分布函数为

$$F_n(x) = \begin{cases} 0, & x \le X_{(1)}; \\ i/n, & X_{(i)} < x \le X_{(i+1)}, i = 1, \cdots, n-1; \\ 1, & x > X_{(n)}, \end{cases}$$

其中$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$为次序统计量. 且已知

$$\mathsf{P}\Big(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0 \Big) = 1.$$

故对分布检验问题

$$H_0 : F(x) = F_0(x),$$

当理论分布$F_0(x)$已知时, 我们可以取

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|$$

为检验统计量, 并且D_n较大时拒绝H_0 —Kolmogorov 检验.

临界值的计算:

$F_0(x)$是单调非降函数, $F_n(x)$是单调非降的阶梯函数, $|F_n(x) - F_0(x)|$的上确界可在n个$X_{(i)}$处找. 又$F_n(X_{(i)}) = \frac{i-1}{n}$, $F_n(X_{(i)}+) = \frac{i}{n}$, 所以

$$D_n = \max_{1\le i\le n} \left\{ \left| F_0(X_{(i)}) - \frac{i-1}{n} \right| \vee \left| F_0(X_{(i)}) - \frac{i}{n} \right| \right\}.$$

D_n的分布: 对$\lambda > -\frac{1}{2n}$, $D_n < \lambda + \frac{1}{2n} \iff$

$$-\lambda - \tfrac{1}{2n} < F_0(X_{(i)}) - \tfrac{i-1}{n} < \lambda + \tfrac{1}{2n},$$
$$-\lambda - \tfrac{1}{2n} < F_0(X_{(i)}) - \tfrac{i}{n} < \lambda + \tfrac{1}{2n},$$
$$i = 1, \cdots, n$$
$$\iff \frac{2i-1}{2n} - \lambda < F_0(X_{(i)}) < \frac{2i-1}{2n} + \lambda,\ i = 1, \cdots, n$$

在H_0下, 且$F_0(x)$连续时, $U_i = F_0(X_i) \sim U(0,1)$, 所以

$$\begin{aligned} &\mathsf{P}\left(D_n < \lambda + \frac{1}{2n}\right) \\ =&\mathsf{P}\left(\frac{2i-1}{2n} - \lambda < U_{(i)} < \frac{2i-1}{2n} + \lambda,\ i = 1, \cdots, n\right). \end{aligned}$$

Theorem

设$F_0(x)$是连续的, 则在原假设H_0为真时,

$$P\left(D_n < \lambda + \frac{1}{2n}\right)$$

$$= \begin{cases} 0, & \text{当 } \lambda \le -\frac{1}{2n}; \\ \int_{\frac{1}{2n}-\lambda}^{\frac{1}{2n}+\lambda} \int_{\frac{3}{2n}-\lambda}^{\frac{3}{2n}+\lambda} \cdots \int_{\frac{2n-1}{2n}-\lambda}^{\frac{2n-1}{2n}+\lambda} f(y_1, \cdots, y_n) dy_1 \cdots dy_n, \\ \qquad \text{当 } -\frac{1}{2n} < \lambda < \frac{2n-1}{2n}; \\ 1, & \text{当 } \lambda \ge \frac{2n-1}{2n}, \end{cases}$$

其中

$$f(y_1, \cdots, y_n) = \begin{cases} n! & \text{当} 0 < y_1 < \cdots < y_n < 1 \\ 0 & \text{其它.} \end{cases}$$

由这个定理可以得到$D_{n,\alpha}$使得

$$\mathsf{P}(D_n > D_{n,\alpha}|H_0) = \alpha \qquad \text{当n不太大时.}$$

书中附表13列出了n不太大时, 柯尔莫哥洛夫检验临界值$D_{n,\alpha}$. 当n充分大时, 我们可用D_n的渐近分布(见下面定理)求$D_{n,\alpha}$.

Theorem

*(Kolmogorov)*设$F_0(x)$是连续的, 则在原假设H_0为真时, 当$n \to \infty$,

$$\begin{aligned} &P\left(\sqrt{n}D_n < \lambda\right) \\ \to &K(\lambda) = \begin{cases} \sum_{j=-\infty}^{\infty}(-1)^j \exp\{-2j^2\lambda^2\}, & \text{当 } \lambda > 0 \\ 0, & \text{当 } \lambda \le 0. \end{cases} \end{aligned}$$

Example

在$\alpha = 0.1$下,检验是否可以认为下列10个数是来自正态分布$N(0,1)$的随机数:

0.4855, -0.0050, -0.2762, 1.2765, 1.8634, -0.5226, 0.1034, -0.8076, 0.6804, -2.3646

取$D_n = \sup_x |F_n(x) - F_0(x)|$为检验统计量, 现$n = 10$, $\alpha = 0.1$, 查附表13, 得拒绝域为

$$D = \{D_n > D_{n,\alpha}\} = \{D_n > 0.369\}.$$

下计算检验统计量的值:

i	$x_{(i)}$	$F_0(x_{(i)})$	$(i-1)/n$	i/n	δ_i
1	-2.3646	0.0090	0	0.1	0.0910
2	-0.8076	0.2096	0.1	0.2	0.1096
3	-0.5226	0.3006	0.2	0.3	0.1006
4	-0.2762	0.3912	0.3	0.4	0.0912
5	-0.0050	0.4980	0.4	0.5	0.0980
6	0.1034	0.5412	0.5	0.6	0.0588
7	0.4855	0.6863	0.6	0.7	0.0863
8	0.6804	0.7519	0.7	0.8	0.0519
9	1.2765	0.8991	0.8	0.9	0.0991
10	1.8634	0.9688	0.9	1.0	0.0688

$D_n = 0.1096 < D_{10,0.10} = 0.369$, 故保留$H_0$, 认为数据是来自$N(0,1)$.

```
data =[0.4855, -0.0050, -0.2762, 1.2765, 1.8634,
      -0.5226, 0.1034, -0.8076, 0.6804, -2.3646];
>> xi=sort(data)
xi =
   -2.3646   -0.8076   -0.5226   -0.2762   -0.0050
    0.1034    0.4855    0.6804    1.2765    1.8634
>> f0xi=normcdf(xi,0,1)
f0xi =
    0.0090    0.2096    0.3006    0.3912    0.4980
    0.5412    0.6863    0.7519    0.8991    0.9688
>> a=(0:9)/10
a =
         0    0.1000    0.2000    0.3000    0.4000
    0.5000    0.6000    0.7000    0.8000    0.9000
```

```
>> b=(1:10)/10
b =
    0.1000    0.2000    0.3000    0.4000    0.5000
    0.6000    0.7000    0.8000    0.9000    1.0000
>> delta=max(abs(f0xi-a),abs(f0xi-b))
delta =
    0.0910    0.1096    0.1006    0.0912    0.0980
    0.0588    0.0863    0.0519    0.0991    0.0688
>> max(delta)
ans =
    0.1096
```

二、Smirnov 检验

检验两总体的分布函数的关系, 如:

$$H_0 : F_1(x) = F_2(x)$$

取检验统计量为

$$D_{n_1,n_2} = \sup_x |F_{1n_1}(x) - F_{2n_2}(x)|.$$

则拒绝域为

$$D = \{D_{n_1,n_2} > C\}.$$

若考虑单边检验问题：

$$H_0 : F_1(x) \leq F_2(x) \longleftrightarrow H_1 : F_1(x) > F_2(x),$$

则取检验统计量为

$$D^+_{n_1,n_2} = \sup_x \big(F_{1n_1}(x) - F_{2n_2}(x)\big).$$

则拒绝域为

$$D = \{D^+_{n_1,n_2} > C\}.$$

Theorem

若$F_1(x) \equiv F_2(x)$, 且连续, 则

$$\lim_{n_1,n_2\to\infty} P\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}D^+_{n_1,n_2} \le x\right) = \begin{cases} 1-e^{-2x^2}, & x>0, \\ 0, & x \le 0. \end{cases}$$

$$\lim_{n_1,n_2\to\infty} P\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}D_{n_1,n_2} \le x\right) = \begin{cases} \sum_{j=-\infty}^{\infty}(-1)^j \exp\{-2j^2x^2\}, & x>0 \\ 0, & x \le 0. \end{cases}$$

§6.5 列联表的独立性检验

Example

	男	女	合计
正常	442	514	956
色盲	38	6	44
合计	480	520	1000

检验性别与色盲这两个特性是否独立, 即

$$H_0: \text{性别与色盲这两个特性独立.}$$

此题中的数据是一个2×2的列联表.

一般地, 假定n个随机试验结果(或n个个体)根据两个特性X_1和X_2进行分类,若X_1中有r类$A_1, \cdots, A_r$, X_2中有s类$B_1, \cdots, B_s$, 属于A_i和B_j的个体的数目为n_{ij}, 那么得到$r \times s$列联表:

	B_1	B_2	$\cdots$	B_s	合计
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\cdot}$
合计	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	n

要检验的是两个特性X_1和X_2是否独立.

为了明确写出检验问题, 记总体为X. 它是二维变量(X_1, X_2), 这里X_1被分成r类$A_1, \cdots, A_r$, X_2被分成s类$B_1, \cdots, B_s$. 并设

$$\mathsf{P}(X \in A_i \cap B_j) := \mathsf{P}(X_1 \in A_i \ \text{且}\ X_2 \in B_j) = p_{ij}, \quad i = 1, \cdots, r; j = 1, \cdots, s.$$

又记

$$p_{i\cdot} = \sum_{j=1}^{s} p_{ij} = \mathsf{P}(X_1 \in A_i), \ \ p_{\cdot j} = \sum_{i=1}^{r} p_{ij} = \mathsf{P}(X_2 \in B_j).$$

这里必有

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

那么当X_1与X_2两个特性独立时,应对所有的i,j,有

$$p_{ij} = p_{i\cdot} \cdot p_{\cdot j}.$$

因此我们的检验问题即为

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}, \ \ \forall i, j \quad H_1 : \text{至少一对}(i, j) \text{ 有} p_{ij} \neq p_{i\cdot}p_{\cdot j} \tag{1}$$

当$p_{i\cdot}$, $p_{\cdot j}$已知时, 考虑到$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-np_{ij})^2}{np_{ij}}$,取检验统计量为

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-np_{i\cdot}p_{\cdot j})^2}{np_{i\cdot}p_{\cdot j}}\xrightarrow{D}\chi^2(rs-1),$$

$$\text{as}\quad n\to\infty,\quad \text{当}H_0\text{成立}.$$

则拒绝域为$\{\chi^2>\chi^2_\alpha(rs-1)\}$.

若$p_{i\cdot}$, $p_{\cdot j}$未知, 那么我们有$r+s$个参数要估计. 但由于$\sum_i p_{i\cdot}=1$, $\sum_j p_{\cdot j}=1$. 因此只有$r+s-2$个独立参数要估计. 而$p_{i\cdot}$, $p_{\cdot j}$ 的MLE为

$$\widehat{p}_{i\cdot}=\frac{n_{i\cdot}}{n},\quad \widehat{p}_{\cdot j}=\frac{n_{\cdot j}}{n}.$$

检验统计量为

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

在H_0为真时, 其极限分布为χ^2分布, 自由度为

$$rs-(r+s-2)-1=(r-1)(s-1).$$

则拒绝域为$\{\chi^2>\chi^2_\alpha((r-1)(s-1))\}$.

色盲与性别问题:

	男	女	合计
正常	442	514	956
色盲	38	6	44
合计	480	520	1000

取检验统计量为

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

则拒绝域为$\{\chi^2 > \chi^2_\alpha((r-1)(s-1))\}$.现$r = s = 2$, 若$\alpha = 0.01$, 查表得$\chi^2_\alpha((r-1)(s-1)) = \chi^2_{0.01}(1) = 6.6349$, 那么拒绝域为$\{\chi^2 > 6.6349\}$

下计算统计量的值. 以$\widehat{n}_{ij} = n\widehat{p}_{i\cdot}\widehat{p}_{\cdot j} = \frac{n_{i\cdot}n_{\cdot j}}{n}$ 构造另一个2×2列联表:

理论频数表

$n\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}$	男	女	合计
正常	458.88	497.12	956
色盲	21.12	22.88	44
合计	480	520	1000

由此得

χ^2值表

$\frac{(n_{ij}-\widehat{n}_{ij})^2}{\widehat{n}_{ij}}$	男	女	合计
正常	0.6209	0.5732	1.1941
色盲	13.4912	12.4534	25.9446
合计	14.1121	13.0266	27.1387

注意到$6.6349 < 27.1387$, 因此拒绝原假设, 认为色盲与性别两特性不独立.
或用p值,

$$\text{p-value} = \mathsf{P}\{\chi^2(1) > 27.1387\} = 1.8937 \times 10^{-7} < 0.01,$$

因此拒绝原假设.

Matlab 程序

```
>> A =[442, 514; 38, 6];
>> a1=sum(A)
a1 =
   480   520
>> a2=sum(A')
a2 =
   956    44
>> B=a2'*a1/1000
B =
  458.8800  497.1200
   21.1200   22.8800
```

```
>> C=(A-B).^2./B
C =
    0.6209    0.5732
   13.4912   12.4534
>> sum(C(:))
ans =
   27.1387

>> chi2inv(0.99,1)
ans =
    6.6349
>> pvalue=1-chi2cdf(sum(C(:)),1)
pvalue =
  1.8936e-007
```

Example

有的零售商店开展上门服务的业务,有的不开展这项业务. 为了了解这项业务的开展是否与其月销售额有关, 调查了363个商店, 结果如下:

		上门服务	不上门服务	合计
月	≤ 1	32	29	61
销	$(1,5]$	111	24	135
售	$(5,10]$	104	6	110
额	$(10,20]$	40	2	42
(千元)	>20	14	1	15
合计		301	62	363

在$\alpha = 0.01$水平上检验服务方式与月销售额是否有关?

根据题意, 检验为

$$H_0: \text{服务方式与月销售额独立.}$$

取检验统计量为

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

其中$\widehat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}$, $\widehat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$, 则拒绝域为

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

现$r = 5$, $s = 2$, $\alpha = 0.01$, 则拒绝域为

$$\{\chi^2 > \chi^2_\alpha(4)\}.$$

以$\widehat{n}_{ij} = n\widehat{p}_{i\cdot}\widehat{p}_{\cdot j} = \frac{n_{i\cdot}n_{\cdot j}}{n}$构造另一个列联表:

理论频数表

$n\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}$	上门服务	不上门服务	合计
≤ 1	50.5813	10.4187	61
$(1, 5]$	111.9421	23.0579	135
$(5, 10]$	91.2121	18.7879	110
$(10, 20]$	34.8264	7.1736	42
> 20	12.4380	2.5620	15
合计	301	62	363

由此得

χ^2值表

$\frac{(n_{ij}-\widehat{n}_{ij})^2}{\widehat{n}_{ij}}$	上门服务	不上门服务	合计
≤ 1	6.8259	33.1387	39.9646
$(1,5]$	0.0079	0.0385	0.0464
$(5,10]$	1.7929	8.7040	10.4969
$(10,20]$	0.7685	3.7312	4.4997
>20	0.1962	0.9523	1.1485
合计	9.5914	46.5647	56.1561

所以$\chi^2_{0.01}((2-1)(5-1)) = 13.2767 < 56.1561 = \chi^2$, 拒绝$H_0$: 服务方式与月销售额独立, 由此认为两者是不独立的.

Matlab 程序

```
>> N=[32,29;111,24;104,6;40,2;14,1];
>> Ncolumn=sum(N')
Ncolumn =
    61   135   110    42    15
>> Nrow=sum(N)
Nrow =
   301    62
>> Ntheory=Ncolumn'*Nrow/sum(N(:))
Ntheory =
   50.5813   10.4187
  111.9421   23.0579
   91.2121   18.7879
   34.8264    7.1736
```

```
>> chisqure=(N-Ntheory).^2./Ntheory
chisqure =
    6.8259   33.1387
    0.0079    0.0385
    1.7929    8.7040
    0.7685    3.7312
    0.1962    0.9523
>> chisq=[chisqure;sum(chisqure)];chisq=[chisq';sum(chisq')]'
chisq =
    6.8259   33.1387   39.9646
    0.0079    0.0385    0.0464
    1.7929    8.7040   10.4969
    0.7685    3.7312    4.4997
    0.1962    0.9523    1.1485
```

```
>> pvalue=1-chi2cdf(sum(chisqure(:)),4)
pvalue =
  1.8596e-011
>> chi2inv(0.99,4)
ans =
   13.2767
```