第4.3节　非参数假设检验法

一、χ^2 拟合优度检验

二、柯尔莫哥洛夫及斯米尔诺夫检验

三、独立性检验

问题引入

第二节涉及到的假设检验问题，都是依赖总体为正态分布。总体服从什么分布，一般预先无法知晓，因而需要对总体的分布进行各种假设。

不依赖于分布的统计方法称为非参数统计方法，本节将主要讨论对总体分布的非参数假设检验问题.

本节主要介绍其中常见的3种方法：拟合优度检验，柯尔莫哥洛夫-斯米尔诺夫检验和独立性检验.

一、χ^2拟合优度检验法

1. χ^2检验法的定义

这是在总体的分布未知的情况下,根据样本 $X_1, X_2, \cdots, X_n$ 来检验关于总体分布的假设

H_0 : 总体 X的分布函数为 $F(x)$,

H_1 : 总体 X的分布函数不是 $F(x)$,

的一种方法.

说明

(1)在这里备择假设H_1可以不必写出.

(2) 若总体 X 为离散型：则上述假设相当于

H_0：总体 X 的分布律为$P\{X = x_i\} = p_i, i = 1, 2, \cdots.$

(3) 若总体 X 为连续型： 则上述假设相当于

H_0：总体 X 的概率密度为 $f(x)$.

(4) 在使用χ^2检验法检验假设 H_0 时,若 $F(x)$ 的形式已知,但其参数值未知,需要先用最大似然估计法估计参数,然后作检验.

2. χ^2检验法的基本思想

将随机试验可能结果的全体Ω分为m个互不相容的事件$A_1, A_2, \cdots, A_m$ ($\sum_{i=1}^{m} A_i = \Omega$, $A_i A_j = \phi, i \neq j$, $i, j = 1, 2, \cdots, m$). 于是在假设 H_0 下, 我们可以计算 $p_{i0} = P(A_i)$, $i = 1, 2, \cdots, k$. 在n次试验中, 事件 A_i 出现的频率$\frac{N_i}{n}$与 p_{i0} 往往有差异,但一般来说, 若 H_0 为真, 且试验次数又多时, 这种差异不应很大.

3.皮尔逊定理

设检验假设 H_0 的统计量为皮尔逊统计量

$$\chi_n^2=\sum_{i=1}^{m}\frac{\left(N_i-np_{i0}\right)^2}{np_{i0}}\left(\text{或}\chi_n^2=\sum_{i=1}^{m}\frac{N_i^2}{np_{i0}}-n\right)$$

定理4.1 若 n 充分大(≥ 50), 则当 H_0 为真时(不论 H_0 中的分布属什么分布), 皮尔逊统计量总是近似地服从自由度为 $m-1$ 的 χ^2 分布.

$$\chi_n^2=\sum_{i=1}^{m}\frac{\left(N_i-np_{i0}\right)^2}{np_{i0}}\overset{\text{近似}}{\sim}\chi^2(m-1)$$

于是，如果在假设 H_0 下，

$$\chi_n^{\ 2}=\sum_{i=1}^{m}\frac{(N_i-np_{i0})^2}{np_{i0}}\geq\chi_\alpha^{\ 2}(m-1),$$

则在显著性水平 α 下拒绝 H_0，否则就接受 H_0.

注意

在使用 χ^2 检验法时，n 要足够大，np_{i0} 不太小.

根据实践，一般 $n\geq 50$，每一个 $np_{i0}\geq 5$.

4. 多项分布的 χ^2 检验法

设总体X为离散型分布，其分布律为

$$P\{X = x_i\} = p_i,\ \ i = 1,2,\cdots,m.$$

设$(X_1, X_2, \cdots, X_n)^T$为来自总体$X$的样本,$(x_1, x_2, \cdots, x_n)^T$为其观测值，$N_i$表示$(X_1, X_2, \cdots, X_n)^T$中取值为$x_i$的个数，且$\sum_{i=1}^{m} N_i = n$,$(N_1, N_2, \cdots, N_m)^T$分布为

$$P\{N_1 = n_1, N_2 = n_2, \cdots, N_m = n_m\} = \frac{n!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m}$$

假设检验的问题为

$$H_0 : p_i = p_{i0} \leftrightarrow H_1 : p_i \neq p_{i0}, i = 1, 2, \cdots, m$$

由前面的分析可以看出，选择皮尔逊统计量

$$\chi_n^2 = \sum_{i=1}^{m} \frac{\left(N_i - np_{i0}\right)^2}{np_{i0}} \left(\text{或} \chi_n^2 = \sum_{i=1}^{m} \frac{N_i^2}{np_{i0}} - n \right)$$

拒绝域为

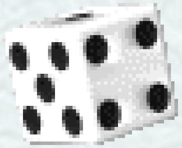
$$W = \{x : \chi_n^2 = \sum_{i=1}^{m} \frac{(N_i - np_{i0})^2}{np_{i0}} \geq \chi_\alpha^2(m-1)\}$$

例1 把一颗骰子重复抛掷 300 次, 结果如下:

出现的点数	1	2	3	4	5	6
出现的频数	40	70	48	60	52	30

试检验这颗骰子的六个面是否匀称? (取 $\alpha = 0.05$)

解 根据题意需要检验假设



H_0: 这颗骰子的六个面是匀称的.

(或 $H_0: P\{X = i\} = \dfrac{1}{6} \quad (i = 1,2,\cdots,6)$)

其中X表示抛掷这骰子一次所出现的点数 (可能值只有6个),

在 H_0 为真的前提下,

$$p_{i0}=\frac{1}{6}\quad(i=1,2,\cdots,6)$$

$$\chi_n^{\,2}=\sum_{i=1}^{6}\frac{(N_i-np_{i0})^2}{np_{i0}}$$

$$=\frac{(40-300\times\frac{1}{6})^2}{300\times\frac{1}{6}}+\frac{(70-300\times\frac{1}{6})^2}{300\times\frac{1}{6}}+\frac{(48-300\times\frac{1}{6})^2}{300\times\frac{1}{6}}+$$

$$\frac{(60-300\times\frac{1}{6})^2}{300\times\frac{1}{6}}+\frac{(52-300\times\frac{1}{6})^2}{300\times\frac{1}{6}}+\frac{(30-300\times\frac{1}{6})^2}{300\times\frac{1}{6}}$$

$\chi_n^{\ 2}=20.16$ 自由度为$6-1=5$,

查卡方分布表得$\chi_{0.05}^2(5)=11.07$, $\chi^2=20.16>11.07$,

所以拒绝 H_0, 认为这颗骰子的六个面不是匀称的.

5. 一般分布的 χ^2 检验法

假设检验的问题为 $H_0 : F(x) = F_0(x),$

任取$m-1$个实数, 使得$-\infty < a_1 < \cdots < a_{m-1} < +\infty,$

$$A_1 = (-\infty, a_1), A_2 = [a_1, a_2), \cdots, A_m = [a_{m-1}, +\infty),$$

$$令 p_{i0} = F_0(a_i) - F_0(a_{i-1}), i = 2, \cdots, m-1,$$

$$p_{10} = F_0(a_1), p_{m0} = 1 - F_0(a_{m-1})$$

设$(X_1, X_2, \cdots, X_n)^T$为来自总体$X$的样本,$(x_1, x_2, \cdots, x_n)^T$为其观测值，$N_i$表示样本值落入每个区间A 的频数，且$\sum_{i=1}^{m} N_i = n, (N_1, N_2, \cdots, N_m)^T$分布为多项分布.

经过上述处理，此问题又转化为检验多项分布问题.

选择皮尔逊统计量

$$\chi_n^2 = \sum_{i=1}^{m} \frac{\left(N_i - np_{i0}\right)^2}{np_{i0}} \left(\text{或} \chi_n^2 = \sum_{i=1}^{m} \frac{N_i^2}{np_{i0}} - n \right)$$

拒绝域为

$$W = \{x : \chi_n^2 = \sum_{i=1}^{m} \frac{(N_i - np_{i0})^2}{np_{i0}} \geq \chi_\alpha^2(m-1)\}$$

例2(p136例4.11)　某盒中装有白球和黑球，现做下面的试验，用返回式抽取方式从盒中取球，直到取到白球为止，记录下抽取的次数，重复如此的试验100次，其结果为：

抽取次数	1	2	3	4	≥ 5
频数	43	31	15	6	5

试问该盒中的白球与黑球的个数是否相等(α=0.05)?

解　从题意可知，该总体服从几何分布，

$$P\{X=k\}=(1-p)^{k-1}p, k=1,2,\cdots,$$

p是盒中白球数量所占的比例。

若黑球白球个数相等，则p=1/2,因此

$$P\{X=1\}=\frac{1}{2},P\{X=2\}=\frac{1}{4},P\{X=3\}=\frac{1}{8},$$

$$P\{X=4\}=\frac{1}{16},P\{X\geq 5\}=\frac{1}{16}$$

由此可知，检验的问题是

$$H_0: p_{10}=\frac{1}{2},p_{20}=\frac{1}{4},p_{30}=\frac{1}{8},p_{40}=\frac{1}{16},p_{50}=\frac{1}{16},$$

计算皮尔逊统计量可得：

$$\chi_n^2=\sum_{i=1}^{m}\frac{\left(N_i-np_{i0}\right)^2}{np_{i0}}=3.2$$

查表可得

$$\chi_{0.05}^2(4)=9.488$$

显然

$$\chi_n^2=\sum_{i=1}^{m}\frac{\left(N_i-np_{i0}\right)^2}{np_{i0}}=3.2<\chi_{0.05}^2(4)=9.488$$

因而接受原假设，黑球白球个数相等.

6. 分布中含有未知参数的 χ^2 检验法

前面主要讨论了多项分布和分布形式完全确定下的检验方法。在很多情况下，只确定了总体分布的类型，而分布中包含未知参数，则需要先利用最大似然估计法对未知参数进行估计，然后带入分布函数。

6. 分布中含有未知参数的 χ^2 检验法

假设检验的问题为

$$H_0: F(x) = F_0(x, \theta_1, \cdots, \theta_r) \leftrightarrow H_1: F(x) \neq F_0,$$

其中F_0的形式已知,参数$\theta_1, \cdots, \theta_r$未知.

设$(X_1, X_2, \cdots, X_n)^T$为来自总体$X$的样本,$(x_1, x_2, \cdots, x_n)^T$为其观测值，用最大似然估计首先得到参数的估计. 由此可以得到$F_0(x, \hat{\theta}_1, \cdots, \hat{\theta}_r)$,令

$$\hat{p}_{10} = F_0(a_1, \hat{\theta}_1, \cdots, \hat{\theta}_r),\ \hat{p}_{m0} = 1 - F_0(a_{m-1}, \hat{\theta}_1, \cdots, \hat{\theta}_r)$$

$$\hat{p}_{i0} = F_0(a_i, \hat{\theta}_1, \cdots, \hat{\theta}_r) - F_0(a_{i-1}, \hat{\theta}_1, \cdots, \hat{\theta}_r)\ i = 2, \cdots, m-1,$$

由此可以看到，此问题又可以转化为多项分布的假设检验问题，其统计量为

$$\chi_n^{\,2}=\sum_{i=1}^{m}\frac{\left(N_i-n\hat{p}_{i0}\right)^2}{n\hat{p}_{i0}}\left(\text{或}\chi_n^{\,2}=\sum_{i=1}^{m}\frac{N_i^{\,2}}{n\hat{p}_{i0}}-n\right)$$

定理4.2 若 n 充分大(≥ 50), 则当 H_0 为真时(不论 H_0 中的分布属什么分布), 皮尔逊统计量总是渐近地服从自由度为 $m-r-1$ 的 χ^2 分布.

$$\chi_n^{\,2}=\sum_{i=1}^{m}\frac{\left(N_i-n\hat{p}_{i0}\right)^2}{n\hat{p}_{i0}}\overset{\text{近似}}{\sim}\chi^2(m-r-1)$$

此种检验法称为χ^2拟合优度检验法.

此类假设检验的拒绝域为

$$W = \{x : \chi_n^{\ 2} = \sum_{i=1}^{m} \frac{(N_i - n\hat{p}_{i0})^2}{n\hat{p}_{i0}} \geq \chi_\alpha^{\ 2}(m-r-1)\}$$

在使用χ^2检验法时, n要足够大, np_i不太小.

根据实践, 一般 $n \geq 50$, 每一个$np_i \geq 5$.

而且np_i最好在10以上，否则可以适当合并区间，使得np_i满足要求。

例3 在一试验中, 每隔一定时间观察一次由某种铀所放射的到达计数器上的 α 粒子数, 共观察了100次, 得结果如下表:

i	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12
N_i	1	5	16	17	26	11	9	9	2	1	2	1	0
A_i	A_0	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}

其中 N_i 是观察到有 i 个 α 粒子的次数. 从理论上

考虑 X 应服从泊松分布 $P\{X=i\}=\dfrac{e^{-\lambda}\lambda^i}{i!}$ $i=0,1,2,\cdots$

问 $P\{X=i\}=\dfrac{e^{-\lambda}\lambda^i}{i!}$ 是否符合实际?($\alpha=0.05$)

解 所求问题为: 在水平0.05下检验假设

H_0：总体 X 服从泊松分布

$$P\{X=i\}=\frac{e^{-\lambda}\lambda^i}{i!}\quad i=0,1,2,\cdots$$

由于在 H_0 中参数 λ 未具体给出, 故先估计 λ.

由最大似然估计法得 $\lambda=\bar{x}=4.2$,

根据题目中已知表格, $P\{X=i\}$有估计

$$\hat{p}_{i0} = \hat{P}\{X = i\} = \frac{e^{-4.2} 4.2^i}{i!} \quad i = 0, 1, 2, \cdots$$

如 $\hat{p}_{00} = \hat{P}\{X = 0\} = e^{-4.2} = 0.015,$

$$\hat{p}_{30} = \hat{P}\{X = 3\} = \frac{e^{-4.2} 4.2^3}{3!} = 0.185,$$

$$\hat{p}_{120} = \hat{P}\{X \geq 12\} = 1 - \sum_{i=1}^{11} \hat{p}_{i0} = 0.002,$$

具体计算结果见下表,

表1　　例3的χ^2拟合检验计算表

A_i	N_i	$\hat{p}_{i0}$	$n\hat{p}_{i0}$	$N_i^2 / n\hat{p}_{i0}$
A_0	1 } 6	0.015 } 0.078	1.5	4.615
A_1	5	0.063	6.3	
A_2	16	0.132	13.2	19.394
A_3	17	0.185	18.5	15.622
A_4	26	0.194	19.4	34.845
A_5	11	0.163	16.3	7.423
A_6	9	0.114	11.4	7.105
A_7	9	0.069	6.9	11.739
A_8	2	0.036	3.6	
A_9	1	0.017	1.7	5.538
A_{10}	2 } 6	0.007 } 0.065	0.7	
A_{11}	1	0.003	0.3	$\sum=106.281$
A_{12}	0	0.002	0.2	

其中有些 $n\hat{p}_{i0} < 5$的组予以合并, 使得每组均有 $np_i \geq 5$, 如表中第二列花括号所示.

并组后 $m = 8$, 故χ^2的自由度为$8-1-1=6$,

$$\chi_{\alpha}^2(m-r-1) = \chi_{0.05}^2(6) = 12.592 > 6.2815,$$

故接受H_0, 认为样本来自泊松分布总体.

例4 自1965年1月1日至1971年2月9日共2231天中，全世界记录到里氏震级4级和4级以上地震共162次，统计如下： $(\alpha = 0.05)$

(X表示相继两次地震间隔天数，Y表示出现的频数)

X	0 – 4	5 – 9	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	≥ 40
Y	50	31	26	17	10	8	6	6	8

试检验相继两次地震间隔天数 X 服从指数分布.

解 所求问题为: 在水平 0.05下检验假设

H_0：X 的概率密度 $f(x)=\begin{cases}\dfrac{1}{\theta}e^{-\frac{x}{\theta}}, & x>0\\ 0, & x\le 0.\end{cases}$

由于在 H_0 中参数 θ 未具体给出，故先估计 θ.

由最大似然估计法得 $\hat{\theta}=\bar{x}=\dfrac{2231}{162}=13.77,$

X 为连续型随机变量，

将 X 可能取值区间 $[0,\infty)$ 分为 $k=9$ 个互不重叠的子区间 $[a_i, a_{i+1})$ $i=1,2,\cdots,9.$ (见下页表)

表2　例4的χ^2拟合检验计算表

A_i	N_i	$\hat{p}_{i0}$	$n\hat{p}_{i0}$	$N_i^2 / n\hat{p}_{i0}$
$A_1: 0 \le x \le 4.5$	50	0.2788	45.1656	55.3519
$A_2: 4.5 < x \le 9.5$	31	0.2196	35.5752	27.0132
$A_3: 9.5 < x \le 14.5$	26	0.1527	24.7374	27.3270
$A_4: 14.5 < x \le 19.5$	17	0.1062	17.2044	16.7980
$A_5: 19.5 < x \le 24.5$	10	0.0739	11.9718	8.3530
$A_6: 24.5 < x \le 29.5$	8	0.0514	8.3268	7.6860
$A_7: 29.5 \le x \le 34.5$	6	0.0358	5.7996	6.2073
$A_8: 34.5 < x \le 39.5$	6	0.0248	4.0176 } 13.2192	14.8269
$A_9: 39.5 < x < \infty$	8	0.0568	9.2016	$\sum = 163.5633$

在 H_0 为真的前提下,

X 的分布函数的估计为 $\hat{F}(x)=\begin{cases}1-e^{-\frac{x}{13.77}}, & x>0\\ 0, & x\le 0.\end{cases}$

概率 $p_i=P(A_i)$有估计

$$\hat{p}_{i0}=\hat{P}(A_i)=\hat{P}\{a_i\le X<a_{i+1}\}=\hat{F}(a_{i+1})-\hat{F}(a_i),$$

如 $\hat{p}_{20}=\hat{P}(A_2)=\hat{P}\{4.5\le X<9.5\}$

$$=\hat{F}(9.5)-\hat{F}(4.5)=0.2196,$$

$$\hat{p}_{90} = \hat{F}(A_9) = 1 - \sum_{i=1}^{8} \hat{F}(A_i) = 0.0568,$$

$$\chi^2 = 163.5633 - 162 = 1.5633, \qquad m = 8, r = 1,$$

$$\chi^2_{\alpha}(m - r - 1) = \chi^2_{0.05}(6) = 12.592 > 1.5633,$$

故在水平0.05下接受H_0,

认为样本服从指数分布.

例5 下面列出了84个依特拉斯坎人男子的头颅的最大宽度(mm), 试验证这些数据是否来自正态总体?

$(\alpha = 0.1)$

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						

解 所求问题为检验假设

H_0：X 的概率密度 $f(x)=\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},-\infty<x<\infty.$

由于在 H_0 中参数 μ,σ^2 未具体给出,故先估计 μ,σ^2.

由最大似然估计法得 $\hat{\mu}=143.8,\quad \hat{\sigma}^2=6.0^2,$

将 X 可能取值区间$(-\infty,\infty)$分为7个小区间,

(见表3)

表3　　例5的 χ^2 拟合检验计算表

A_i	N_i	$\hat{p}_{i0}$	$n\hat{p}_{i0}$	$N_i^2 / n\hat{p}_{i0}$
$A_1: x \le 129.5$	1	0.0087	0.73	
$A_2: 129.5 < x \le 134.5$	4	0.0519	4.36	
	$\}$ 5 (merged)		$\}$ 5.09	4.91
$A_3: 134.5 < x \le 139.5$	10	0.1752	14.72	6.79
$A_4: 139.5 < x \le 144.5$	33	0.3120	26.21	41.55
$A_5: 144.5 < x \le 149.5$	24	0.2811	23.61	24.40
$A_6: 149.5 < x \le 154.5$	9	0.1336	11.22	10.02
$A_7: 154.5 < x < \infty$	3	0.0375	3.15	
			$\}$ 14.37	$\Sigma = 87.67$

在 H_0 为真的前提下， X 的概率密度的估计为

$$\hat{f}(x)=\frac{1}{\sqrt{2\pi}\times 6}e^{-\frac{(x-143.8)^2}{2\times 6^2}},-\infty<x<\infty.$$

概率 $p_i=P(A_i)$有估计

如 $\hat{p}_{20}=\hat{P}(A_2)=\hat{P}\{129.5\le x<134.5\}$

$$=\Phi\left(\frac{134.5-143.8}{6}\right)-\Phi\left(\frac{129.5-143.8}{6}\right)$$

$$=\Phi(-1.55)-\Phi(-2.38)=0.0519\ .$$

$$\chi_{\alpha}^2(m-r-1)=\chi_{0.1}^2(5-2-1)=\chi_{0.1}^2(2)=4.605>3.67,$$

故在水平0.1下接受H_0, 认为样本服从正态分布.

二、柯尔莫哥洛夫及斯米尔诺夫检验

1. χ^2 检验法的缺点

此种检验依赖于区间划分，划分的巧合可能导致检验的错误,例如

$$H_0: F(x) = F_0(x)\text{不成立}$$

但是当划分巧合时，也可能会出现

$$F(a_i) - F(a_{i-1}) = F_0(a_i) - F_0(a_{i-1}) = p_{i0}, i = 1, \cdots, m.$$

这样的结果不会影响皮尔逊统计量的值，因而可以导致接受错误的假设.

柯尔莫哥洛夫-斯米尔诺夫检验法更为精确，不仅可以检验经验分布是否服从某种理论分布，也可以检验两个样本是否来自同一个总体，但总体的分布要求连续.

2. 柯尔莫哥洛夫检验

设总体分布函数为$F(x)$，假定是 x 的连续函数，设$(X_1, X_2, ..., X_n)^T$是容量为n的样本，其经验分布函数为$F_n(x)$. 由格列汶科定理可得

$$P\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} |F_n(x) - F(x)| = 0\} = 1.$$

即$D_n = \sup\limits_{-\infty<x<+\infty} |F_n(x) - F(x)|$以概率$1$是无穷小的。

定理4.3 设F是连续的分布函数，则

$$P\{D_n = \sup_{-\infty<x<+\infty} |F_n(x)-F(x)| \le y+\frac{1}{2n}\}$$

$$=\begin{cases} 0, & y\le 0, \\ \displaystyle\int_{\frac{1}{2n}-y}^{\frac{1}{2n}+y}\int_{\frac{3}{2n}-y}^{\frac{13}{2n}+y}\cdots\int_{\frac{2n-1}{2n}-y}^{\frac{2n-1}{2n}+y} f(x_1,x_2,\cdots,x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n, & 0<y<\dfrac{2n-1}{2n}, \\ 1, & y\ge\dfrac{2n-1}{2n}, \end{cases}$$

其中$f(x_1,x_2,\cdots,x_n)=\begin{cases} n!, & 0<x_1<x_2<\cdots<x_n<1, \\ 0, & \text{其他}, \end{cases}$

定理4.4 设F是连续的分布函数，则

$$\lim_{n\to\infty} P\{\sqrt{n} \sup_{-\infty<x<+\infty} |F_n(x)-F(x)|< y\} = K(y)$$

$$=\begin{cases} 0, & y \le 0, \\ \sum\limits_{k=-\infty}^{\infty} (-1)^k \mathrm{e}^{-2k^2 y^2}, & y > 0. \end{cases}$$

上述两个定理证明略。它们将是柯尔莫哥洛夫检验法的理论基础.

假设检验的问题为

$$H_0: F(x) = F_0(x) \leftrightarrow H_1: F(x) \neq F_0(x),$$

其中$F(x)$为连续分布函数。设$(X_1, X_2, \cdots, X_n)^T$为来自总体$X$的样本,$(x_1, x_2, \cdots, x_n)^T$为其观测值，统计量选为 $D_n = \sup\limits_{-\infty<x<+\infty} |F_n(x) - F_0(x)|$

只要原假设不真，则统计量的值就会偏大，因而给定显著性水平α，可以选择临界值使得

$$P\{D_n > D_{n,\alpha}\} = \alpha$$

其中临界值$D_{n,\alpha}$可以查表(参见P303附表6).

则此检验法的拒绝域为

$$W = \{x : \hat{D}_n(x) > D_{n,\alpha}\}$$

当n >100时，利用极限分布定理4.4可得

$$D_{n,\alpha} \approx \frac{\lambda_{1-\alpha}}{\sqrt{n}},(\lambda_{1-\alpha}\text{可由附表7得到})$$

例6(p141例4.13) 某矿区煤层厚度的123个数据的频数分布如下表所示，试用柯尔莫哥洛夫检验法检验煤层的厚度是否服从正态分布？

组号	厚度间隔/m	组中值 x_i	频数 n_i	组号	厚度间隔	组中值 x_i	频数 n_i
1	0.20-0.50	0.35	1	6	1.70-2.00	1.85	24
2	0.50-0.80	0.65	6	7	2.00-2.30	2.15	25
3	0.80-1.10	0.95	5	8	2.30-2.60	2.45	19
4	1.10-1.40	1.25	12	9	2.60-2.90	2.85	20
5	1.40-1.70	1.55	19	10	2.90-3.20	3.05	2

解 用X表示煤层厚度，欲假设检验

$$H_0: \text{总体}X\text{服从正态分布}N(\mu,\sigma^2)\text{分布.}$$

由于参数未知，因而首先对参数进行估计

$$\hat{\mu} = \overline{x} = 1.884, \qquad \hat{\sigma}^2 = s_n^2 = 0.576^2$$

则 H_0：总体X服从正态分布$N(1.884, 0.576^2)$.

$$F(x_i) = P\{X \le x_i\} = P\{\frac{X - 1.884}{0.576} \le \frac{x_i - 1.884}{0.576}\}$$

$$= \Phi(\frac{x_i - 1.884}{0.576})$$

$$F_n(x_i) = \frac{v_n(x_i)}{n}$$

$$D_n = \sup_{-\infty < x_i < +\infty} |F_n(x_i) - F_0(x_i)| = 0.034$$

查附表7，取$\alpha = 0.05, D_{n,\alpha} \approx \dfrac{\lambda_{1-\alpha}}{\sqrt{123}} = \dfrac{1.36}{\sqrt{123}} = 0.123,$

显然 $D_{n,\alpha} \approx 0.123 > \hat{D}_n = 0.0343,$

因此接受原假设，认为煤层厚度服从正态分布.

表 4.7

组号	厚度间隔/m	频率 n_i/n	组上限 x	$u=\dfrac{x-1.884}{0.576}$	经验分布函数值 $F_n(x_i)$	理论分布函数值 $F(x_i)$	$\lvert F(x_i)-F_n(x_i)\rvert$
1	0.20～0.50	0.008	0.50	−2.40	0.008	0.008 2	0.000 2
2	0.50～0.80	0.049	0.80	−1.88	0.057	0.030 1	0.026 9
3	0.80～1.10	0.041	1.10	−1.36	0.098	0.086 9	0.011 1
4	1.10～1.40	0.098	1.40	−0.84	0.196	0.200 5	0.004 5
5	1.40～1.70	0.154	1.70	−0.32	0.350	0.374 5	0.024 5
6	1.70～2.00	0.195	2.00	0.20	0.545	0.579 3	0.034 3(D_n)
7	2.00～2.30	0.203	2.30	0.72	0.748	0.764 2	0.016 2
8	2.30～2.60	0.154	2.60	1.24	0.902	0.892 5	0.009 5
9	2.60～2.90	0.081	2.90	1.76	0.983	0.960 8	0.022 2
10	2.90～3.20	0.017	3.20	2.28	1.000	0.988 7	0.011 3

式 (4.31) 除可以用作检验法外，还可用来估计未知分布函数$F(x)$. 给定置信度$1-\alpha$,可由附表7查得$\lambda_{1-\alpha}$, n充分大时，$P\{D_n < \frac{\lambda_{1-\alpha}}{\sqrt{n}}\} \approx 1-\alpha$, 即

$$F_n(x) - \frac{\lambda_{1-\alpha}}{\sqrt{n}} < F(x) < F_n(x) + \frac{\lambda_{1-\alpha}}{\sqrt{n}}.$$

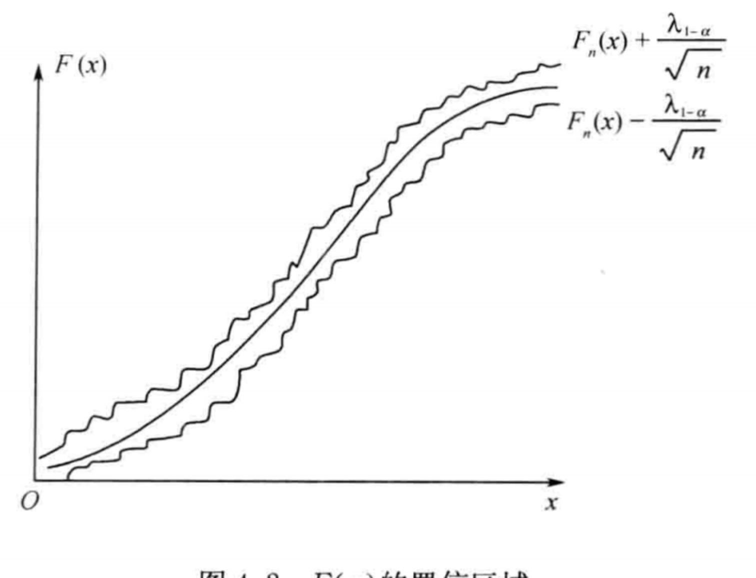


图 4.2 $F(x)$的置信区域

3. 斯米尔诺夫检验

假设检验的问题为

$$H_0: F(x) = G(x) \leftrightarrow H_1: F(x) \neq G(x),$$

其中$F(x)$、$G(x)$为两个总体的连续分布函数。设$(X_1, X_2, \cdots, X_{n_1})^T$为来自总体$F(x)$的样本,$(Y_1, Y_2, \cdots, Y_{n_2})^T$为来自总体$G(x)$的样本，并且假设两个总体独立，统计量选为

$$D_{n_1, n_2} = \sup_{-\infty < x < +\infty} |F_{n_1}(x) - G_{n_2}(x)|$$

其中$F_{n_1}(x)$与$G_{n_2}(x)$分别是两个总体的经验分布函数.

为了得到显著性水平下的拒绝域，需要如下定理：

定理4. 5 如果$F(x)=G(x)$,且$F(x)$是连续函数，则

$$P\{D_{n_1,n_2} = \sup_{-\infty<x<+\infty} |F_{n_1}(x) - G_{n_2}(x)| \le x\}$$

$$= \begin{cases} 0, & x \le \dfrac{1}{n}, \\ \displaystyle\sum_{j=-[\frac{n}{c}]}^{[\frac{n}{c}]} (-1)^j \frac{C_{2n}^{n-j}}{C_{2n}^{n}}, & \dfrac{1}{n} < x \le 1 \\ 1, & x > 1, \end{cases}$$

其中$c = -[-xn]$.

定理4. 6如果$F(x)=G(x)$,且F是连续函数，则

$$\lim_{n\to\infty} P\left\{\sqrt{\frac{n_1 n_2}{n_1 - n_2}} \sup_{-\infty<x<+\infty} |F_{n_1}(x) - G_{n_2}(x)| \le x\right\} = K(x)$$

$$= \begin{cases} 0, & x \le 0, \\ \sum\limits_{k=-\infty}^{\infty} (-1)^k \mathrm{e}^{-2k^2x^2}, & x > 0. \end{cases}$$

上述两个定理证明略。它们将是斯米尔诺夫检验法的理论基础.

只要原假设不真，则统计量的值就会偏大，因而给定显著性水平α，可以选择临界值使得

$$P\{D_{n_1,n_2} \geq D_{n_1,n_2,\alpha}\} = P\{D_{n_1,n_2} \geq D_{n,\alpha}\} = \alpha$$

其中$n = \dfrac{n_1 n_2}{n_1 + n_2}$，临界值$D_{n,\alpha} \approx \dfrac{\lambda_{1-\alpha}}{\sqrt{n}}$可以查表

(参见P304附表7)得到.

则此检验法的拒绝域为

$$W = \{x : \hat{D}_{n_1,n_2}(x) \geq D_{n,\alpha}\}$$

例7（p144例4. 14）在自动车床上加工某一零件，在工人刚接班时，先抽取150个零件作为样本，在自动车床工作两小时后，再抽取100个零件作为第二次样本，测得每个零件距离标准的偏差X，其数值列入下表，试比较两个样本是否来自同一总体?

偏差X的测量区间/μm	频数		偏差X的测量区间/μm	频数	
	n_{1j}	n_{2j}		n_{1j}	n_{2j}
[-15, -10)	10	0	[10, 15)	8	15
[-10, -5)	27	7	[15, 20)	1	1
[-5, 0)	43	17	[20, 25)	0	1
[0, 5)	38	30	$\sum$	$n_1=150$	$n_2=100$
[5, 10)	23	29			

解 欲假设检验

$$H_0: F(x) = G(x) \leftrightarrow H_1: F(x) \neq G(x),$$

计算两个样本对应的经验分布函数

$$F_{n_1}(x_i) = \frac{v_{n_1}(x)}{n_1} \qquad G_{n_2}(x) = \frac{v_{n_2}(x)}{n_2}$$

$$D_{n_1,n_2} = \sup_{-\infty<x<+\infty} |F_{n_1}(x) - G_{n_2}(x)| = 0.293$$

表 4.9

$x/\mu\mathrm{m}$	频数		累积频数		$F_{n_1} = \frac{n_1(x)}{n_1}$	$G_{n_2}(x) = \frac{n_2(x)}{n_2}$	$\lvert F_{n_1}(x) - G_{n_2}(x) \rvert$
	n_{1j}	n_{2j}	$n_1(x)$	$n_2(x)$			
−10	10	0	10	0	0.067	0.000	0.067
−5	27	7	37	7	0.247	0.070	0.177
0	43	17	80	24	0.533	0.240	0.293
5	38	30	118	54	0.787	0.540	0.247
10	23	29	141	83	0.940	0.830	0.110
15	8	15	149	98	0.993	0.980	0.013
20	1	1	150	99	1.000	0.990	0.010
25	0	1	150	100	1.000	1.000	0.000

查附表7，取$\alpha = 0.05, n = \dfrac{n_1 n_2}{n_1 + n_2} = 60,$

$$D_{n,\alpha} \approx \frac{\lambda_{1-\alpha}}{\sqrt{60}} = 0.17231$$

显然

$$D_{n,\alpha} \approx 0.17231 < \hat{D}_{n_1,n_2} = 0.293,$$

因此拒绝原假设，认为不是同一分布.

三、独立性检验

在社会调查中，调查人员怀疑男人和女人对某种提案将会有不同的反应，根据被调查者的性别和对某种提案的态度进行分类。

态度 性别	赞成	反对	弃权
男	1154	475	243
女	1083	442	362

三、独立性检验

1. 列联表的形式

假设有一个二元总体(X,Y)，将X,Y的取值范围分别划分为m个和k个互不相交的区间$A_1,A_2,\cdots,A_m$和$B_1,B_2,\cdots,B_k$。设从该总体中抽取一个容量为n的样本(X_1,Y_1)，$(X_2,Y_2),\cdots,(X_n,Y_n)$，用$n_{ij}$表示样本值中其$X$坐标落于$A_i$而$Y$坐标落于$B_j$中的个数$(i=1,2,\cdots,m,\ j=1,2,\cdots,k)$；又记

$$n_{i\bullet}=\sum_{j=1}^{k}n_{ij},\qquad n_{\bullet j}=\sum_{i=1}^{m}n_{ij}$$

上述问题可以用一个表格－列联表来表示如下

列联表

$X \backslash Y$		Y				$n_{i\bullet}=\sum_{j=1}^{k} n_{ij}$
		B_1	B_2	$\cdots$	B_k	
X	A_1	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1\bullet}$
	A_2	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	A_m	n_{m1}	n_{m2}	$\cdots$	n_{mk}	$n_{m\bullet}$
$n_{\bullet j}=\sum_{i=1}^{m} n_{ij}$		$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet k}$	

检验的问题为:

H_0 :总体的两个指标X和Y是相互独立的.

记 $p_{ij}=P\{X\in A_i, Y\in B_j\}, i=1,2,...,m, j=1,2,...,k,$

$p_{i\bullet}=P\{X\in A_i\}, i=1,2,...,m,$

$p_{\bullet j}=P\{Y\in B_j\}, i=1,2,...,k.$

那么

$$p_{i\bullet}=\sum_{j=1}^{k}p_{ij}, p_{\bullet j}=\sum_{i=1}^{m}p_{ij},$$

$$\sum_{i=1}^{m}p_{i\bullet}=\sum_{j=1}^{k}p_{\bullet j}=1.$$

如果H_0为真，则有$p_{ij} = p_{i\bullet} \bullet p_{\bullet j}$

因此列联表中的独立性检验就是要检验假设

$$H_0: p_{ij} = p_{i\bullet} \bullet p_{\bullet j}, i = 1,2,...,m, j = 1,2,...,k.$$

$p_{i\bullet}$和$p_{\bullet j}$是未知的，想要用χ^2检验验证H_0，须首先按照最大似然估计法，从样本中估计这些未知参数的值，结果如下

$$\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n}, \quad i = 1,2,...m; \quad \hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}, \quad j = 1,2,...k.$$

统计量选择为

$$\chi^2 = n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{n_{i\bullet}n_{\bullet j}}$$

$$\chi^2 = n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{n_{i\bullet}n_{\bullet j}} \overset{\text{近似}}{\sim} \chi^2((m-1)(k-1))$$

拒绝域为

$$W=\{x:\chi^2 = n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{n_{i\bullet}n_{\bullet j}} \ge \chi^2_{\alpha}((m-1)(k-1))\}$$

例7(p143例4.15)调查339名50岁以上的吸烟者与慢性气管炎的关系，结果如下表

Y / X	患慢性气管炎者	未患慢性气管炎者	合计	患病率%
吸烟	43	162	205	21.0
不吸烟	13	121	134	9.7
合计	56	283	339	16.5

试问吸烟者与不吸烟者患慢性支气管炎疾病是否有所不同(α=0.01)?

解　检验的问题为：

H_0 : 总体的两个指标X和Y是相互独立的.

统计量为　$$\chi^2 = n\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{n_{i\bullet}n_{\bullet j}}$$

观察值为　$$\hat{\chi}^2 = n\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{n_{i\bullet}n_{\bullet j}} = 7.48$$

上侧分位数　$\chi^2_{0.01}((m-1)(k-1)) = \chi^2_{0.01}(1) = 6.635$

显然　$\chi^2_{0.01}(1) = 6.635 < \hat{\chi}^2 = 7.48$

因而拒绝原假设，即认为慢性气管炎与吸烟有关

Thank You!