§7.4　非参数假设检验
(分布的假设检验)

1964年我国某研究所对一批数据提出问题:“苏联的轴承寿命服从对数正态分布，美国的轴承寿命服从威布尔分布，我国的轴承寿命服从什么分布?”

经过多年研究，最后确定我国的轴承寿命服从两参数威布尔分布。随后也选定了估计其中两个参数的估计方法——最好线性无偏估计(BLUE)。

这类问题在国内外经常出现，又如一种新的电子元件设计和制造出来了，它的平均寿命的0.95单侧置信下限是多少?这对其销售量影响很大，因此就先要确定该元件的寿命分布。

分布的检验问题一般只给出原假设，因为它所涉及的备择假设很多，不可能全部列出，也说不清楚。如原假设为正态分布，那么一切非正态分布都可以作为备择假设。

一、正态性检验

二、χ^2拟合优度检验

三、科尔莫戈罗夫(Kolmogorov)拟合检验

四、科尔莫戈罗夫-斯米尔诺夫
两子样检验（K-S检验）

一、正态性检验

一个样本是否来自正态分布的检验称为正态性检验。在这种检验中“样本来自正态分布”是作为原假设而设立的，在H_0为真时，人们根据正态分布特性构造一个统计量或一种特定方法，观察其是否偏离正态性。若偏离到一定程度就拒绝原假设，否则就接受原假设。

- **正态概率图**

根据正态分布性质构造一张图，样本在其上明显不在一条直线上，就认为该样本偏离正态性，从而拒绝正态性假设。

- **夏皮洛-威尔克(Shapiro-Wilk)检验($8 \le n \le 50$)**

- **爱泼斯-普利(Epps Pully)检验($n \ge 8$)**

正态概率图是一个简单、快速检验正态性的方法，值得首先使用。后两个检验方法对各种非正态分布偏离正态性较为有效，已被国际标准化组织(ISO)认可，形成国际标准ISO 5479 - 1997， 我国也采用这两种方法，形成国家标准GB/T4882 - 2001, 推广使用。

四段

二、χ^2拟合检验法

χ^2拟合优度检验是著名英国统计学家老皮尔逊(K. Pearson, 1857 - 1936年)于1900年结合检验分类数据的需要而提出的，然后又用于分布的拟合检验与列联表的独立性检验上去。

1、总体可分为有限类，其分布不含未知参数

$H_0 : X \sim F(x) = F_0(x).$（分布的检验问题一般只给出原假设）

基本思想：

将总体X的值域分成k个互不相容的区间$A_1 = (a_0, a_1], A_2 = (a_1, a_2], \cdots, A_k = (a_{k-1}, a_k],$

H_0成立时，$p_i = P(A_i) = P(a_{i-1} < X \le a_i) = F_0(a_i) - F_0(a_{i-1}),$

设n个样本观测值落入区间A_i中的频数为n_i.

则事件A_i出现的频率为$\dfrac{n_i}{n}$.

$p_i \leftrightarrow \dfrac{n_i}{n}$

渐近

在 n 次试验中，事件 A_i 出现的频率 $\dfrac{n_i}{n}$ 与 p_i 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.

np_i — 理论频数

n_i — 频数

Pearson χ^2 - 统计量

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\sim\chi^2(k-1)(n\to\infty)\quad(Th7.1)$$

$$P(\chi^2 \geq \chi^2_{1-\alpha}(k-1)) = \alpha$$

k=2时怎么证明？

Pearson χ^2 −统计量 $\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\sim\chi^2(k-1)(n\to\infty)$ (Th7.1)

检验步骤:

(1)根据题意给出$H_0: X$服从某一分布;

(2)选取Pearson χ^2检验统计量;

(3)在给定α下,查表得$\chi^2_{1-\alpha}(k-1)$,即$P(\chi^2\geq\chi^2_{1-\alpha}(k-1))=\alpha$

(4)在观测值$(x_1,x_2,\cdots,x_n)$下计算出χ_0^2,

若$\chi_0^2\geq\chi^2_{1-\alpha}(k-1)$,则拒绝$H_0$,否则接受$H_0$

注意 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$np_i \geq 5$.

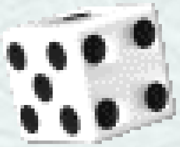
χ^2拟合检验法一般要求$np_i \geq 5$（只有两个分类），或者至少有80%的$np_i \geq 5$（有两个以上的分类）.

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

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),

取 $A_i=\{X=i\}\quad(i=1,2,\cdots,6)$

在 H_0 为真的前提下, $p_i=P(A_i)=\frac{1}{6},\quad(i=1,2,\cdots,6)$

$$\chi^2=\sum_{i=1}^{6}\frac{(n_i-np_i)^2}{np_i}$$

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

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

$\chi^2 = 20.16,$ 自由度为 $6-1=5,$

查表得 $\chi^2_{1-0.05}(5) = 11.071,$ $\chi^2 = 20.16 > 11.071,$

所以拒绝 $H_0,$

认为这颗骰子的六个面不是匀称的.

例2:

一位大学教育专家想知道,对于小型私立大学的四个年级,各年级学生数是否相等?为此,从某校随机抽取4500个本科生,一年级1200人,二年级1100人,三年级1150人,四年级1050人,$\alpha = 0.05$.试利用这些资料做出χ^2拟合检验.

要检验假设H_0：各年级人数相等

年级	频数 n_i	p_i	np_i	$\frac{(n_i-np_i)^2}{np_i}$
一	1200	0.25	1125	5
二	1100	0.25	1125	0.56
三	1150	0.25	1125	0.56
四	1050	0.25	1125	5
合计	4500	1.00	4500	11.12

查表知 $\chi^2_{1-0.05}(4-1)=7.815<11.12$, 故不能认为各年级人数相等.

例3 下表列出了某一地区在夏季的一个月中由100个气象站报告的雷暴雨的次数.

i	0	1	2	3	4	5	≥ 6
n_i	22	37	20	13	6	2	0

其中 n_i 是报告雷暴雨次数为 i 的气象战数. 试用 χ^2 拟合检验法检验雷暴雨的次数 X 是否服从均值 $\lambda = 1$ 的泊松分布(取显著性水平 $\alpha = 0.05$).

解 按题意需检验假设

$$H_0: \quad P\{X=i\}=\frac{\lambda^i e^{-\lambda}}{i!}=\frac{e^{-1}}{i!}, \quad i=0,1,\cdots.$$

在H_0下X所有可能取的值为 $\Omega=\{0,1,2,\cdots\}$, 将Ω分成如表所示的两两不相交的子集 $A_0, A_1, \cdots, A_6$, 则有$P\{X=i\}$为

$$p_i=P\{X=i\}=\frac{e^{-1}}{i!}, \quad i=0,1,\cdots 5.$$

例如 $p_0=P\{X=0\}=e^{-1}=0.36788$

$$p_3 = P\{X = 3\} = \frac{e^{-1}}{3!} = 0.06131$$

$$p_6 = P\{X \geq 6\} = 1 - \sum_{i=0}^{5} p_i = 0.059$$

A_i	n_i	p_i	np_i	$n_i^2/(np_i)$
$A_0:\{X=0\}$	22	e^{-1}	36.788	13.16
$A_1:\{X=1\}$	37	e^{-1}	36.788	37.21
$A_2:\{X=2\}$	20	$e^{-1}/2$	18.394	21.75
$A_3:\{X=3\}$	13	$e^{-1}/6$	6.131	
$A_4:\{X=4\}$	6	$e^{-1}/24$	1.533	
$A_5:\{X=5\}$	2	$e^{-1}/120$	0.307 } 1.899	33.7
$A_6:\{X\geq 6\}$	0	$1-\sum_{i=0}^{5} p_i$	0.059	

还不够大！

A_i	n_i	p_i	np_i	$n_i^2/(np_i)$
$A_0:\{X=0\}$	22	e^{-1}	36.788	13.16
$A_1:\{X=1\}$	37	e^{-1}	36.788	37.21
$A_2:\{X=2\}$	20	$e^{-1}/2$	18.394	21.75
$A_3:\{X=3\}$	13	$e^{-1}/6$	6.131	
$A_4:\{X=4\}$	6	$e^{-1}/24$	1.533 $\}$ 8.03	54.92
$A_5:\{X=5\}$	2	$e^{-1}/120$	0.307	
$A_6:\{X\ge 6\}$	0	$1-\sum_{i=0}^{5} p_i$	0.059	

127.04

要重新计

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

计算结果如表所示，其中有些 $np_i < 5$的组予以适当合并，使得每组均有 $np_i \geq 5$，如表中第4列花括号所示，并组后 $k = 4$，

χ^2的自由度为 $k-1 = 4-1 = 3$

$$\chi^2_{0.95}(k-1) = \chi^2_{0.95}(3) = 7.815$$

现在

$$\chi^2 = 27.04 > 7.815$$

故在显著性水平 0.05 下拒绝 H_0，认为样本不是来自均值 $\lambda = 1$的泊松分布.

2、总体可分为有限类，但其分布含未知参数

$H_0: X \sim F(x) = F_0(x).$

设总体X的真实分布函数为$F_0(x;\theta_1,\cdots,\theta_m)$,

其中$\theta_1,\cdots,\theta_m$为m个未知参数.

用$\theta_1,\cdots,\theta_m$的极大似然估计$\hat{\theta}_{1L},\cdots,\hat{\theta}_{mL}$代替$\theta_1,\cdots,\theta_m$,

取$A_1=(a_0,a_1], A_2=(a_1,a_2],\cdots, A_k=(a_{k-1},a_k]$,

记样本观测值落入A_i中的频数为n_i,总体X落入A_i中的概率用$\hat{p}_i$代替,

$$\hat{p}_i = F_0(a_i;\ \hat{\theta}_{1L},\cdots,\hat{\theta}_{mL}) - F_0(a_{i-1};\ \hat{\theta}_{1L},\cdots,\hat{\theta}_{mL}).$$

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-n\hat{p}_i)^2}{n\hat{p}_i} \sim \chi^2(k-m-1)(n\to\infty) \quad (Th7.2)$$

例4：

下表是上海1875年到1955年的81年间, 根据其中的63 年观察记录到的一年中(5月到9月)下暴雨次数的整理资料：

一年中暴雨次数i	0	1	2	3	4	5	6	7	8	≥ 9
实际年数n_i	4	8	14	19	10	4	2	1	1	0

试检验一年中下暴雨的次数X 是否服从*Possion* 分布.($\alpha = 0.05$)

$H_0: \xi \sim P(\lambda)$, 极大似然估计得$\hat{\lambda}_L = \bar{\xi} = 2.86.$

概率论与数理统计

$$\hat{p}_1 = \frac{2.86^0}{0!}e^{-2.86},\quad \hat{p}_2 = \frac{2.86^1}{1!}e^{-2.86},\cdots$$

区间	频数 n_i	$\hat{p}_i = F_0(a_i) - F_0(a_{i-1})$	$n\hat{p}_i$	$(n_i - n\hat{p}_i)^2$	$(n_i - n\hat{p}_i)^2 / n\hat{p}_i$
[0,0]	4	0.057	3.591	0.167	0.0465
(0,1]	8	0.164	10.332	5.438	0.526
(1,2]	14	0.234	14.742	0.551	0.037
(2,3]	19	0.230	14.49	20.34	1.404
(3,4]	10	0.164	10.332	0.110	0.011
(4,5]	4	0.094	5.922	3.694	0.624
(5,6]	2	0.045	2.835	0.697	0.246
(6,7]	1	0.010	0.63	0.1369	0.217
$(7,\infty)$	1	0.002	0.126	0.764	6.06
合计	63	1	63		9.1715

查表知 $\chi^2_{1-0.05}(9-1-1) = 14.067 > 9.1715$, 故可以认为服从$Poisson$分布.

区间	频数 n_i	$\hat{p}_i = F_0(a_i) - F_0(a_{i-1})$	$n\hat{p}_i$	$(n_i - n\hat{p}_i)^2$	$(n_i - n\hat{p}_i)^2 / n\hat{p}_i$
[0,1]	12	0.221	13.923	3.6979	0.2656
(1,2]	14	0.234	14.742	0.551	0.037
(2,3]	19	0.230	14.49	20.34	1.404
(3,4]	10	0.164	10.332	0.110	0.011
$(4, \infty)$	8	0.151	9.513	2.289	0.241
合计	63	1	63		1.9466

查表知 $\chi^2_{1-0.05}(5-1-1) = 7.815 > 1.9466$, 故可以认为服从*Poisson*分布.

例5：($p352$)

研究混凝土抗压强度的分布.200件混凝土制件的抗压强度以分组形式列出.

压强区间kg/cm^2	频数n_i	组中值
190 ~ 200	10	195
200 ~ 210	26	205
210 ~ 220	56	215
220 ~ 230	64	225
230 ~ 240	30	235
240 ~ 250	14	245

试问抗压强度ξ是否服从正态分布.($\alpha = 0.05$)

$H_0: \xi \sim F(x) = F_0(x) = \Phi(x;\mu,\sigma^2)$, 其中$\mu,\sigma^2$为未知参数.

μ和σ^2的极大似然估计为$\hat{\mu}_L = \overline{x}$和$\hat{\sigma}_L^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

设x_i^*表示第i组区间的中点值，计算

$$\hat{\mu}_L = \frac{1}{n}\sum_{i=1}^{n} x_i^* n_i = 221,\quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^* - \overline{x})^2 n_i = 152,\ \hat{\sigma} = 12.33.$$

在正态分布$N(221, 12.33^2)$的分布下，计算每个区间理论概论值的估计：

$$\hat{p}_i = P(a_{i-1} < \xi \le a_i) = P\left(\frac{a_{i-1} - 221}{12.33} < \frac{\xi - 221}{12.33} \le \frac{a_i - 221}{12.33}\right)$$

$$= \Phi_0\left(\frac{a_i - 221}{12.33}\right) - \Phi_0\left(\frac{a_{i-1} - 221}{12.33}\right) =: \Phi_0(u_i) - \Phi_0(u_{i-1}).$$

区间 $(a_{i-1}, a_i]$	频数	标准化区间 $(u_{i-1}, u_i]$	$\hat{p}_i = \Phi_0(u_i) - \Phi_0(u_{i-1})$	$n\hat{p}_i$	$(n_i - n\hat{p}_i)^2$	$(n_i - n\hat{p}_i)^2 / n\hat{p}_i$
(-∞,200]	10	(-∞,-1.7]	0.045	9.0	1.00	0.11
(200,210]	26	(-1.7,-0.89]	0.142	28.4	5.76	0.20
(210,220]	56	(-0.89,-0.08]	0.281	56.2	0.04	0.00
(220,230]	64	(-0.08,0.73]	0.299	59.8	17.64	0.29
(230,240]	30	(0.73,1.54]	0.171	34.2	17.64	0.52
(240, ∞)	14	(1.54, ∞)	0.062	12.4	2.56	0.23
合计	200		1.000	200		1.35

查表知 $\chi^2_{1-0.05}(6-2-1) = 7.815 > 1.35$, 故可以认为服从正态分布.

练习：在股票投资中有一个流行的说法: 盈利、持平和亏损的比例为1 : 2: 7。2003 年2月8日《上海青年报》第16版上发表了一个调查数据，在1270位被调查的股民中盈利者273人，持平者240人，亏损者757人。这些调查数据能否认可流行的说法?

这个问题归结为检验如下假设的问题:

H_0 : P(盈)=0.1, P(平)=0.2，P(亏)=0. 7

股民盈亏数据的 χ^2 检验计算表

i	O_i	p_i	$E_i = np_i$	$\lvert O_i - E_i \rvert$	$\frac{(O_i - E_i)^2}{E_i}$
1	273	0.1	127	145	165.55
2	240	0.2	254	14	0.77
3	757	0.7	889	132	19.60
和	1 270	1.0	1 270	/	185.92

$$\chi^2 = 188.21 > \chi^2_{0.95}(2) = 5.99$$

对“拟合优度”作一些说明：

在分布检验中常要问:

(1) 实际数据与理论分布是否符合?

在分布检验中对原假设H_0作出“拒绝”或“接受”的判断。

(2) 若符合，符合程度如何? 能否提供一个(介于0~1之间的)数字作为符合程度的数量指标?

老皮尔逊研究了这个问题，找到了这个数量指标，并称之为“拟合优度”(goodness of fit)。

拟合优度——分布检验中的p值。

在例5中

$$p = P\{\chi^2 \geq \chi_0\} = P\{\chi^2 \geq 1.35\} = 0.72$$

拟合优度(即p值)越大，表示实际数据与理论分布拟合得越好，该理论分布就获得更多实际数据支持。而显著性水平a只是人们设置的一个门槛，当拟合优度低于a时拒绝H_0，拟合优度越低，人们放弃H_0越放心;当拟合优度高于a时，接受H_0，若取a=0.05，当p=0.06或p=0.90时虽都接受H_0，但后者使数据对理论分布的支持比前者强得多，前者勉强过关，后者接近完美。

三、科尔莫戈罗夫(Kolmogorov)拟合检验

克服了卡方检验依赖于区间划分的缺点，但要求总体分布必须假定为连续.

定理7.3 (P355) $D_n = \sup\limits_{x \in R} \left| F_n(x) - F(x) \right|$

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

$$= \begin{cases} 0, & \lambda < 0 \\ \int_{\frac{1}{2n}-\lambda}^{\frac{1}{2n}+\lambda} \int_{\frac{3}{2n}-\lambda}^{\frac{3}{2n}+\lambda} \cdots \int_{\frac{2n-1}{2n}-\lambda}^{\frac{2n-1}{2n}+\lambda} f(y_1, \cdots, y_n)\, dy_1 \cdots dy_n, & 0 \leqslant \lambda < \dfrac{2n-1}{2n} \\ 1, & \lambda \geqslant \dfrac{2n-1}{2n} \end{cases} \tag{7.33}$$

其中

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

$$\lim_{n\to\infty} P\{\sqrt{n}D_n \le \lambda\}$$

$$\to K(\lambda) = \begin{cases} \sum_{j=-\infty}^{+\infty} (-1)^j \exp(-2j^2\lambda^2), & \lambda > 0 \\ 0, & \lambda \le 0 \end{cases}$$

步骤：(P356)

$$H_0: \text{X服从某连续型分布} F(x)$$

经验分布函数：

$$F_n(x)=\begin{cases}0, & x<x_{(1)}\\ \dfrac{j-0.5}{n}, & x_{(j)}\le x<x_{(j+1)}, j=1,2,\dots,n-1\\ 1, & x\ge x_{(n)}\end{cases}$$

$$D_n=\sup_{x\in R}\left|F_n(x)-F(x)\right|$$

$$=\max_{1\le j\le n}\left|F_n\left(x_{(j)}\right)-F\left(x_{(j)}\right)\right|,\left|F_n\left(x_{(j-1)}\right)-F\left(x_{(j)}\right)\right|$$

四、科尔莫戈罗夫-斯米尔诺夫两子样检验（K-S检验）

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

定理7.4 (P361) $D_{n_1,n_2} = \sup_{x \in R} \left| F_{1n_1}(x) - F_{1n_2}(x) \right|$

$$\lim_{n_1,n_2 \to \infty} P\{\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1,n_2} \le \lambda\}$$

$$\to K(\lambda) = \begin{cases} \sum_{j=-\infty}^{+\infty} (-1)^j \exp(-2j^2\lambda^2), & \lambda > 0 \\ 0, & \lambda \le 0 \end{cases}$$

作业： p378. 7.23，7.24