

# 4.3 非参数假设检验法

## 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。但实际中，有时总体服从什么分布，一般预先无法知晓，因而需要对总体的分布进行各种假设。

本节将主要讨论对总体分布的假设检验问题，此类问题通常称为非参数统计方法。

本文主要介绍其中常见的3种方法。

一、 $\chi^2$  拟合优度检验

二、柯尔莫哥洛夫<sup>及</sup>斯  
米尔诺夫检验

\*三、独立性检验



# 一、 $\chi^2$ 拟合优度检验法

卡方拟合优度检验是著名英国统计学家老皮尔逊于1900年结合检验分类数据的需要而提出，然后又用于分布的拟合检验和独立性检验。

卡方拟合优度检验又称卡方检验。



## 1. $\chi^2$ 检验法的定义

这是在总体的分布未知 的情况下,根据样本  $X_1, X_2, \dots, X_n$  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) 这里备择假设  $H_1$  可不必写出.

(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  $P\{X = x_i\} = p_i, i = 1, 2, \dots$ .

(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  $\chi^2$  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. 问题的背景

19世纪生物学家孟德尔 (Mendel) 按颜色与形状把豌豆分为4类,

$A_1$  := 黄而圆的,  $A_2$  = 青而圆的

$A_3$  = 黄而有角的,  $A_4$  = 青而有角的

孟德尔根据遗传学的理论指出, 这4类豌豆个数之比为9:3:3:1, 这相当于说任取一粒豌豆, 它属这4类的概率分别为:

$$p_1 = \frac{9}{16}, \quad p_2 = \frac{3}{16}, \quad p_3 = \frac{3}{16}, \quad p_4 = \frac{1}{16},$$

孟德尔在一次收获的 $n$ 等于556粒豌豆的观察中，发现4类豌豆的个数分别为

$$N_1 = 315, \quad N_2 = 108, \quad N_3 = 101, \quad N_4 = 32,$$

$$\text{显然, } N_1 + N_2 + N_3 + N_4 = 556 = n.$$

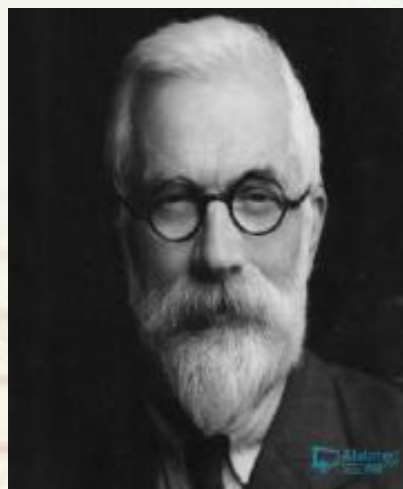
但由于样本值的随机性，诸观察数  $N_i$  不会恰好呈现9:3:3:1的比例，因此就需要根据这些观察数据对孟德尔的遗传学说进行统计检验。

孟德尔的实践向统计学家提出一个很有意义的问题：一组实际数据与一个给定的多项分布的拟合程度。

老皮尔逊研究了这个问题，提出了卡方拟合优度检验，解决了这类问题。后经过英国统计学家费希尔推广，这个检验更加趋于完善，最终开创了假设检验的理论与实践。



Karl Pearson,  
1857~1936



R. A. Fisher,  
1890~ 1962



## 2. 多项分布的 $\chi^2$ 检验法构造

设总体 $X$ 为离散型分布，其分布律为

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

设 $(X_1, X_2, \dots, X_n)^T$ 为来自总体 $X$ 的样本， $(x_1, x_2, \dots, x_n)^T$ 为其观测值， $N_i$ 表示 $(X_1, X_2, \dots, X_n)^T$

中取值为 $x_i$ 的个数，且 $\sum_{i=1}^m N_i = n, (N_1, N_2, \dots, N_m)^T$

分布为

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

多项分布

假设检验的问题为

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

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.

则频率  $\frac{N_i}{n}$  与  $p_{i0}$  之间的差异程度可反映出  $(p_{10}, p_{20}, \dots, p_{m0})$  是否是总体的真实分布。

频率  $\frac{N_i}{n}$  与  $p_{i0}$  间差异  $\longleftrightarrow$  频数  $N_i$  与  $np_{i0}$  间差异

皮尔逊提出用如下统计量

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

-----皮尔逊统计量

来衡量  $\frac{N_i}{n}$  与  $p_{i0}$  之间的差异程度。

$$\begin{aligned} \sum_{i=1}^m \frac{N_i^2 - 2N_i np_{i0} + n^2 p_{i0}^2}{np_{i0}} &= \sum_{i=1}^m \frac{N_i^2}{np_{i0}} - \sum_{i=1}^m 2N_i + \sum_{i=1}^m np_{i0} \\ &= \sum_{i=1}^m \frac{N_i^2}{np_{i0}} - 2n + n \sum_{i=1}^m p_{i0} = \sum_{i=1}^m \frac{N_i^2}{np_{i0}} - 2n + n = \sum_{i=1}^m \frac{N_i^2}{np_{i0}} - n \end{aligned}$$

### 3.皮尔逊定理

定理4.1 若  $n$  充分大( $\geq 50$ ), 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  $\chi^2$  分布.

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

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  $\alpha$  下拒绝  $H_0$ , 否则就接受  $H_0$ .

**注意:** 定理要求  $n$  足够大,  $np_{i0}$  不要太小,  $n \geq 50, np_{i0} \geq 5$

## 解决背景问题

$$N_1 = 315, \quad N_2 = 108, \quad N_3 = 101, \quad N_4 = 32,$$

$$N_1 + N_2 + N_3 + N_4 = 556 = n.$$

$$p_1 = \frac{9}{16}, \quad p_2 = \frac{3}{16}, \quad p_3 = \frac{3}{16}, \quad p_4 = \frac{1}{16},$$

$$\begin{aligned} \chi_n^2 = \sum_{i=1}^4 \frac{(N_i - np_{i0})^2}{np_{i0}} &= \frac{\left(315 - 556 \times \frac{9}{16}\right)^2}{556 \times \frac{9}{16}} + \frac{\left(108 - 556 \times \frac{3}{16}\right)^2}{556 \times \frac{3}{16}} \\ &+ \frac{\left(101 - 556 \times \frac{3}{16}\right)^2}{556 \times \frac{3}{16}} + \frac{\left(32 - 556 \times \frac{1}{16}\right)^2}{556 \times \frac{1}{16}} = 0.47 < 7.81 \end{aligned}$$

自由度为  $4-1=3$ ,  
查得  $\chi_{0.05}^2(3) = 7.81$ ,

所以接受  $H_0$ , 孟德尔的遗传学说是可接受的.



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  $\alpha = 0.05$ )

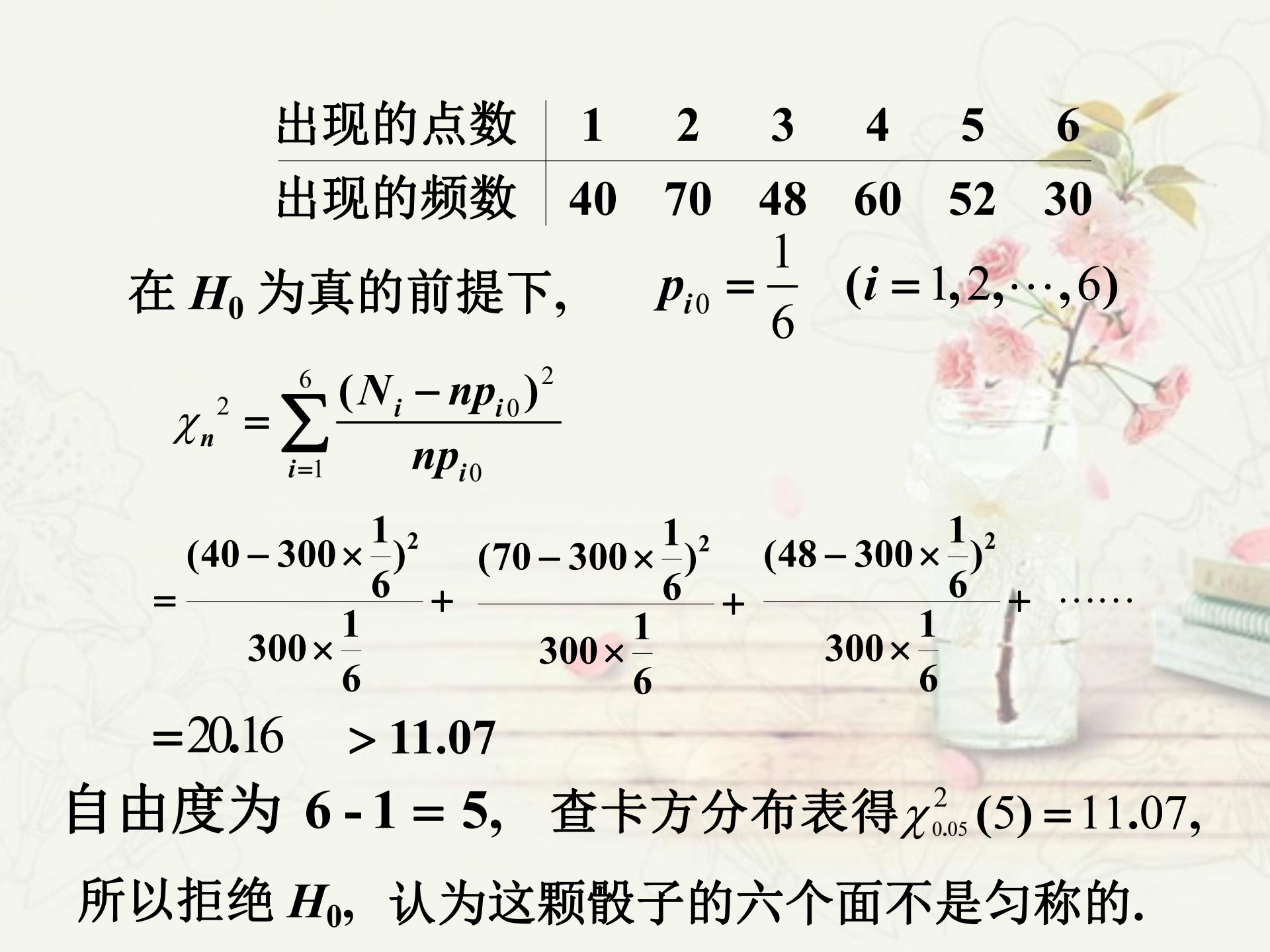
**解** 根据题意需要检验假设

$H_0$ : 这颗骰子的六个面是匀称的.

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

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),





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

在  $H_0$  为真的前提下,  $p_{i0} = \frac{1}{6} \quad (i = 1, 2, \dots, 6)$

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

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

$$= 20.16 > 11.07$$

自由度为  $6 - 1 = 5$ , 查卡方分布表得  $\chi_{0.05}^2(5) = 11.07$ ,

所以拒绝  $H_0$ , 认为这颗骰子的六个面不是匀称的.

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	$\geq 5$
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( $\alpha=0.05$ )?

**解** 从题意可知，该总体服从几何分布，

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

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

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

$$m=5,$$

计算皮尔逊统计量可得：

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

查表可得

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

显然

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

因而接受原假设，黑球白球个数相等.



## 4. 一般分布的 $\chi^2$ 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

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

拒绝域为

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

## 5. 分布中含有未知参数的 $\chi^2$ 检验法

假设检验的问题为

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

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

设  $(X_1, X_2, \dots, X_n)^T$  为来自总体  $X$  的样本,  $(x_1, x_2, \dots, x_n)^T$  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  $F_0(x, \hat{\theta}_1, \dots, \hat{\theta}_r)$ , 令

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

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

$$i = 2, \dots, m-1,$$

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。

**定理4.2** 若  $n$  充分大( $\geq 50$ ), 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  $\chi^2$  分布.

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

未知参数个数

此种检验法称为  $\chi^2$  拟合优度检验法.

此类假设检验的拒绝域为

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



**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  $\alpha$  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	$\geq 12$
$N_i$	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  $\alpha$  粒子的次数.从理论上

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

问  $P\{X=i\} = \frac{e^{-\lambda} \lambda^i}{i!}$  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

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

由于在  $H_0$  中参数  $\lambda$  未具体给出，故先估计  $\lambda$ 。

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

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

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 $\chi^2$  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	0.015	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	5.538
$A_9$	1	0.017	1.7	
$A_{10}$	2	0.007	0.7	
$A_{11}$	1	0.003	0.3	
$A_{12}$	0	0.002	0.2	$\Sigma = 106.281$

其中有些  $n\hat{p}_{i0} < 5$  的组予以合并, 使得每组均有  $np_i \geq 5$ .

并组后  $m = 8$ , 故  $\chi^2$  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	$\geq 40$
$Y$	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 
$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}}, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

由于在  $H_0$  中参数  $\theta$  未具体给出, 故先估计  $\theta$ .

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

$X$  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  $[a_i, a_{i+1})$   $i = 1, 2, \dots, 9$ . (见下页表)

表2 例4的 $\chi^2$  拟合检验计算表

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

在  $H_0$  为真的前提下,

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

概率  $p_i = P(A_i)$  有估计

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

$$\begin{aligned} \text{如 } \hat{p}_{20} &= \hat{P}(A_2) = \hat{P}\{4.5 \leq X < 9.5\} \\ &= \hat{F}(9.5) - \hat{F}(4.5) = 0.2196, \end{aligned}$$

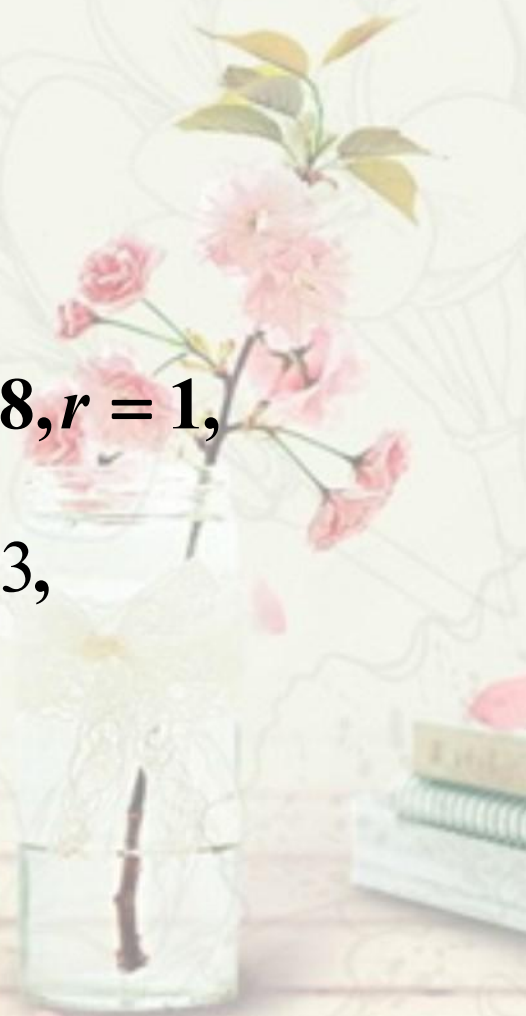
$$\hat{p}_{90} = \hat{F}(A_9) = 1 - \sum_{i=1}^8 \hat{F}(A_i) = 0.0568,$$

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

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

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

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

由于在  $H_0$  中参数  $\mu, \sigma^2$  未具体给出, 故先估计  $\mu, \sigma^2$ .

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

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  $\chi^2$  拟合检验计算表

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

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

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

$$\begin{aligned} &= \Phi\left(\frac{134.5 - 143.8}{6}\right) - \Phi\left(\frac{129.5 - 143.8}{6}\right) \\ &= \Phi(-1.55) - \Phi(-2.38) = 0.0519. \end{aligned}$$

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

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. $\chi^2$ 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

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

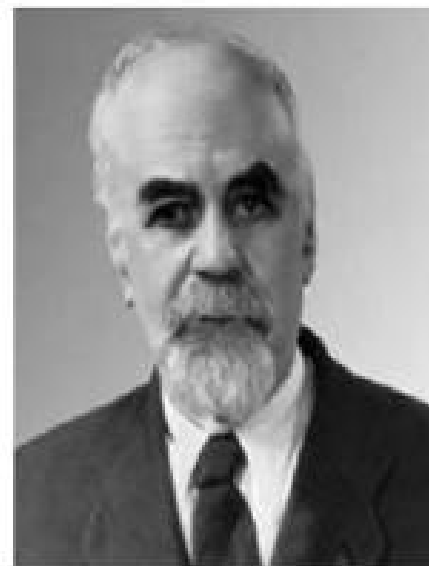
这样的结果不会影响皮尔逊统计量的值，因而可以导致接受错误的假设。

本节将介绍柯尔莫哥洛夫—斯米尔诺夫检验法，  
柯尔莫哥洛夫检验法可检验经验分布是否服从某种理论分布。  
斯米尔诺夫检验法可检验两个样本是否服从同一分布。



柯尔莫哥洛夫, A. H.

苏联最伟大的数学家之一，也是20世纪最伟大的数学家之一  
<http://maths.hust.edu.cn/info/1187/3361.htm>



斯米尔诺夫（1887.6.10-1974.2.11）苏联数学家、物理学家



## 2. 柯尔莫哥洛夫检验

首先回忆格里汶科定理（1.1节），

对于任一实数  $x$ , 当  $n \rightarrow \infty$  时,  $F_n(x)$  以概率 1 一致收敛于分布函数  $F(x)$ , 即

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

再看两个定理，这是由柯尔莫哥洛夫给出的检验定理。

**定理4.3（精确分布）** 设 $F$ 是连续的分布函数，则

$$P\{D_n = \sup_{-\infty < x < +\infty} |F_n(x) - F(x)| \leq y + \frac{1}{2n}\} \\ = \begin{cases} 0, & y \leq 0, \\ \int_{\frac{1}{2n}-y}^{\frac{1}{2n}+y} \int_{\frac{3}{2n}-y}^{\frac{3}{2n}+y} \cdots \int_{\frac{2n-1}{2n}-y}^{\frac{2n-1}{2n}+y} f(x_1, x_2, \cdots, x_n) dx_1 \cdots dx_n, & 0 < y < \frac{2n-1}{2n}, \\ 1, & y \geq \frac{2n-1}{2n}, \end{cases}$$

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

**定理4.4（极限分布）** 设 $F$ 是连续的分布函数，则

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

$$= \begin{cases} 0, & y \leq 0, \\ \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 y^2}, & y > 0. \end{cases}$$

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 $(X_1, X_2, \dots, X_n)^T$ 为来自总体 $X$ 的样本,  $(x_1, x_2, \dots, x_n)^T$ 为其观测值, 统计量选为  $D_n = \sup_{-\infty < x < +\infty} |F_n(x) - F_0(x)|$

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 $\alpha$ , 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

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

由于参数未知，因而首先对参数进行估计

$$\hat{\mu} = \bar{x} = 1.884, \quad \hat{\sigma}^2 = s_n^2 = 0.576^2$$

则  $H_0$  : 总体  $X$  服从正态分布  $N(1.884, 0.576^2)$ .

$$\begin{aligned} F(x_i) &= P\{X \leq x_i\} = P\left\{\frac{X - 1.884}{0.576} \leq \frac{x_i - 1.884}{0.576}\right\} \\ &= \Phi\left(\frac{x_i - 1.884}{0.576}\right) \end{aligned}$$

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

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

看书  
表4.7

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

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.

### 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 $(X_1, X_2, \dots, X_{n_1})^T$ 为来自总体 $F(x)$ 的样本,  $(Y_1, Y_2, \dots, Y_{n_2})^T$ 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（精确分布）**如果 $F(x)=G(x)$ ,且 $F(x)$ 是连续函数，

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

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

其中  $c = -[-xn]$ .



**定理4.6（极限分布）**如果 $F(x)=G(x)$ ,且 $F$ 是连续函数, 则

$$\lim_{n \rightarrow \infty} P\left\{\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup_{-\infty < x < +\infty} |F_{n_1}(x) - G_{n_2}(x)| \leq x\right\} = K(x)$$
$$= \begin{cases} 0, & x \leq 0, \\ \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 x^2}, & x > 0. \end{cases}$$

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 $\alpha$ ，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 $n = \frac{n_1 n_2}{n_1 + n_2}$ ，临界值 $D_{n, \alpha} \approx \frac{\lambda_{1-\alpha}}{\sqrt{n}}$ 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 / $\mu\text{m}$	频 数		偏差 $X$ 的 测量区间 / $\mu\text{m}$	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	$\Sigma$	$n_1 = 150$	$n_2 = 100$
[5, 10)	23	29			

$$\frac{10}{150} = 0.067$$

表 4.9

$$\frac{37}{150} = 0.247$$

$x/\mu\text{m}$	频 数		累积频数		$F_{n_1} = \frac{n_1(x)}{n_1}$	$G_{n_2}(x) = \frac{n_2(x)}{n_2}$	$ F_{n_1}(x) - G_{n_2}(x) $
	$n_{1j}$	$n_{2j}$	$n_1(x)$	$n_2(x)$			
-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	<u>0.293</u>
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

**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

$$F_{n_1}(x) = \frac{v_{n_1}(x)}{n_1}$$

$$G_{n_2}(x) = \frac{v_{n_2}(x)}{n_2}$$

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



查附表6, 取  $\alpha = 0.05, n = \frac{n_1 n_2}{n_1 + n_2} = \frac{150 \cdot 100}{150 + 100} = 60,$

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

显然

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

因此拒绝原假设, 认为不是同一分布.



# 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 $A_1, A_2, \dots, A_m$ 和 $B_1, B_2, \dots, B_k$ 。设从该总体中抽取一个容量为 $n$ 的样本 $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ , 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( $i = 1, 2, \dots, m, j = 1, 2, \dots, k$ ); 又记

$$n_{i\cdot} = \sum_{j=1}^k n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^m n_{ij}$$

上述问题可以用一个表格—**列联表**来表示如下

# 列联表

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

检验的问题为:

$H_0$ : 总体的两个指标  $X$  和  $Y$  是相互独立的.

统计量选择为

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

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

拒绝域为

$$W = \{x : \chi^2 = n \sum_{i=1}^m \sum_{j=1}^k \frac{(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n})^2}{n_{i\cdot} n_{\cdot j}} \geq \chi_{\alpha}^2((m-1)(k-1))\}$$

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

$X \backslash Y$	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 ( $\alpha=0.01$ )?

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

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

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

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

因而拒绝原假设，即认为慢性气管炎与吸烟有关

**Thank You!**

