3.5 拟合优度和独立性检验

在前面的课程中，我们已经了解了假设检验的基本思想，并讨论了当总体分布为正态时，关于未知参数的假设检验问题．

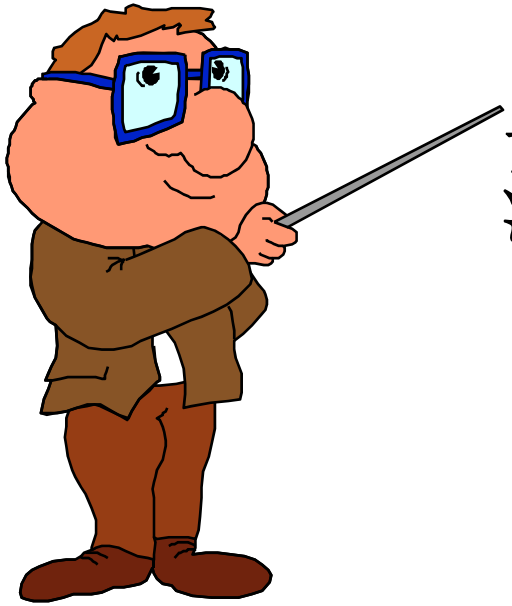
然而可能遇到这样的情形，总体服从何种理论分布并不知道，要求我们直接对总体分布提出一个假设．

例如，从1500到1931年的432年间，每年爆发战争的次数可以看作一个随机变量，据统计，这432年间共爆发了299次战争，具体数据如下：

战争次数X	发生 X次战争的年数
0	223
1	142
2	48
3	15
4	4

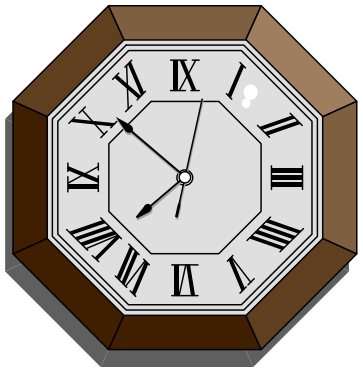
在概率论中，大家对泊松分布产生的一般条件已有所了解，容易想到，每年爆发战争的次数，可以用一个泊松随机变量来近似描述．也就是说，我们可以假设每年爆发战争次数分布X近似泊松分布.

现在的问题是：



上面的数据能否证实X 具有泊松分布的假设是正确的？

又如，某钟表厂对生产的钟进行精确性检查，抽取**100**个钟作试验，拨准后隔**24**小时以后进行检查，将每个钟的误差（快或慢）按秒记录下来.



问该厂生产的钟的误差是否服从正态分布？

再如，某工厂制造一批骰子，声称它是均匀的.

也就是说，在投掷中，出现1点，2点，…，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，统计各点出现的频率与1/6的差距.

问题是：得到的数据能否说明“骰子均匀”的假设是可信的？

解决这类问题的工具是英国统计学家***K.***皮尔逊在**1900**年发表的一篇文章中引进的所谓χ^2检验法**.**

这是一项很重要的工作，不少人把它视为近代统计学的开端**.**

K.皮尔逊

χ^2检验法是在总体X的分布未知时，根据来自总体的样本，检验关于总体分布的假设的一种检验方法.

使用 χ^2检验法对总体分布进行检验时，我们先提出原假设：

H_0：总体X的分布函数为$F(x)$

然后根据样本的经验分布和所假设的理论分布之间的吻合程度来决定是否接受原假设.

这种检验通常称作拟合优度检验，它是一种非参数检验.

在用χ^2检验法检验假设H_0时，若在H_0下分布类型已知，但其参数未知，这时需要先用极大似然估计法估计参数，然后作检验.

分布拟合的χ^2检验法的基本原理和步骤如下：

1. 将总体**X**的取值范围分成k个互不重迭的小区间,记作$A_1, A_2, \ldots, A_k$.

2.把落入第i个小区间A_i的样本值的个数记作f_i， 称为实测频数. 所有实测频数之和$f_1+ f_2+ \ldots+ f_k$等于样本容量n.

3.根据所假设的理论分布,可以算出总体X的值落入每个A_i的概率p_i,于是np_i就是落入A_i的样本值的理论频数.

实测频数　　理论频数

$$f_i - np_i$$

标志着经验分布与理论分布之间的差异的大小.

皮尔逊引进如下统计量表示经验分布与理论分布之间的差异:

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

在理论分布已知的条件下, np_i是常量

统计量 χ^2 的分布是什么?

皮尔逊证明了如下定理：

若原假设中的理论分布$F(x)$已经完全给定，那么当$n \to \infty$时，统计量

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

的分布渐近$(k-1)$个自由度的χ^2分布.

如果理论分布$F(x)$中有r个未知参数需用相应的估计量来代替，那么当$n \to \infty$时，统计量χ^2的分布渐近$(k-r-1)$个自由度的χ^2分布.

根据这个定理，对给定的显著性水平α，查χ^2分布表可得临界值χ_α^2，使得

$$P(\chi^2 > \chi_\alpha^2) = \alpha$$

得拒绝域：$\chi^2 > \chi_\alpha^2(k-1)$ (不需估计参数)

$$\chi^2 > \chi_\alpha^2(k-r-1)$$ (估计r个参数)

如果根据所给的样本值$X_1, X_2, \ldots, X_n$算得统计量χ^2的实测值落入拒绝域，则拒绝原假设，否则就认为差异不显著而接受原假设.

皮尔逊定理是在n无限增大时推导出来的，因而在使用时要注意n要足够大，以及np_i不太小这两个条件.

根据计算实践，要求n不小于50，以及np_i都不小于 5. 否则应适当合并区间，使np_i满足这个要求 .

拟合优度检验

1.多项总体拟合优度检验
2.泊松分布拟合优度检验
3.正态分布拟合优度检验

多项总体

多项总体：总体中的每一个个体被分配到几个类别中的一个且被分配到一个类别中的情况。

例：在过去的一年中，A公司的市场份额稳定在30%，B公司稳定在50%,C公司稳定在20%，最近C公司开发了一种“新型改进”产品，以取代当前市场上该公司所售产品。启典市场调查公司受雇于C公司，目的是判断新产品是否使市场份额发生了改变。

需要检验的是一个多项总体：每一个顾客按照他所购买A公司、B公司还是C公司的产品来进行分类。

拟合优度检验（比例检验）
(goodness of fit test)

1. 检验多个比例是否相等
2. 检验的步骤
   - 提出假设
     - H_0：$\pi_1 = \pi_2 = \ldots = \pi_j$；
       H_1：$\pi_1, \pi_2, \ldots, \pi_j$
       不全相等
   - 计算检验的统计量

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

   - 进行决策
     - 根据显著性水平α和自由度$(k-1)$查出临界值χ_α^2
     - 若$\chi^2 > \chi_\alpha^2$，拒绝H_0；若$\chi^2 < \chi_\alpha^2$，接受H_0

拟合优度检验(比例检验）

- 【例】为了提高市场占有率，A公司和B公司同时开展了广告宣传。在广告宣传战之前，A公司的市场占有率为45%，B公司的市场占有率为40%，其他公司的市场占有率为15%。为了了解广告战之后A、B和其他公司的市场占有率是否发生变化，随机抽取了200名消费者，其中102人表示准备购买A公司产品，82人表示准备购买B公司产品，另外16人表示准备购买其他公司产品。检验广告战前后各公司的市场占有率是否发生了变化 (α =0.05)

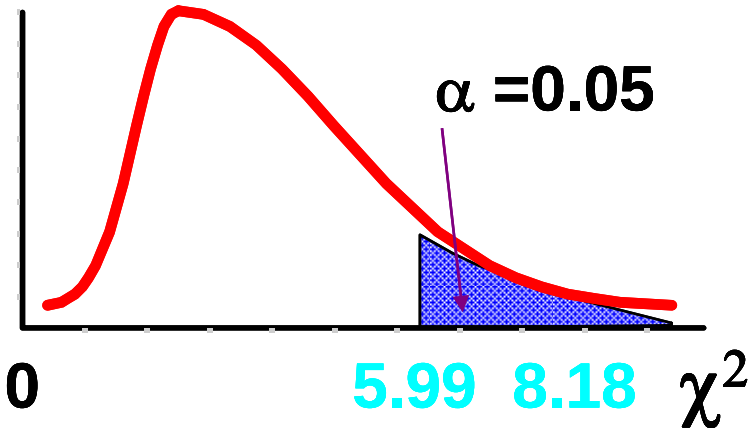
χ^2检验统计量的计算过程

类别	假设比例	观察频数 f	期望频数np	差	差的平方
A公司	0.45	102	90	12	144
B公司	0.40	82	80	2	4
其他公司	0.15	16	30	-14	196
合计	1	200	200	\	\

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} = 8.18$$

拟合优度检验

- **H0: π_1=0.45 π_2=0.4 π_3= 0.15**
- **H1:**原假设中至少有一个不成立
- **α = 0.05**
- **df =(3-1)= 2**
- 临界值**(s):** $\chi^2_{0.05}(2) = 5.99$



统计量:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} = 8.18$$

决策:

在 α = 0.05的水平上拒绝H_0

结论:

可以认为广告后各公司产品市场占有率发生显著变化

拟合优度检验---泊松分布

$$f(x) = e^{-u}u^{x} \Big/ x!$$ **x=0,1,2,3,4**

让我们回到开始的一个例子，检验每年爆发战争次数分布是否服从泊松分布.

提出假设H_0: X服从参数为λ的泊松分布

根据观察结果，得参数λ的极大似然估计为

$$u = \overline{X} = 0.69$$

按参数为0.69的泊松分布，计算事件$X=i$的概率p_i，p_i的估计是

$$f(x) = e^{-0.69}\,0.69^{x} \Big/ x!\quad i=0,1,2,3,4$$

将有关计算结果列表如下:

战争次数 x	0	1	2	3	4	$\sum$
实测频数 f_i	223	142	48	15	4	
$\hat{p}_i$	0.50	0.35	0.12	0.03	0.005	
$n\hat{p}_i$	216.7	149.5	51.6	12.0	2.16	
					14.16	
$\dfrac{(f_i - np_i)^2}{np_i}$	0.183	0.376	0.251	1.623		2.43

将$n\hat{p}_i$<5的组予以合并，即将发生3次及4次战争的组归并为一组.

因H_0所假设的理论分布中有一个未知参数，故自由度为4-1-1=2.

按α=0.05，自由度为4-1-1=2查χ^2分布表得

$$\chi^2_{0.05}(2)=5.991$$

由于统计量χ^2 的实测值

$$\chi^2=2.43<5.991,$$

未落入否定域.

故认为每年发生战争的次数X服从参数为0.69的泊松分布.

例2：p299表12-6资料

检验5分钟时间段内进入该超市的顾客数是否服从泊松分布，以便合理进行员工规划。

表12-6 由126个5分钟时间段超市顾客到达的观察频数

顾客到达数	观察频数	顾客到达数	观察频数	顾客到达数	观察频数
0	2	4	18	8	12
1	8	5	22	9	6
2	10	6	22	合计	128
3	12	7	16		

$u = \overline{X}$ =5

泊松分布拟合优度检验总结

1.建立零假设和备择假设

H0:总体服从泊松概率分布

H1:总体不服从泊松概率分布

2.抽取一个随机样本，并且

a.对于泊松随机变量的每个值记录观察频数*fi*

b.计算发生次数的平均值*U*

3. 计算发生次数的期望频数，即样本容量与泊松随机变量取每个值的概率的乘积。如果期望频数小于5，则将相邻的数值合并，同时减少类别个数。

4.计算检验统计量

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

5.拒绝法则：*p*值法：如果$p \le \alpha$，则拒绝H0

临界值法：若$\chi 2 \ge \chi\alpha$，则拒绝H0

式中，α为显著性水平，自由度为$k-2$.

拟合优度检验---正态分布

连续分布情形下

例：随机抽取50名求职者的能力测验分数

71	66	61	65	54	93	60	86	70	73
73	55	63	56	62	76	54	82	79	76
68	53	58	85	80	56	61	61	64	65
62	90	69	76	79	77	54	64	74	65
70	65	61	56	63	80	56	71	79	84

$\bar{X}$ **=68.42**

S=10.41 (STDEV)

H0:测验分数总体服从均值为68.42和标准差为10.41的正态分布

H1:测验分数总体不服从均值为68.42和标准差为10.41的正态分布

对连续概率分布，可利用测验分数的区间来定义类别，且要满足每个区间中期望频数至少为5的法则

因样本频数为50，以每10%的概率为划分的依据，进行区间估计

正态分布的10各等概率区间

最低10%：68.42-1.28×10.41=55.10 NORMSINV(0.1)=-1.28	最高40%：68.42+0.25×10.41=71.02
最低20%：68.42-0.84×10.41=59.68 NORMSINV(0.2)=-0.84	最高30%：68.42+0.52×10.41=73.83
最低30%：68.42-0.52×10.41=63.01 NORMSINV(0.3)=-0.52	最高20%：68.42+0.84×10.41=77.16
最低40%：68.42-0.25×10.41=65.82 NORMSINV(0.4)=-0.25	最高10%：68.42+1.28×10.41=81.74
中间分数：68.42-0.00×10.41=68.42	

NORMSINV(0.1)=-1.28 ➔P(X<-1.28)=0.1

χ2统计量的计算过程

测验分数	观察频数	期望频数	差的平方	差的平方/期望频数
55.10以下	5	5	0	0.0
55.10-59.68	5	5	0	0.0
59.68-63.01	9	5	16	3.2
63.01-65.82	6	5	1	0.2
65.82-68.42	2	5	9	1.8
68.42-71.02	5	5	0	0.0
71.02-73.83	2	5	9	1.8
73.83-77.16	5	5	0	0.0
77.16-81.74	5	5	0	0.0
81.74以上	6	5	1	0.2
合计	50	50	/	χ^2 =7.2

按α=0.10，自由度为10-2-1=7查χ^2分布表得

$$\chi^2_{0.10}(7)=18.475$$

CHIINV(0.10,7)=12.017

由于统计量χ^2的实测值

χ^2=7.2<12.017，

未落入否定域.

P值计算：

CHIDIST(7.2，7)=0.4084>**0.05**

故认为不能拒绝测验分数服从正态分布的假设

正态分布拟合优度检验总结

1.建立零假设和备择假设

H0:总体服从正态分布

H1:总体不服从正态分布

2.抽取一个随机样本，并且

a.计算样本均值和样本标准差

b.确定取之区间使得每个区间中的期望频数至少为5。

c.对于每个确定好的区间记录观察频数的数据值

3.对于步骤2（b)中确定的每个区间，计算发生次数的期望频数，即样本容量与正态随机变量落入每个区间的概率的乘积

4.计算检验统计量

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

5.拒绝法则：*p*值法：如果$p \leq a$，则拒绝H0

临界值法：若$\chi 2 \geq \chi\alpha$，则拒绝H0

式中，*a*为显著性水平，自由度为*k-3*.

独立性检验
(test of independence)

独立性检验

【例】某啤酒厂生产三种类型的啤酒：淡啤酒、普通啤酒和黑啤酒。在一次对三种啤酒市场份额的分析中，公司市场研究小组提出了一个问题：男性与女性饮酒者对于三种啤酒的偏好是否存在差异。随机抽取150名饮酒者进行检验，结果如下表。检验性别与啤酒偏好之间是否存在依赖关系（α =0.05)

		啤酒偏好		
性别	淡啤酒	普通啤酒	黑啤酒	合计
男性	**20**	**40**	**20**	**80**
女性	**30**	**30**	**10**	**70**
合计	**50**	**70**	**30**	**150**

独立性检验

【例】一种原料来自三个不同的地区，原料质量被分成三个不同等级。从这批原料中随机抽取500件进行检验，结果如下表。检验各地区与原料之间是否存在依赖关系（$\alpha=0.05$）

地区	一级	二级	三级	合计
甲地区	52	64	24	140
乙地区	60	59	52	171
丙地区	50	65	74	189
合计	162	188	150	500

列联表
(contingency table)

1. 由两个以上的变量交叉分类的频数分布表
2. 行变量的类别用 r 表示， r_i 表示第 i 个类别
3. 列变量的类别用 c 表示， c_j 表示第 j 个类别
4. 每种组合的观察频数用 f_{ij} 表示
5. 表中列出了行变量和列变量的所有可能的组合
6. 一个 r 行 c 列的列联表称为 $r \times c$ 列联表

列联表的结构
(2 × 2 列联表)

行 (r_i) \ 列(c_j)	列(c_j)		合计
	j =1	j =2	
i =1	f_{11}	f_{12}	$f_{11}+f_{12}$
i =2	f_{21}	f_{22}	$f_{21}+f_{22}$
合计	$f_{11}+f_{21}$	$f_{12}+f_{22}$	n

列联表的结构

($r \times c$ 列联表的一般表示)

行(r_i) \ 列(c_j)	列(c_j)			合计
	$j=1$	$j=2$	…	
$i=1$	f_{11}	f_{12}	…	r_1
$i=2$	f_{21}	f_{22}	…	r_2
:	:	:	:	:
合计	c_1	c_2	…	n

f_{ij} 表示第 i 行第 j 列的观察频数

列联表的分布

观察值的分布

1. 边缘分布
   - 行边缘分布
     - 行分布观察值的的合计数
     - 例如，男性共有**80**人，女性共有**70**人
   - 列边缘分布
     - 列观察值的合计数的分布
     - 例如，偏好淡啤酒的**50**人，偏好普通啤酒的**70**人，偏好黑啤酒的**30**人
2. 条件分布与条件频数
   - 变量 ***X*** 条件下变量 ***Y*** 的分布，或在变量 ***Y*** 条件下变量 ***X*** 的分布
   - 每个具体的观察值称为条件频数

观察值的分布
(图示)

条件频数:每个具体的观察值称为条件频数

行边缘分布:行分布观察值的的合计数

		啤酒偏好		
性别	淡啤酒	普通啤酒	黑啤酒	合计
男性	20	40	20	80
女性	30	30	10	70
合计	50	70	30	150

列边缘分布:列观察值的合计数的分布

期望频数的分布
(例题分析)

在全部150个调查人中，偏好淡啤酒的有50人，占到总数的33.3%；偏好普通啤酒的占到46.7%；偏好黑啤酒的占到20%。

如果性别与啤酒偏好独立，

那么男性饮酒者中也应该有**33.3%**偏好淡啤酒；**46.7%**偏好普通啤酒；**20%**偏好黑啤酒。那么男性中偏好淡啤酒的应该有：

80*33.3%=26.67 人

期望频数的分布
(例题分析)

		淡啤酒	普通啤酒	黑啤酒
男性	实际频数	20	40	20
	期望频数	26.67	37.33	60
女性	实际频数	30	30	10
	期望频数	23.33	32.67	14

χ^2 统计量

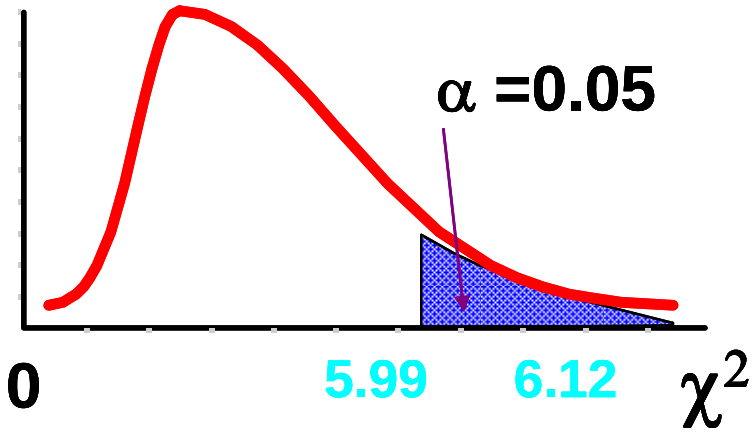
性别	实际频数 (f_{ij})	期望频数 (e_{ij})	$f_{ij} - e_{ij}$	$(f_{ij} - e_{ij})^2$	$\frac{(f_{ij} - e_{ij})^2}{e}$
男性	20	26.67	-6.67	44.44	1.67
男性	40	37.33	2.67	7.11	0.19
男性	20	16	4	16	1
女性	30	23.33	6.67	44.44	1.9
女性	30	32.67	-2.67	7.11	0.22
女性	10	14	-4	16	1.14
合计	150	150	/	/	χ^2 =6.12

临界值计算：CHIINV(0.05，2)=5.99　　χ2=6.12>5.99

P值计算：CHIDIST(6.12，2)=0.047<**0.05**

拟合优度检验
(例题分析)

- **H0:**性别与啤酒偏好之间独立
- **H1:**性别与啤酒偏好之间不独立
- **α = 0.05**
- **df = (2-1)(3-1)= 2**
- 临界值(s):



统计量:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} = 6.12$$

决策:

在 α = 0.05的水平上拒绝H_0

结论:

性别和啤酒偏好之间存在依赖关系

独立性检验

1. 检验列联表中的行变量与列变量之间是否独立
2. 检验的步骤为
   - 提出假设
     - H_0：行变量与列变量独立
     - H_1：行变量与列变量不独立
   - 计算检验的统计量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$$

   - 进行决策
     - 根据显著性水平α和自由度$(r-1)(c-1)$查出临界值χ_α^2
     - 若$\chi^2 \geq \chi_\alpha^2$，拒绝$H_0$；若$\chi^2 < \chi_\alpha^2$，接受$H_0$

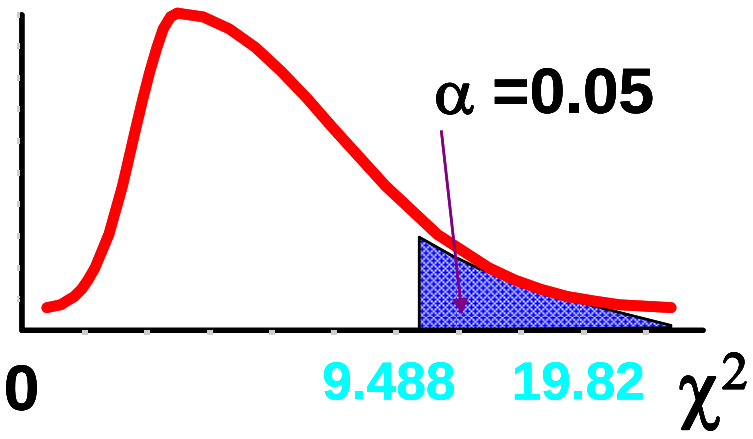
独立性检验
(例题分析)

1. 提出假设
   - H_0：地区与原料等级之间独立
   - H_1：地区与原料等级之间不独立
2. 计算检验的统计量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} = 19.82$$

拟合优度检验
(例题分析)

- **H0:**地区与原料等级之间独立
- **H1:**地区与原料等级之间不独立
- **α = 0.05**
- **df = (3-1)(3-1)= 4**
- 临界值**(s):**



统计量:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} = 19.82$$

决策:

在 α = 0.05的水平上拒绝H_0

结论:

地区和原料等级之间存在依赖关系

独立性检验总结

1.建立零假设和备择假设

H0:列变量与行变量独立

H1:列变量与行变量独立

2.抽取一个随机样本，记录列联表中每个单元格的观察频数

3.计算每个单元格的期望频数

4.计算检验统计量

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$$

其自由度为$(r-1)(c-1)$

5.拒绝法则：*p*值法：如果$p \le \alpha$，则拒绝H0

临界值法：若$\chi 2 \ge \chi\alpha$，则拒绝H0