第六讲　总体分布的假设检验

6.1 拟合优度的χ^2检验

The Chi-Square Goodness -of- Fit Test

检验目的：总体被分为*m*类;

检验观测频次与期望频次是否吻合？

H_0: 总体在第 1, 2, …, *m*类中的比率分别是 $p_1, p_2, \cdots, p_m$.

$$p_1 + p_2 + \cdots + p_m = 1$$

H_1:上述比率中至少有一个是不正确的.

例:

电视台有六种儿童节目。随机抽取300名常看电视的儿童，询问“最喜欢哪一个节目”（每人只能选一种节目）。

H_0：儿童对电视台提供的六种节目无偏好

$$p_1 = p_2 = \cdots = p_m = 1/6$$

第 i 类	o_i	p_i	e_i	$(o_i - e_i)$	$(o_i - e_i)^2/e_i$
1	85	1/6	50	35	24.5
2	80	1/6	50	30	18.0
3	55	1/6	50	5	0.5
4	10	1/6	50	-40	32.0
5	40	1/6	50	-10	2.0
6	30	1/6	50	-20	8.0
Total	300	1.0	300	0.0	$\chi^2 = 85.0$

o_i: 观测频次; p_i : 期望比率

e_i: 期望频次

拟合优度的χ^2检验

H_0: 在总体中，在第 1,2,…, m 类中的比率分别是

$p_1, p_2, \ldots, p_m$.

H_1:上述比率中至少有一个是不正确的。

$$\chi^2 = \sum_{i=1}^{m} \frac{(o_i - e_i)^2}{e_i} \overset{H_0}{\sim} \chi^2(m-1)$$

取 α=0.05:

如果H_0 为真: $P\{\chi^2(5) > \lambda\} = 0.05 \Rightarrow \lambda = 11.07$

由于 χ^2=85.0> 11.07

因此，拒绝 H_o.

$$\chi^2 = \sum_{i=1}^{m} \frac{n}{p_i}(\hat{p}_i - p_i)^2 \overset{H_0}{\sim} \chi^2(m-1)$$

其中， $\hat{p}_i = \frac{o_i}{n}$， $p_i = \frac{e_i}{n}$

例：瑞典一年分为四季，观察在全年出生的的新生儿的一个容量为88 的样本。

H_0:在瑞典，一年四季中新生儿出生是同等频繁的。

季节	x_i	p_i	np_i	x_i-np_i	$\frac{(x_i - np_i)^2}{np_i}$
(91)春4~6月	27		22.0	5	1.14
(62)夏7~8月	20		15.0	5	1.67
(61)秋9~10月	8		14.7	-6.7	3.05
(151)冬11~3月	33		36.3	-3.3	0.30
合计	88	1.0	88	0	$\chi^2 = 6.16$

自由度 df = (K-1) = （4-1）= 3

取 α =0.05

$$P(\chi^2(3) > \lambda) = 0.05$$

查表： $\lambda = 7.815$

$$\chi^2(3) = 6.16 < 7.815$$

不能拒绝H_0。

即在瑞典，一年中新生儿的出生分布还是相当均匀的。

总体分布的假设检验

$H_0: X \sim F(x)$　　$H_1: F(x)$不是 X 的分布函数

类似直方图绘制方法，对样本观测值 $X_1, X_2, \cdots, X_n$ 所在区间进行划分，得到互不相交的区间： $I_1, I_2, \cdots, I_m$

$$\hat{p}_i = \frac{\#\left\{k \middle| X_k \in I_j\right\}}{n}$$

$$p_i = P\left(X \in I_j\right)$$

区间划分	频率 $\hat{p}_j$	概率 p_j
I_1	$\hat{p}_1$	p_1
I_2	$\hat{p}_2$	p_2
...	...	...
I_m	$\hat{p}_m$	p_m

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i}(\hat{p}_i - p_i)^2 \sim \chi^2(m-1)$$

注意：

（1）样本容量的大小与区间划分要满足条件：

$$\forall 0 \le j \le 1: \quad np_j \ge 5$$

否则，将 I_j 与临近区间合并

（2）如果总体分布F(*x*)中**有 *r* 个未知参数**，就需要用观测数据先计算出未知参数的估计值。

这时，在H_0假设的条件下，当*n* 比较大时，有：

$$\chi^2 = \sum_{i=1}^{m} \frac{n}{p_i} (\hat{p}_i - p_i)^2 \overset{H_0}{\sim} \chi^2(m - r - 1)$$

例题1.4（P217） 某图书馆在一年中，通过随机抽样调查了60天的读者借书数（数据见P126的例3.1），能否认为这批数据是正态总体的样本？

$$\hat{\mu}=\bar{X}_n=403.5 \qquad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_j-\hat{\mu}\right)^2=83.12^2$$

$H_0:\ X\sim N\left(403.5,83.12^2\right)$

借出书数 I_j	频率	概率
（200,250）	0.0500	0.0252
(250,300)	0.0333	0.0741
(300,350)	0.2000	0.1534
(350,400)	0.2333	0.2233
(400,450)	0.2000	0.2280
(450,500)	0.1833	0.1651
(500,550)	0.0500	0.0838
(550,600)	0.0500	0.0300

$\because\ np_1=60\times0.0252=1.512<5\ \therefore\ I_1$与$I_2$合并

$np_8=60\times0.0300=1.8<5$ I_7与I_8合并

借出书数Ij	频率	概率
（200,300)	0.0833	0.0993
(300,350)	0.2000	0.1534
(350,400)	0.2333	0.2233
(400,450)	0.2000	0.2280
(450,500)	0.1833	0.1651
(500,600)	0.1000	0.1138

扣除2个自由度：

查表：$\chi^2_{0.05}(6-2-1)=7.815$

$$\chi^2=\sum_{i=1}^{k}\frac{n}{p_i}(\hat{p}_i-p_i)^2=0.1471<7.815$$

所以，不能拒绝总体来自正态分布 $N\left(403.5,83.12^2\right)$

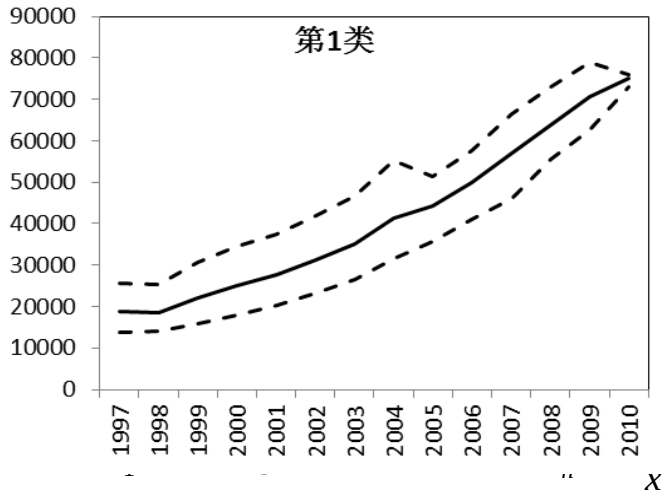
直观判断方法： Q-Q 图

$$X \leftarrow X_1 \le X_2 \le \cdots \le X_n \quad i.i.d$$

经验分布：

$$\hat{F}_n(x) = \frac{\#\left\{ j \middle| X_j \le x \right\}}{n} = \frac{1}{n}\sum_{j=1}^{n} I\left[X_j \le x \right]$$

$$\hat{F}_n(x) = \begin{cases} 0 & x < X_1 \\ \dfrac{j}{n} & x \in [X_j, X_{j+1}),\ j = 1, 2, \cdots, n-1 \\ 1 & x \le X_n \end{cases}$$



可以证明：

$$\lim_{n\to\infty} \hat{F}_n(x) = F(x) \quad a.s$$

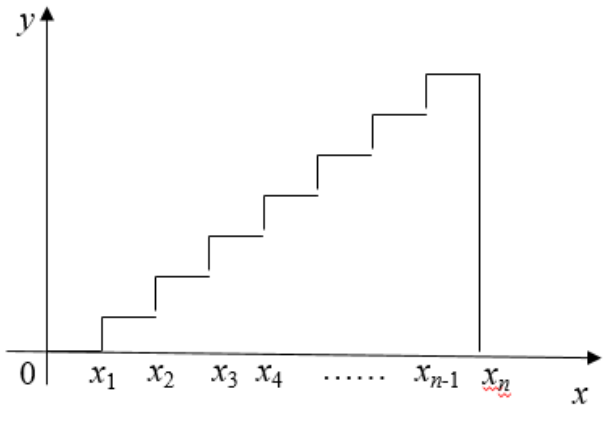
$$\lim_{n\to\infty} \sup_x \left| \hat{F}_n(x) - F(x) \right| = 0 \quad a.s$$

经验分布与理论分布

两条曲线趋于重合

经验分布的反函数：

$$\hat{F}_n^{-1}\left(\frac{j}{n}\right)=X_j\ ,\ j=1,2,\cdots,n\text{-}1$$



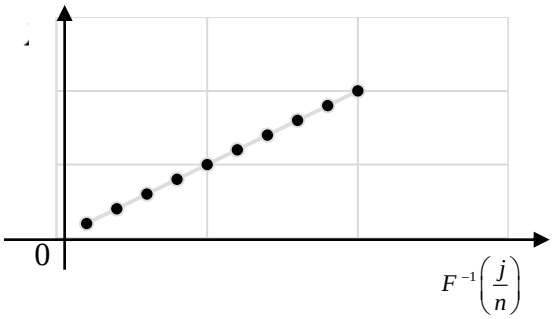
对比：　理论分布的反函数值——经验分布的反函数值

$$F^{-1}\left(\frac{j}{n}\right)$$

$$\hat{F}_n^{-1}\left(\frac{j}{n}\right)=X_j$$

绘制散点图：　Q-Q 图　(quantile-quantile plot)

$$\left(F^{-1}\left(\frac{j}{n}\right),X_j\right)\ ,\ j=1,2,\cdots,n-1$$



$H_0:\ X\sim F(x)$

当图中的点都在对角线附近

6.2 列联表独立性检验

Test of Independence of Contingency Tables

列联表（ Contingency table ）

两个定性变量的相关关系

例：对电视节目的选择与工资收入是否相关？

Income	Type of TV Show			Total
	Hockey	Movie	News	
Low	143	70	37	250
Medium	90	67	43	200
High	17	13	20	50
Total	250	150	100	500

Income	Type of TV Show			Total
	Hockey	Movie	News	
Low				
o_{ij}	143	70	37	250
p_{ij}	0.5	0.3	0.2	
e_{ij}	125	75	50	
Medium				
o_{ij}	90	67	43	200
p_{ij}	0.5	0.3	0.2	
e_{ij}	100	60	40	
High				
o_{ij}	17	13	20	50
p_{ij}	0.5	0.3	0.2	
e_{ij}	25	15	10	
Total	250/500 =0.5	150/500 =0.3	100/500 =0.2	500

250×100/500

$$e_{11} = \frac{r_1 c_1}{n} = \frac{250 \times 250}{500} = 125$$

$$e_{23} = \frac{r_2 c_3}{n} = \frac{200 \times 100}{500} = 40$$

$$e_{ij} = \frac{r_i c_j}{n}$$

Chi-square 检验统计量为：

$$\chi^2 = \sum_{i=1}^{H} \sum_{j=1}^{K} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$\chi^2 = \frac{(143-125)^2}{125} + \frac{(70-75)^2}{75} + \cdots + \frac{(20-10)^2}{10} = 21.74$$

H_0: 对电视节目的选择与工资收入无关.

H_1:对电视节目的选择与工资收入相关.

$$\chi^2 = \sum_{i=1}^{H}\sum_{j=1}^{K}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \overset{H_0}{\sim} \chi^2\left((H-1)(K-1)\right)$$

取α=0.05, df = (H-1)(K-1)=(3-1)(3-1)=4

查表： $\chi^2(4)=9.49$

观察的χ^2值为： χ^2= 21.174>9.49

因此，拒绝 H_0 .

收入与电视选择具有相关性.

通过条件概率观察相关关系

原始数据表

Income	Type of TV Show			Total
	Hockey	Movie	News	
Low	143	70	37	250
Medium	90	67	43	200
High	17	13	20	50
Total	250	150	100	500

条件概率表

Income	Type of TV Show			Total
	Hockey	Movie	News	
Low	0.572	0.280	0.148	250
Medium	0.450	0.335	0.215	200
High	0.340	0.260	0.400	50
Total	250	150	100	500

$\frac{90}{200}$

例：在电视收视率调查中，得到性别与收视习惯的列联表如下。试分析性别与收视习惯的相互关系。

习惯 \ 性别	男	女	$x_{i\cdot}$
几乎天天看	38	24	62
偶　尔　看	31	7	38
$x_{\cdot j}$	69	31	100

H_0: 性别与收视习惯无关系。

H_1: 性别与收视习惯有关系。

习惯 \ 性别	男	女	$x_i.$
几乎天天看	a	b	$a+b$
偶 尔 看	c	d	$c+d$
$x_j.$	$a+c$	$b+d$	n

$$\chi^2 = \frac{n(ad-cb)^2}{(a+c)(b+d)(a+b)(c+d)}$$

例：在电视收视率调查中，得到性别与收视习惯的列联表如下。试分析性别与收视习惯的相互关系。

习惯 \ 性别	男	女	$x_{i\cdot}$
几乎天天看	0.55 38	24 0.77	62
偶 尔 看	0.45 31	7 0.23	38
$x_{j\cdot}$	69	31	100

$$\chi^2 = \frac{100(38\times 7 - 24\times 31)^2}{69\times 31\times 62\times 38} = 4.54$$

$$df = (2-1)(2-1) = 1 \quad \Rightarrow \chi^2_{0.05}(1) = 3.84$$

$$\chi^2 = 4.54 > 3.84, \quad \text{拒绝} H_0 \text{。}$$

例：1969年美国恢复抽签方法来决定谁服兵役，用随机机制决定年轻人去越南战场的可能性。（用征兵号码对应生日。1969年9月14日为1号,…）

出生的月份 / 征兵号码	1~6 月	7~12 月	行和
1~183	73	110	183
184~366	109	74	183
列和	182	184	366

例：奶酪种类——地区；地区——竞选者；

报纸杂志种类—社会阶层；婆媳关系—住房条件；

市民情绪分布：年龄—情绪特征

阅读与练习

拟合优度检验SPSS 软件应用

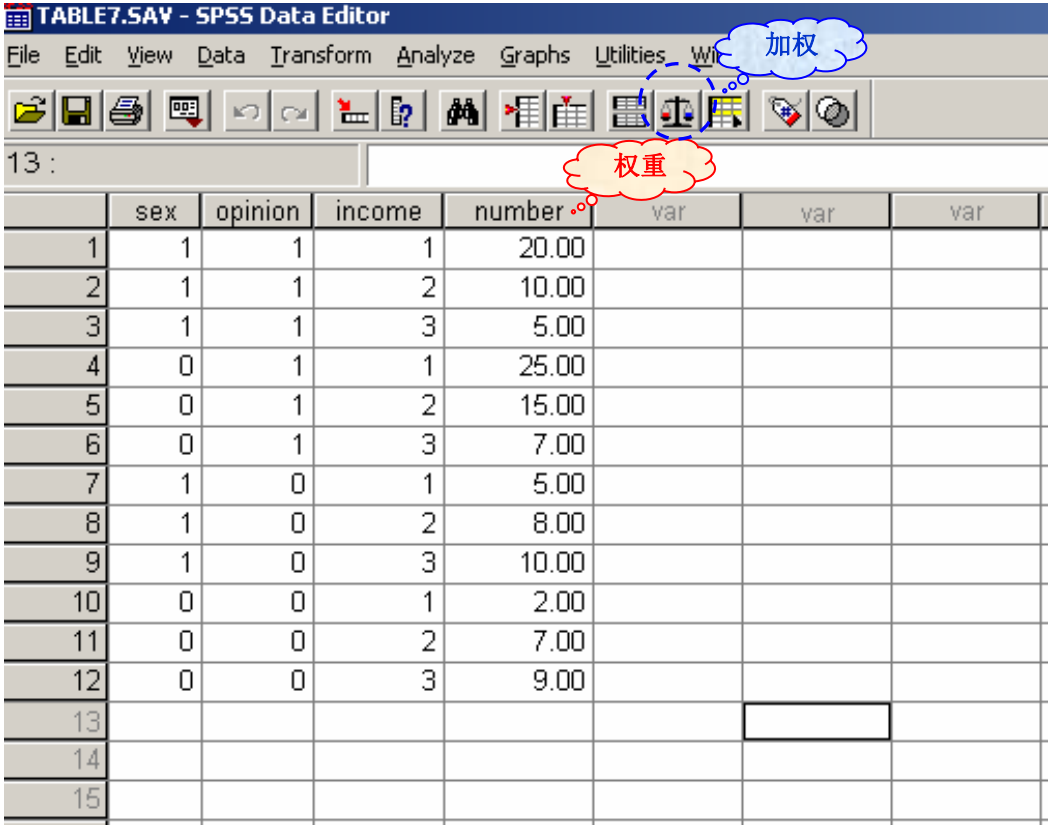
列联表检验SPSS 软件应用

拟合优度检验SPSS软件应用：数据table7.sav

共有 3 个变量：性别（sex)、观点（opinion）、收入（income）

每种组合的权重（weight）（即列联表中的频数）在number那一列

	观点：反对			观点：赞成		
性别	低收入	中收入	高收入	低收入	中收入	高收入
男	5	8	10	20	10	5
女	2	7	9	25	15	7



	sex	opinion	income	number	var	var	var
1	1	1	1	20.00			
2	1	1	2	10.00			
3	1	1	3	5.00			
4	0	1	1	25.00			
5	0	1	2	15.00			
6	0	1	3	7.00			
7	1	0	1	5.00			
8	1	0	2	8.00			
9	1	0	3	10.00			
10	0	0	1	2.00			
11	0	0	2	7.00			
12	0	0	3	9.00			
13							
14							
15							

计算过程：

（1）加权：点击图标中的小天平，出现对话框之后点击**Weight cases**，然后把“**number**”选入即可。

（2）选项：**Analyze－Nonparametric Tests－Chi Square**

◆选择想要检验的变量（比如**income**）

➢如要检验其水平是否相等，则在**Expected Values**选**All categories equal**作为零假设

➢如要检验其水平是否为某比例，则在下面**Values**逐个输入比例（例如**5：4：1**）

➢如果选入的变量多于一个，则检验的是水平相等的零假设（不能分别输入比例）

◆ 点**Exact**时打开的对话框，可以选择精确方法（**Exact**）

◆ 最后**OK**即可。

如要检验其水平是否相等：

Expected values:

All categories equal

Test Statistics

	income	sex	opinion
Chi-Square[a,b]	5.415	.398	13.667
df	2	1	1
Asymp. Sig.	.067	.528	.000
Exact Sig.	.069	.589	.000
Point Probability	.005	.118	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 41.0.

b. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 61.5.

只有“观点”变量是显著的

SPSS还分别给出对每个变量的O_i和E_i

income

	Observed N	Expected N	Residual
1	52	41.0	11.0
2	40	41.0	-1.0
3	31	41.0	-10.0
Total	123		

sex

	Observed N	Expected N	Residual
0	65	61.5	3.5
1	58	61.5	-3.5
Total	123		

opinion

	Observed N	Expected N	Residual
0	41	61.5	-20.5
1	82	61.5	20.5
Total	123		

如果想知道收入的比例是否是 5：4：1（零假设）

- 在 Values中逐个输入比例（例如5：4：1）

给出了的O_i和 E_i（分别为下表中的Observed N和Expected N）：

income

	Observed N	Expected N	Residual
1	52	61.5	-9.5
2	40	49.2	-9.2
3	31	12.3	18.7
Total	123		

$$Q=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}$$

得到各种检验结果如下：

Test Statistics

	income
Chi-Square[a]	31.618
df	2
Asymp. Sig.	.000
Exact Sig.	.000
Point Probability	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 12.3.

列联表检验SPSS软件应用：数据table7.sav

- 使用变量：观点（opinion）和收入（income）
- 每种组合的权重（weight）（即列联表中的频数）在number那一列

计算过程：

（1）加权：点击图标中的小天平，出现对话框之后点击Weight cases，然后把“number”选入即可。

（2）选项：Analyze－Descriptive Statistics－Crosstabs

- 分别把opinion和income分别选入Row（行）和Column（列）
- 在Statistics中选择Chi-square
- 在Exect 中选择Exect
- 点击OK

opinion * income Crosstabulation

Count

		income			Total
		1	2	3	
opinion	0	7	15	19	41
	1	45	25	12	82
Total		52	40	31	123

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	20.456[a]	2	.000	.000		
Likelihood Ratio	21.190	2	.000	.000		
Fisher's Exact Test	20.713			.000		
Linear-by-Linear Association	20.290[b]	1	.000	.000	.000	.000
N of Valid Cases	123					

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.33.

b. The standardized statistic is -4.504.