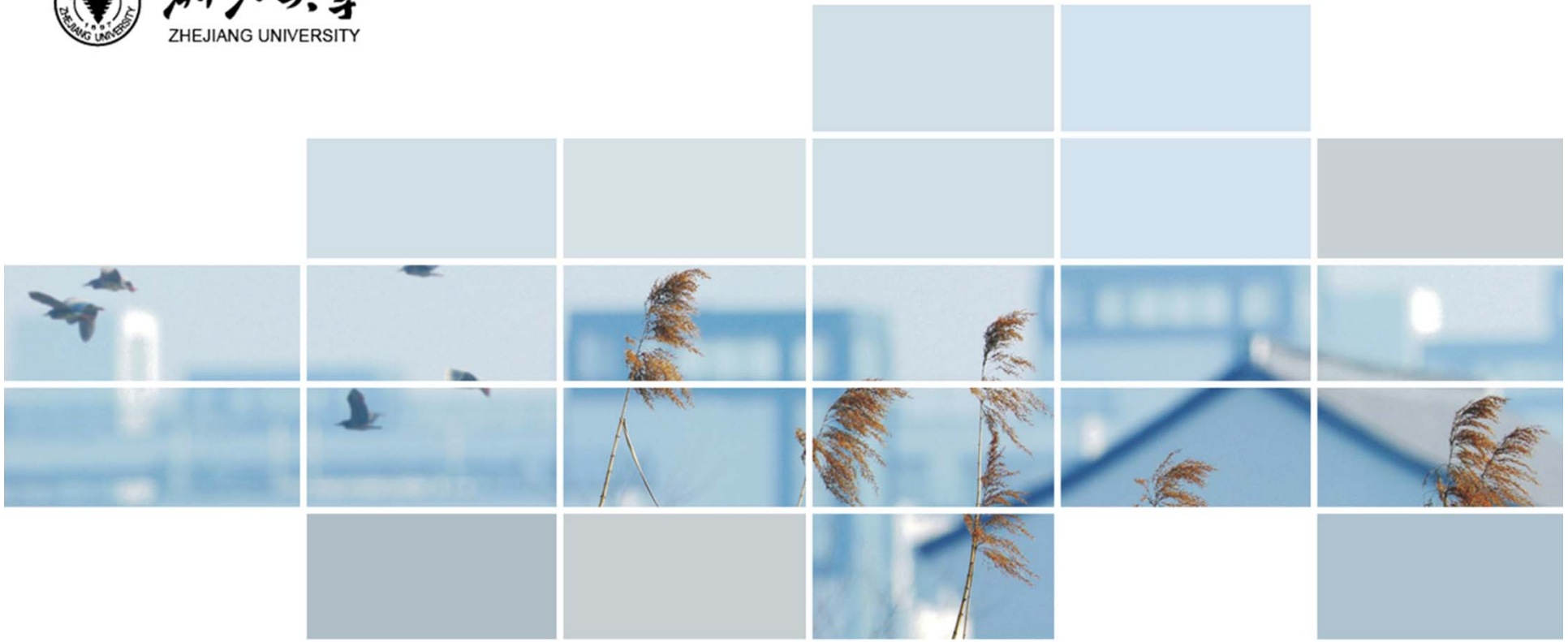
浙江大学
ZHEJIANG UNIVERSITY



第60讲：拟合优度检验

前面介绍的各种检验都是在总体服从正态分布前提下，对参数进行假设检验的。实际中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1: 一淘宝店主搜集了一年中每天的订单数X，除去春节期间及双十一前后外，按330天计，具体数据如下:

订单数X	0	1	2	3	4	5	6	7
天数	3	6	21	46	48	61	52	42

订单数X	8	9	10	11	12	13	16
天数	27	11	6	4	1	1	1

通常认为每天的订单数服从泊松分布，以上的数据是否支持这个结论?

问题：根据上面的数据是否可以得出每天的订单数服从泊松分布假设？

记：$F(x)$为总体X的未知的分布函数，

假设：$F_0(x)$是形式已知，但含有若干个未知参数的分布函数，

检验假设：

$$H_0 : F(x) = F_0(x) \qquad \forall x \in R$$

注：若总体X为离散型，则假设H_0为：

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

若总体X为连续型，则假设H_0为：

H_0：总体X的概率密度为$f(x)$。

检验方法：拟合优度检验

拟合优度检验的基本原理和步骤：

1. 在H_0下，总体X取值的全体分成k个两两不相交的子集$A_1,\ldots,A_k$.

2. 以$n_i(i=1,\ldots,k)$记样本观察值$x_1,\ldots,x_n$中落在A_i的个数（实际频数）.

3.当H_0为真且$F_0(x)$完全已知时，计算事件A_i发生的概率$p_i = P_{F_0}(A_i), i = 1,\ldots,k;$

当$F_0(x)$含有r个未知参数时，先利用极大似然法估计r个未知参数，然后求得p_i的估计$\hat{p}_i$.

此时称np_i(或$n\hat{p}_i$)为**理论频数.**

4.检验统计量$\sum_{i=1}^{k} h_i(n_i - np_i)^2$，　$h_i = ?$

检验的拒绝域形式为：　$\sum_{i=1}^{k} h_i(n_i - np_i)^2 \geq c.$

定理： 若n充分大，则当H_0为真时，统计量

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \overset{\text{近似}}{\sim} \chi^2(k-1)$$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} \overset{\text{近似}}{\sim} \chi^2(k-r-1)$$

其中k为分类数，r为$F_0(x)$中被估未知参数个数.

所以，取检验统计量为 $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$

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

$$= \sum_{i=1}^{k} \frac{n_i^2}{np_i} - 2\sum_{i=1}^{k} n_i + n\sum_{i=1}^{k} p_i$$

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

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

即在显著水平α下拒绝域为

$$\chi^2 = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n \geq \chi_\alpha^2(k-1), \quad \text{（没有参数需要估计）}$$

$$\chi^2 = \sum_{i=1}^{k} \frac{n_i^2}{n\hat{p}_i} - n \geq \chi_\alpha^2(k-r-1), \quad \text{（有}r\text{个参数需要估计）}$$

注：χ^2拟合检验使用时必须注意：

n要足够大，np_i(或$n\hat{p}_i$)不能太小。

根据实践，要求$n \geq 50$，np_i（或$n\hat{p}_i$）≥ 5，

否则应适当合并相邻的类，以满足要求。

例1的Excel实现见实验24.

解：$H_0: X \sim \pi(\lambda)$，λ未知，总订单数为1749，

所以，平均每天订单数$\hat{\lambda} = \overline{X} = 1749/330 = 5.3.$

概率估计（大于11的订单次数较小，所以将大于等于11的合并）：

需注意！

$$\hat{p}_i = \frac{\hat{\lambda}^i e^{-\hat{\lambda}}}{i!}, i = 0,1,...,10, \quad \hat{p}_{11} = \sum_{j=11}^{\infty} \frac{\hat{\lambda}^j e^{-\hat{\lambda}}}{j!} = 1 - \sum_{i=0}^{10} \hat{p}_i.$$

理论频数：$n\hat{p}_i, i = 0,1,...,10,11, n\hat{p}_0 = 1.65 < 5,$

将$x = 0$与$x = 1$合并. 具体结果为

订单数X	0	1	2	3	4	5
天数	3	6	21	46	48	61
概率估计	0.005	0.026	0.070	0.124	0.164	0.174
理论频数	1.65	8.73	23.13	40.87	54.16	57.41

10.23

订单数X	6	7	8	9	10	≥11
天数	52	42	27	11	6	7
概率估计	0.154	0.116	0.077	0.045	0.024	0.021
理论频数	50.71	38.39	25.44	14.98	7.94	6.60

检验统计量的值为

$$\chi^2 = \sum_{i=1}^{k} \frac{n_i^2}{n\hat{p}_i} - n = \sum_{i=1}^{11} \frac{n_i^2}{n\hat{p}_i} - 330 = 3.97$$

即在显著性水平$\alpha = 0.05$下临界值

$$\chi_\alpha^2(k-r-1) = \chi_{0.05}^2(11\text{-}1\text{-}1) = 16.92$$

于是，$3.97 < 16.92$,接受原假设。

例2:孟德尔遗传理论断言，当两个品种的豆杂交时，圆的和黄的、起皱的和黄的、圆的和绿的、起皱的和绿的豆的频数将以比例9：3：3：1发生。在检验这个理论时，孟德尔分别得到频数315、101、108、32、这些数据是否支持该理论？

解：定义$X=\begin{cases}1, & \text{若豆子是圆的和黄的}\\ 2, & \text{若豆子是起皱的和黄的}\\ 3, & \text{若豆子是圆的和绿的}\\ 4, & \text{若豆子是起皱的和绿的}\end{cases}$

$$H_0: p_1=P(X=1)=\frac{9}{16}, p_2=P(X=2)=\frac{3}{16},$$

$$p_3=P(X=3)=\frac{3}{16}, p_4=P(X=4)=\frac{1}{16}.$$

豆子状态 x	1	2	3	4
实测频数 n_i	315	101	108	32
概率 p_i	9/16	3/16	3/16	1/16
理论频数 np_i	312.75	104.25	104.25	34.75

$$\chi^2 = \sum_{i=1}^{4} \frac{n_i^2}{np_i} - n = 0.47 < \chi^2_{0.05}(3) = 7.815,$$

不拒绝原接受，即数据支持该理论.