5.3：拟合优度检验

中国科学技术大学统计与金融系

第五章：假设检验

5.3 拟合优度检验

前面的假设检验基本上是在假定总体是正态的条件下做的, 但是这个假设本身不一定成立, 需要收集样本 $(X_1, \cdots, X_n)$ 来检验它. 一般地, 检验

$$H_0 : X\text{服从某种分布}F$$

可以采用 Karl Pearson 提出的 χ^2 拟合优度检验.

基本想法: 基于样本得到 F 的估计 $\hat{F}_n$, 计算某种偏差 $D(\hat{F}_n, F)$, 例如 $sup_{x\in R}|\hat{F}_n(x) - F(x)|$. 当 H_0 正确时, 由于 $\hat{F}_n$ 是 F 的相合估计, 偏差 $D(\hat{F}_n, F)$ 应该很小.

Karl Pearson 对离散分布 F 提出一种检验方法, 即拟合优度检验方法或者称为 Pearson 卡方检验方法.

5.3.1 离散总体情形

(1) 理论总体分布不含未知参数的情形

设某总体 X 服从一个离散分布,

X	a_1	$\dots$	a_k
P	p_1	$\dots$	p_k

$p_1, \cdots, p_k$ 完全已知. 现从该总体抽得一个样本量为 n 的样本, 其落在类别 $a_1, \cdots, a_k$ 的观测数分别为 $n_1, \cdots, n_k$. 感兴趣的问题是检验理论频率是否正确, 即下面假设是否正确:

$$H_0 : P(X = a_1) = p_1, \cdots, P(X = a_k) = p_k.$$

这类问题只提零假设而不提对立假设, 相应的检验方法称为拟合优度检验. 显然, 在零假设下, 各类别的理论频数分别为 $np_1, \cdots, np_k$, 将理论频数和观测频数列于下表:

类别	a_1	a_2	$\cdots$	a_k
理论频数	np_1	np_2	$\cdots$	np_k
观测频数	n_1	n_2	$\cdots$	n_k

由大数定律知, 在零假设成立时, n_i/n 依概率收敛于 p_i, 故理论频数 np_i 与观测频数 n_i 接近. Pearson 提出检验统计量

$$T = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum \frac{(O-E)^2}{E}.$$

可以严格地证明, 在一定的条件下, 当 H_0 成立时, T 的极限分布就是自由度为 $k-1$ 的 χ^2 分布.

拒绝域: $T > \chi^2_\alpha(k-1)$

下面给出一个例子来说明拟合优度检验的应用.

有人制造一个含 6 个面的骰子, 并声称是均匀的. 现设计一个实验来检验此命题: 连续投掷 600 次, 发现出现六面的频数分别为 97, 104, 82, 110, 93, 114. 问能否在显著性水平 0.2 下认为骰子是均匀的?

↑Example

↓Example

解: 该问题设计的总体是一个有 6 个类别的离散总体, 记出现六个面的概率分别为 $p_1, \cdots, p_6$, 则零假设可以表示为

$$H_0 : p_i = 1/6, i = 1, \cdots, 6.$$

在零假设下, 理论频数都是 100, 故检验统计量 χ^2 的取值为

$$\frac{(97-100)^2}{100} + \frac{(104-100)^2}{100} + \frac{(82-100)^2}{100} + \frac{(110-100)^2}{100} + \frac{(93-100)^2}{100} + \frac{(114-100)^2}{100} = 6.94,$$

跟自由度为 $6 - 1 = 5$ 的 χ^2 分布的上 0.05 分位数 $\chi^2_5(0.2) \approx 7.29$ 比较, 不能拒绝零假设, 即可在显著性水平 0.2 下认为骰子是均匀的.

孟德尔 (Mendel) 豌豆杂交试验。纯黄和纯绿品种杂交，因为黄色对绿色是显性的，在 Mendel 第一定律 (自由分离定律) 的假设下，二代豌豆中应该有 75%是黄色的，25%是绿色的。在产生的 $n = 8023$ 个二代豌豆中，有 $n_1 = 6022$ 个黄色，$n_2 = 2001$ 个绿色。我们的问题是检验这些这批数据是否支持 Mendel 第一定律，要检验的假设是

$$H_0 : \quad \pi_1 = 0.75, \quad \pi_2 = 0.25$$

↑Example

↓Example

解: 在 Mendel 第一定律 (H_0) 下，黄色和绿色的个数期望值为

$\mu_1 = n\pi_1 = 8023*0.75 = 6017.25$, $\mu_2 = n\pi_2 = 8023*0.25 = 2005.75$

则 Pearson χ^2 统计量为

$$Z=\sum \frac{(O-E)^2}{E}=(6022-6017.25)^2/6017.25+(2001-2005.75)^2/2005.75=0.015$$

自由度 $df = 1$，$p - value$ 为 0.99996. 因此可以认为这些数据服从 Mendel 第一定律。Fisher 基于 Mendel 的这些数据，发现其数据与理论值符合的太好，$p - value = 0.99996$，但这么好的拟合在几千次试验中才发生一次，因而 Fisher 断定数据可能有伪造的嫌疑。

(2) 理论总体分布含若干未知参数的情形设某总体 X 服从一个离散分布,

X	a_1	$\dots$	a_k
P	p_1	$\dots$	p_k

$p_i = p_i(\theta_1, \dots, \theta_r), i = 1, \dots, k$ 依赖于 r 个未知参数 $\theta_1, \dots, \theta_r$. 此时理论频数 np_i 一般也与这些参数有关, 从而使用最大似然估计代替这些参数以得到 p_i 的最大似然估计 $\hat{p}_i$, 得到的统计量记为

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

拟合优度检验的提出者 Karl Pearson 最初认为在零假设下, 检验统计量的 χ^2 的极限分布仍等于自由度为 $k-1$ 的 χ^2 分布, R. A. Fisher 发现自由度应该等于 $k-1$ 减去估计的独立参数的个数 r, 即 $k-1-r$.

从某人群中随机抽取 100 个人的血液, 并测定他们在某基因位点处的基因型. 假设该位点只有两个等位基因 A 和 a, 这 100 个基因型中 AA, Aa 和 aa 的个数分别为 30, 40, 30, 则能否在 0.05 的水平下认为该群体在此位点处达到 Hardy-Weinberg 平衡态?

↑Example

↓Example

解: 取零假设为

$$H_0 : \text{Hardy-Weinberg 平衡态成立.}$$

设人群中等位基因 A 的频率为 p, 则该人群在此位点处达到 Hardy-Weinberg 平衡态指的是在人群中 3 个基因型的频率分别为 $P(AA) = p^2$, $P(Aa) = 2p(1-p)$ 和 $P(aa) = (1-p)^2$, 即零假设可等价地写成

$$H_0 : P(AA) = p^2, P(Aa) = 2p(1-p), P(aa) = (1-p)^2.$$

在 H_0 下, 3 个基因型的理论频数为 $100 \times \hat{p}^2$, $100 \times 2 \times \hat{p}^2(1-\hat{p})$ 和 $100 \times (1-\hat{p})^2$, 其中 $\hat{p}$ 等于估计的等位基因频率 0.5, 代入 χ^2 统计量表达式, 得统计量的值等于 4. 该统计量的值大于自由度为 $3-1-1=1$ (恰好一个自由参数被估计) 的 χ^2 分布上 0.05 分位数 3.84, 故可在 0.05 的水平下认为未达到 Hardy-Weinberg 平衡态.

5.3.2 列联表的独立性和齐一性检验

(1) 独立性检验

下面考虑很常用的列联表. 列联表是一种按两个属性作双向分类的表. 例如肝癌病人可以按所在医院 (属性 A) 和是否最终死亡 (属性 B) 分类. 目的是看不同医院的疗效是否不同. 又如婴儿可按喂养方式 (属性 A, 分两个水平: 母乳喂养与人工喂养) 和小儿牙齿发育状况 (属性 B, 分两个水平: 正常与异常) 来分类. 这两个例子中两个属性都只有两个水平, 相应的列联表称为 “四格表”, 一般地, 如果第一个属性有 a 个水平, 第二个属性有 b 个水平, 称为 $a \times b$ 表 (见教材 p268) . 实际应用中, 常见的一个问题是考察两个属性是否独立. 即零假设是

$$H_0 : \text{属性 A 与属性 B 独立.}$$

这是列联表的独立性检验问题.

假设样本量为 n, 第 (i, j) 格的频数为 n_{ij}. 记

$$p_{ij} = P(\text{属性 A, B 分别处于水平} i, j), \tag{5.1}$$

$$u_i = P(\text{属性 A 有水平} i), \tag{5.2}$$

$$v_i = P(\text{属性 B 有水平} j) \tag{5.3}$$

则零假设等价于

$$H_0 : p_{ij} = u_i v_j \quad \forall i, j$$

将 u_i 和 v_j 看成参数, 则总的独立参数有 $a - 1 + b - 1 = a + b - 2$ 个. 它们的极大似然估计为

$$\hat{u}_i = \frac{n_{i\cdot}}{n}, \hat{v}_j = \frac{n_{\cdot j}}{n}.$$

正好是它们的频率 (证明参看教材) . 其中 $n_{i\cdot} = \sum_{j=1}^{b} n_{ij}$, $n_{\cdot j} = \sum_{i=1}^{a} n_{ij}$. 在 H_0 下, 第 (i, j) 格的理论频数为 $n\hat{p}_{ij} = n_{i\cdot} n_{\cdot j}/n$, 因此在 H_0 下, $\sum_{i=1}^{a} \sum_{j=1}^{b} (n_{ij} - n\hat{p}_{ij})$ 应该较小. 故取检验统计量为

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

在零假设下 χ^2 的极限分布是有自由度为 $k - 1 - r = ab - 1 - (a + b - 2) = (a - 1)(b - 1)$ 的 χ^2 分布. 对于四格表, 自由度为 1.

(2) 齐一性检验

跟列联表有关的另一类重要的检验是齐一性检验, 即检验某一个属性 A 的各个水平对应的另一个属性 B 的分布全部相同, 这种检验跟独立性检验有着本质的区别. 独立性问题中两属性都是随机的; 而齐一性问题中属性 A 是非随机的, 这样涉及到的分布实际上是条件分布. 虽然如此, 所采用的检验方法跟独立性检验完全一样.

下面表是甲乙两医院肝癌病人生存情况. 需要根据这些数据判断两医院的治疗效果是否一样.

↑Example

甲、乙两院肝癌的近期疗效

	生存	死亡	合计
甲院	150(n_{11})	88(n_{12})	238($n_{1.}$)
乙院	36(n_{21})	18(n_{22})	54($n_{2.}$)
合计	186($n_{.1}$)	106($n_{.2}$)	292(n)

↓Example

解: 这是一个齐一性检验问题. 检验统计量 χ^2 的观测值为 0.2524, 远远小于自由度为 1 的 χ^2 分布的上 0.05 分位数, 故可以接受零假设, 即在水平 0.05 下可以认为两个医院的疗效无差别的.

5.3.3 连续总体情形

设 $(X_1, \cdots, X_n)$ 是取自总体 X 的一个样本, 记 X 的分布函数为 $F(x)$, 需要检验的那种分布中含有 r 个总体参数 $\theta_1, \cdots, \theta_r$. 我们要在显著性水平 α 下检验

$$H_0 : F(x) = F_0(x; \theta_1, \cdots, \theta_r),$$

其中 $F_0(x; \theta_1, \cdots, \theta_r)$ 表示需要检验的那种分布的分布函数. 例如, 当我们要检验

$$H_0 : X \sim N(\mu, \sigma^2)$$

时, $r=2$, $\theta_1=\mu, \theta_2=\sigma^2$.

$$F_0(x;\mu,\sigma^2)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(t-\mu)^2\right\}dt.$$

上述假设可以通过适当的离散化总体分布, 采用拟合优度法来做检验. 首先把实数轴分成 k 个子区间 $(a_{j-1},a_j]$, $j=1,\cdots,k$, 其中 a_0 可以取 $-\infty$, a_k 可以取 ∞. 这样构造了一个离散总体, 其取值就是这 k 个区间. 记

$$\begin{aligned} p_j &= P_{H_0}(a_{j-1}<X\le a_j) \\ &= F_0(a_j;\theta_1,\cdots,\theta_r)-F_0(a_{j-1};\theta_1,\cdots,\theta_r), j=1,\cdots,k. \end{aligned}$$

如果 H_0 成立, 则概率 p_j 应该与数据落在区间 $(a_{j-1},a_j]$ 的频率 $f_j=n_j/n$ 接近, 其中 n_j 表示相应的频数. 当 p_i 的取值不含未知参数时, 取检验统计量

$$\chi^2=\sum_{j=1}^{k}\frac{(n_j-np_j)^2}{np_j},$$

否则取

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - n\hat{p}_j)^2}{n\hat{p}_j},$$

其中 $\hat{p}_i$ 是将 p_i 中的未知参数换成适当的估计后得到的 p_i 的估计. 拒绝域取为

$$\left\{\chi^2 > \chi^2_{k-r-1}(\alpha)\right\}.$$

如果 p_i 中不含未知参数, 则 $r = 0$.

使用 χ^2 进行拟合优度检验时一般要求 $n \geq 50$, $n\hat{p}_j \geq 5, j = 1, \cdots, k$, 如果不满足这个条件, 最好把某些组作适当合并.

从某连续总体中抽取一个样本量为 100 的样本, 发现样本均值和样本标准差分别为 -0.225 和 1.282, 落在不同区间的频数如下表所示:

↑Example

区间	$(-\infty,-1)$	$[-1,-0.5)$	$[-0.5,0)$	$[0,0.5)$	$[0.5,1)$	$[1,\infty)$
观测频数	25	10	18	24	10	13
理论频数	27	14	16	14	12	17

可否在显著性水平 0.05 下认为该总体服从正态分布?

↓Example

解: 设理论正态分布的均值和方差分别为 μ 和 σ^2, 记第 i 个区间为 $(a_{i-1}, a_i,\ i = 1, \cdots, 6,$ 则样本落在第 i 个格子的理论概数为 $100P(a_{i-1} < X \leq a_i)$, 其中 $X \sim N(\mu, \sigma^2)$. 将 $\mu = -0.225$ 和 $\sigma^2 = \frac{99}{100} \times 1.282^2 = 1.622$ 代入得到估计的理论频数, 列于上表中.

H_0 : 总体服从正态分布

由此算得检验统计量 χ^2 的值约为 9.25, 与自由度为 6-1-2=3 的 χ^2 分布的上 0.05 分位数 $\chi^2_3(0.05) \approx 7.81$ 比较可以拒绝零假设, 即可以在显著性水平 0.05 下认为该总体不服从正态分布.

P 值　若检验的拒绝域为 $T(\mathbf{X}) > \tau$, 对两组不同的样本 $\mathbf{X}_1$ 和 $\mathbf{X}_2$, 若它们均落在拒绝域:

$$T(\mathbf{X}_1) > \tau, \qquad T(\mathbf{X}_2) > \tau$$

则它们否定原假设的程度一样吗? 如何区分这个差异?

P值 $= P($在H_0 ,得到如检验统计量$T(\mathbf{X})$的值$T(\mathbf{x})$这么大或者更极端$)$

从而可以通过 P 值来比较样本的支持程度.

对不同的水平 α 检验方法, 可以通过比较它们的功效 (二型错误) 来评比优劣.