概率论与数理统计

第九章

韩潇
xhan011@ustc.edu.cn

第九章： 非参数假设检验

大纲:

- 拟合优度检验
- Wilcoxon秩和检验
- 符号检验
- 其他非参数检验概述

9.1拟合优度检验

- 第8章讨论的假设检验问题:总体分布已知，部分参数未知
- 总体分布未知时，可以先判断总体是否来自于某种类型的分布
- 例：H_0:总体为正态分布
- 样本对于原假设的总体分布“拟合”较好：不拒绝原假设.反之拒绝原假设. -拟合优度检验(goodness of fit test)

9.1.1理论分布完全已知且只取有限个值

- 总体的值域为有限集$\{\mathrm{a}_1, \ldots, \mathrm{a}_k\}$
- 简单随机样本，样本量为n, 其中有n_i次取值$\mathrm{a}_i, i =$

  $1, \ldots k. \sum_{i=1}^{k} \mathrm{n}_i = n.$ 给定一个分布律为

$$P(X = \mathrm{a}_i) = p_i, i = 1, \ldots k. \; p_i \text{已知}$$

讨论如下检验问题

$$H_0: P(X = \mathrm{a}_i) = p_i, i = 1, \ldots k \leftrightarrow H_1: \exists j \; s.t. P\left(X = \mathrm{a}_j\right) \neq p_j$$

- 由大数定律，有$\mathrm{n}_i \approx np_i \to \mathrm{a}_i$这个类的理论值或期望值
- $\mathrm{n}_i \to$ 观测值

9.1.1理论分布完全已知且只取有限个值

类别	a_1	a_2	$\cdots$	a_i	$\cdots$	a_k
理论值(E)	np_1	np_2	$\cdots$	np_i	$\cdots$	np_k
观测值(O)	n_1	n_2	$\cdots$	n_i	$\cdots$	n_k
$np_i - n_i$	$np_1 - n_1$	$np_2 - n_2$	$\cdots$	$np_i - n_i$	$\cdots$	$np_k - n_k$

- 检验的想法：最后一行的值越小，则与H_0的相似程度越高 – 用差值来构造统计量

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

9.1.1理论分布完全已知且只取有限个值

定理 9.1 $Pearson-\chi^2$ 检验

如果原假设 H_0 成立, 则当样本量 $n \to \infty$ 时, Z 的分布趋于自由度为 $k-1$ 的 χ^2 分布, 即 χ^2_{k-1}.

♡

由定理9. 1, 可以对H_0做检验。显然，当$Z > C$时拒绝H_0，$Z \leq C$时不拒绝H_0. Z在原假设下近似分布为χ^2_{k-1}, 则$C = \chi^2_{k-1}(\alpha)$. 因此检验为

$$\varphi: Z > \chi^2_{k-1}(\alpha)\text{时拒绝}H_0, \text{否则不能拒绝}H_0$$

9.1.1理论分布完全已知且只取有限个值

- 假定根据一组数据算得$Z = Z_0$:如果原假设成立，出现像 Z_0这样大的差异或更大差异的概率有多大?
- 近似地，我们有

$$\mathrm{p}(Z_0) = P(Z \geq Z_0) = 1 - F_{\chi^2_{k-1}}(Z_0)$$

- $\mathrm{p}(Z_0)$越大, 说明在原假设成立时, 出现Z_0这样大的差异就越不奇怪，从而就越使人们相信原假设的正确性–“拟合优度”
- 因此检验φ等价于$\mathrm{p}(Z_0) < \alpha$时，拒绝原假设.

9.1.1理论分布完全已知且只取有限个值

例 9.1 在一个三班制生产的工厂中, 本月出了 30 次事故, 其中早、中、晚班事故数分别为 12, 6, 12 次. 问事故与班次是否有关?

例 9.2 考虑一个骰子是否均匀问题. 设随机变量 X 取值 $1,\ldots,6$, 事件 $\{X=i\}$ 表示掷出 i 点. 如果骰子是均匀的, 相当于

$$H_0:\mathbb{P}(X=i)=1/6, i=1,\ldots,6,$$

设已作了 $n=6\times10^{10}$ 次投掷, 设得到各点出现的次数分别为

$$\begin{aligned} &n_1=10^{10}-10^6, \quad n_2=10^{10}+1.5\times10^6, \quad n_3=10^{10}-2\times10^6,\\ &n_4=10^{10}+4\times10^6, \quad n_5=10^{10}-3\times10^6, \quad n_6=10^{10}+10^6/2. \end{aligned} \tag{9.5}$$

能否认为骰子是均匀的?

9.1.1理论分布完全已知且只取有限个值

例 9.3 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

$$H_0: \quad p_1 = 0.75, \ p_2 = 0.25$$

9.1.2理论分布完全已知但含有有限个未知参数

- 总体的值域为有限集$\{\mathrm{a}_1,\ldots,\mathrm{a}_k\}$,但分布中有r个未知参数$\theta_1,\ldots,\theta_\mathrm{r}$

$$H_0'\colon P(X=\mathrm{a}_i)=p_i(\theta_1,\ldots,\theta_\mathrm{r}),i=1,\ldots k$$

其中$\mathrm{a_i},i=1,\ldots k$已知，且两两不同$\mathrm{r}<k-1$

- 样本为$(X_1,\ldots,X_\mathrm{n})$,在原假设下，记$\hat{\theta}_1,\ldots,\hat{\theta}_\mathrm{r}$为$\theta_1,\ldots,\theta_\mathrm{r}$的最大似然估计，从而$p_i$ 的最大似然估计为$\hat{p}_i=p_i\left(\hat{\theta}_1,\ldots,\hat{\theta}_\mathrm{r}\right),i=1,\ldots k$
- n_i为样本取a_i的次数,$i=1,\ldots k.$

9.1.2理论分布完全已知但含有有限个未知参数

- 构造统计量

$$Z=\sum\frac{(O-\hat{E})^2}{\hat{E}}=\sum_{i=1}^{k}\frac{(n_i-n\hat{p}_i)^2}{n\hat{p}_i}=\sum_{i=1}^{k}\frac{n_i^2}{n\hat{p}_i}-n$$

- 我们有如下定理

定理 9.2

在一定的条件下，若原假设 H_0' 成立，则当 $n\to\infty$ 时，Z 的分布趋于自由度为 $k-r-1$ 的 χ^2 分布，即 χ^2_{k-r-1}.

♡

9.1.2理论分布完全已知但含有有限个未知参数

- 此时检验为

$$\phi'：当Z > \chi^2_{k-r-1}(\alpha)时，拒绝H_0，否则不能拒绝H_0'$$

例 9.4 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?

9.1.2理论分布完全已知但含有有限个未知参数

- 总体X有无穷多个取值，但其分布仅含有有限个未知参数:

$$H_0'': X \sim F_\theta(x), x \in R,$$

- $\theta = (\theta_1, \ldots, \theta_r)$为未知参数. 例: $X \sim N(\mu, \sigma^2)$
- 离散化:将总体的取值分割为k段

$$(x_0, x_1], \ldots, (x_{k-2}, x_{k-1}], (x_{k-1}, x_k), x_0 = -\infty, x_k = \infty$$

- 则定义离散型随机变量，$Y = a_i$,若$x_{i-1} \le X \le x_i, i = 1, \ldots k$. 当原假设成立时，Y的分布为

$$P(Y = a_i) = p_i(\theta_1, \ldots, \theta_r) = F_\theta(x_i) - F_\theta(x_{i-1})$$

- 转换(非等价)为H_0'的检验问题-拒绝H_0',则有理由拒绝H_0''

9.1.2理论分布完全已知但含有有限个未知参数

例 9.5 在一高速路的收费站记录了 106 分钟内在每一分钟内到达收费站的车辆个数. 数据如下表 9.1, 若用 X 表示每一分钟内到达收费站的车辆个数, 试问 X 是否服从某个 Poisson 分布?

表 9.1: 一分钟内达到的车辆个数

车辆个数 (x_i)/m	出现的次数	x_i/m	出现的次数	x_i/m	出现的次数
0	0	7	12	14	4
1	0	8	8	15	5
2	1	9	9	16	4
3	3	10	13	17	0
4	5	11	10	18	1
5	7	12	5		n=106
6	13	13	6		

9.1.2理论分布完全已知但含有有限个未知参数

例 9.6 调查某企业 745 人的收入 (元) 状况如下：

每月收入	≤ 1500	$(1500, 2500]$	$(2500, 3500]$	$(3500, 5000]$	$(5000, 7500]$	> 7500
人数	150	200	220	100	50	25

问该企业的收入能否用正态分布来拟合? ($\alpha = 0.05$)

9.1.2列联表检验

- 理论分布类型已知,但有若干参数未知的检验常用于列联表(contingency table)检验
- 列联表是一种按两个属性作双向分类的表-例:一群人按照吸烟和不吸烟(属性A)和是否患肺癌(属性B)分类
- 记属性A有a个不同水平, 属性B有b个不同水平,则有

$$n_{i\cdot} = \sum_{j=1}^{b} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{a} n_{ij}$$

分别为A处于水平i和属性B处于水平j的个体数.

9.1.2列联表检验

- X:属性A的水平, $X = 1, \dots, a$; Y:属性B的水平, $Y = 1, \dots, a$
- 令$p_{ij} = P(X = i, Y = j) = P$(属性$A, B$分别处于水平$i, j$)
- 将他们的频数罗列出来，可以得到列联表
- 目的：H_0: A, B两属性独立

9.1.2列联表检验

表 9.3: $a \times b$ 列联表

B \ A	1	2	$\cdots$	i	$\cdots$	a	和
1	n_{11}	n_{21}	$\cdots$	n_{i1}	$\cdots$	n_{a1}	$n_{.1}$
2	n_{12}	n_{22}	$\cdots$	n_{i2}	$\cdots$	n_{a2}	$n_{.2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	n_{1j}	n_{2j}	$\cdots$	n_{ij}	$\cdots$	n_{aj}	$n_{.j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
b	n_{1b}	n_{2b}	$\cdots$	n_{ib}	$\cdots$	n_{ab}	$n_{.b}$
和	$n_{1.}$	$n_{2.}$	$\cdots$	$n_{i.}$	$\cdots$	$n_{a.}$	n

9.1.2列联表检验

- H_0成立时，我们有

$$p_{ij} = \mathbb{P}(X = i)\mathbb{P}(Y = j) = p_{i\cdot}p_{\cdot j}, \quad i = 1, \ldots, a, j = 1, \ldots, b$$

- 因此，$\sum_{i=1}^{a} p_{i\cdot} = 1, \sum_{j=1}^{b} p_{\cdot j} = 1, p_{i\cdot}, p_{\cdot j} > 0$
- H_0成立时,独立参数的个数: $\mathrm{r} = (\mathrm{a} - 1) + (\mathrm{b} - 1) = a + b - 2$
- 最大似然估计法可以得到$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}, \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$
- 统计量Z为

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

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

9.1.2列联表检验

- 当$n \to \infty$时, Z的渐近分布是自由度为$k-1-r=ab-1-(a+b-2)=(a-1)(b-1)$的$\chi^2$分布，即$\chi^2_{(a-1)(b-1)}$.

例 9.7 为了了解吸烟是否与患肺癌有关, 在 6000 人中作了调查, 数据如下：

	不吸烟	吸烟	$n_{i\cdot}$
无肺癌	3397	2585	5982
患肺癌	3	15	18
$n_{\cdot j}$	3400	2600	6000

问根据以上数据, 吸烟是否与患肺癌有关? ($\alpha = 0.001$)

9.1.2列联表检验

例 9.8 据报道, 不同时代出生的人在请朋友吃饭时的人均消费是不同的, 大体上是 60 岁以上倾向于人均低一点的消费. 为了证实这一报道是否正确, 某机构作了如下一个关于请朋友吃饭人均消费 (元) 的调查：

年龄段 \ 人均消费	[50.80)	[80,120)	[120,150)	[150,200)	>200	$n_{i\cdot}$
(25,45]	11	26	35	20	9	101
(45,60]	21	35	50	30	5	141
>60	20	38	30	15	1	104
$n_{\cdot j}$	52	99	115	65	15	346

问报道的消息是否正确? ($\alpha = 0.05$)