问题回复

这个问题可以归结为两个多项分布概率是否相等的检验问题或者列联表的齐性检验

假设

疾病组的数据$n_1, n_2, \dots, n_r$服从多项分布$\mathcal{M}(n, p)$

对照组的数据$k_1, k_2, \dots, k_r$服从多项分布$\mathcal{M}(k, q)$

其中:

$$n_1 + n_2 + \cdots + n_r = n,\ k_1 + k_2 + \cdots + k_r = k$$

$$p = (p_1, p_2, \dots, p_r),\ q = (q_1, q_2, \dots, q_r)$$

$$p_1 + \cdots + p_r = 1, q_1 + \cdots + q_r = 1$$

$$p_j, q_j \geq 0, j = 1, \dots, r$$

(在您的论文r = 8)

论文的问题归结为：利用上述数据进行假设检验

$$\mathrm{H}_0: p = q \leftrightarrow H_1: p \neq q$$

检验方法为：令

$$v_{\mathrm{j}} = \frac{n_j + k_j}{n + k}, j = 1, \dots, r$$

$$X^2 = (n + k)\left\{\frac{(n_1 - nv_1)^2}{nv_1} + \frac{(n_2 - nv_2)^2}{nv_2} + \cdots + \frac{(n_r - nv_r)^2}{nv_r} + \frac{(k_1 - kv_1)^2}{kv_1} + \frac{(k_2 - kv_2)^2}{kv_2} + \cdots + \frac{(k_r - kv_r)^2}{kv_r}\right\}$$

则当样本量比较大时X^2近似服从自由度为$\mathrm{r} - 1$的卡方分布。

给定检验水平α，查找到卡方分布(下)分位点$X_{1-\alpha}^2(r-1)$，当$X^2 > X_{1-\alpha}^2(r-1)$时否定原假设。

参考资料：

杨振海，《拟合优度检验》，安徽教育出版社，1993（P71, 推论 3.1）

注：1. 书中问题为列联表的齐性检验

2. 估计有些英文书也有

差”三级，而年龄分为儿童、中青年和老年三组，抽查结果汇总下面数据表：

效果 \ 年龄	儿童	中青年	老年	效果总计
显著	58	38	32	128
一般	28	44	45	117
较差	23	18	14	55
年龄总计	109	100	91	300

由上表及公式（3.14）计算得

$$X^2=\Sigma\Sigma(Z_{ij}-\frac{z_{ij}z_{ij}}{300})\cdot\frac{300}{z_{ij}z_{ij}}=13.9$$

取显著水平$\alpha=0.05$，查表得$\chi^2{}_4(0.05)=9.49<13.9$，即有充分证据说明疗效与年龄有关，事实上$13.9>\chi^2{}_4(0.01)=12.277$，显著水平$<0.01$.

（2）二维列联表：齐性检验

另一类重要问题是齐性检验，数据形式和独立性检验相同，但统计概念是完全不同的．所谓齐性检验是检验若干总体是否有相同的分布．这一问题描述如下：设有c个离散总体，每个总体均取r个值，不失一般性这些值是1，2，$\cdots r$，第i个总体的概率函数是$\mathbf{P}_i$：

$$\mathbf{P}_i=(p_{i1},\ \cdots p_{ir})^T,\ \sum_{i=1}^{r}p_{ij}=1,\ p_{ij}>0,i=1,\cdots,c$$

今将第i个总体记为W_i，则

$$P(W_i=j)=p_{ij}\qquad 1\leqslant j\leqslant r$$

对W_i观察n_i次，W_i取j值的次数记为X_{ij}，$j=1,\cdots,r$则$\sum_{i=1}^{r}X_{ij}=n_i$．将c个总体的独立观察结果可列成表3.3：

$$n=\sum_{i=1}^{c}n_i,\ X_{\cdot j}=\sum_{i=1}^{c}X_{ij}$$

表3.3

总体 \ 值	1	2	…	r	样本量
1	X_{11}	X_{12}	…	X_1r	n_1
2	X_{21}	X_{22}	…	X_2r	n_2
⋮	⋮	⋮		⋮	
C	Xc_1	Xc_2	…	Xcr	nr
和	$X_{\cdot 1}$	$X_{\cdot 2}$		X.r	n

上表的形式和表3.1相同，但统计概念是不相同的。两者差别在于对齐性检验数据，各行数据观察彼此是独立的，可人为控制观察第几个总体，而对独立性检验数据，没有这个性质。观察一个对象要同时记录两个属性（例如3.2例，同时观察疗效和年龄段），简言之，$n_i=\sum_{j=1}^{r} x_{ij}$不是随机的，而对独立性检验，该量是随机的。因此对齐性检验不能套用定理3.2。对独立性检验，不能去掉任何一行数据，去掉了数据就不完全，就成了基于不完全数据的独立性检验，是完全不同的统计问题。而对齐性检验，去掉任何一行均不影响问题的性质和处理方法，仅是用$c-1$代替c罢了。

齐性检验就是根据表 2 的数据，检验各总体是否有相同分布。

精确提法如下：令

$$\mathbf{X}=(X_{11},X_{12},\cdots,X_{1r},X_{21}\cdots,X_{2r},\cdots,x_{c1},\cdots x_{c1})^T$$

则

$$P(\mathbf{X}=\mathbf{x})=P(\mathbf{x},\ \theta)=\prod_{i=1}^{c}\left[n_i!\prod_{j=1}^{r}\frac{1}{x!}p_{ij}^{x_{ij}}\right]$$

其中：

$$\theta=(p_{11},\ p_{12},\ \cdots p_{1r},\ \cdots,p_{c1},\ \cdots,p_{cr})^T=(\mathbf{P}_1{}^T,\ \mathbf{P}_2{}^T,\ \cdots,\ \mathbf{P}_c{}^T)^T$$

那么齐性检验问题是参数的检验问题：

$H_0:\mathbf{P}_i=\mathbf{P}_0\ \ i=1,\cdots,c;\ \ H_1:\mathbf{P}_i\ \ 1\leqslant i\leqslant c$不全相等

其中：$\mathbf{P}_i=(p_{i1}, p_{i2},\cdots,p_{ir})^T$，$0\leqslant i\leqslant c$。

今用似然比检验解决这一问题。对给定$\mathbf{X}=\mathbf{x}$时，对数似然函数为

$$l(\theta)=\text{constant}+\sum_{i=1}^{c}\sum_{j=1}^{r}x_{ij}\log p_{ij} \qquad (3.15)$$

且满足约条件

$$\sum_{j=1}^{r}p_{ij}=1,\ i=1,\ \cdots,\ c$$

用乘子法可求p_{ij}的MLE为

$$\hat{p}_{ij}=x_{ij}/n_i \quad i=1,\cdots,c,\ j=1,\cdots,r$$

当$H_0:\mathbf{P}_i=\mathbf{P}$ $1\leqslant i\leqslant c$成立时，对数似然函数为

$$l_0(\theta)=\text{constant}\sum_{j=1}^{r}x_{ij}\log p_i$$

且满足约束条件

$$\sum_{j=1}^{r}p_i=1$$

于是求得p_{0i}的MLE为

$$\hat{p}_{0i}=x_{\cdot i}/n,\ j=1,\cdots,r,n=\sum_{i=1}^{c}n \qquad (3.16)$$

因此，似然比统计量为

$$LR=2l(\hat{\theta}_n)-2l_0(\hat{\theta}_0)$$

$$=2\left\{\sum_{i=1}^{c}\sum_{j=1}^{r}x_{ij}\log\left(\frac{x_{ij}}{nl}\right)-\sum_{j=1}^{r}x_{ij}\log\frac{x_{\cdot i}}{n}\right\} \qquad (3.17)$$

其中$\hat{\theta}_n$是θ的MLE，$\hat{\theta}_0$是在H_0成立时θ的MLE。当LR过大时拒绝H_0。现给出（3.17）式的等价形式，并H_0成立的条件下，求其极限分布。

令 $\mathbf{P}=\sum_{i=1}^{c}\frac{n_i}{n}\mathbf{P}_i=(p_1,\cdots,p_r)$，$H_0$成立意味着$\mathbf{P}=\mathbf{P}_0$但反之不对。

由大数定理

$$\frac{x_{ij}}{n_i} \to p_{ij} \quad a,\ s,\ n_i \to \infty \quad \begin{matrix} i=1,\cdots,c, \\ j=1,\cdots,r \end{matrix} \qquad (3.18)$$

$$\frac{x_{\cdot j}}{n} \to p_j \quad a,\ s,\ n \to \infty \quad j=1,\cdots,r$$

$$\mathrm{LR}=2\{\sum_{j=1}^{r}\sum_{i=1}^{c} x_{ij}\log\frac{x_{ij}}{n_i} - \sum_{j=1}^{r} x_{\cdot j}\log\frac{x_{\cdot j}}{n}\}$$

$$=2\sum_{j=1}^{r}\sum_{i=1}^{c} x_{ij}\log\frac{x_{ij}}{n_i p_i} + 2\sum_{j=1}^{r}\sum_{i=1}^{c} x_{ij}\log p_i$$

$$-2\sum_{j=1}^{r} x_{\cdot j}\log\frac{x_{\cdot j}}{n}$$

$$=2\sum_{j=1}^{r}\sum_{i=1}^{c} x_{ij}\log\frac{x_{ij}}{n_i p_i} - 2\sum_{j=1}^{r} x_{\cdot j}\log\frac{x_{ij}}{np_i}$$

由(3.18)和*Talor*公式，若n_i $1\leqslant i\leqslant c$同时$\to\infty$时，

$$2\sum_{j=1}^{r}\sum_{i=1}^{c} x_{ij}\log\frac{x_{ik}}{n_i p_i} = 2\sum_{k=1}^{r}\sum_{i=1}^{c} x_{ij}\log\left(1+\frac{x_{ij}-n_i p_i}{n_i p_i}\right)$$

$$=2\sum_{j=1}^{r}\sum_{i=1}^{c} x_{ij}\left\{\left(\frac{x_{ij}-n_i p_i}{n_i p_i}\right) - \frac{1}{2}\frac{(x_{ij}-n_i p_i)^2}{(n_i p_i)^2} + O_p\left(\left(\frac{x_{ij}-n_i p_i}{n_i p_i}\right)^3\right)\right\}$$

$$\doteq 2\sum_{j=1}^{r}\sum_{i=1}^{c}\frac{(x_{ij}-n_i p_i)^2}{n_i p_i} - \sum_{j=1}^{r}\sum_{i=1}^{c}\frac{(x_{ij}-n_i p_i)^2}{n_i p_i}\cdot\left(1+\frac{x_{ij}-n_i p_i}{n_i p_i}\right)$$

$$\overset{e}{=}\sum_{j=1}^{r}\sum_{i=1}^{c}(x_{ij}-n_i p_i)^2/n_i p_i$$

同理可证

$$2\sum_{j=1}^{r} x_{ij}\log\frac{x_{\cdot j}}{n} \overset{e}{=} \sum_{j=1}^{r}(x_{\cdot j}-np_i)^2/np_i$$

于是当n_i，$1\leqslant i\leqslant c$，$n\to\infty$时

$$\mathrm{LR}\overset{e}{=}\sum_{i=1}^{c}\sum_{j=1}^{r}(x_{ij}-np_i)^2/n_i p_i - \sum_{i=1}^{r}(x_{\cdot i}-np_i)^2/np_i \qquad (3.19)$$

关于LR的极限分布有以下定理.

定理3.4 若

(1) $H_0:\mathbf{P}_2=\mathbf{P}\quad 1\leqslant i\leqslant c$, $\mathbf{P}=(p_1, \cdots, p_r)^T$成立,

(2) $\lim n_i/n=\alpha_i$, $1\leqslant i\leqslant c$.

则LR的极限分布为$\chi^2_{(r-1)(c-1)}$.

证明 令

$$Y_{ij}=\sqrt{n_i}\left(\frac{X_{ij}}{n_i}-p_j\right)/\sqrt{p_j}$$

$$\mathbf{Y}_i=(\mathbf{Y}_{i1}, \mathbf{Y}_{i2}\cdots, \mathbf{Y}_{ir})^T \qquad 1\leqslant i\leqslant c$$

$$\mathbf{Y}=(\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_c)^T$$

则(3.19)式可写作

$$\mathrm{LR}\overset{e}{=}\sum_{i=1}^{c}\mathbf{Y}_i{}^T\mathbf{Y}_i-\left(\sum_{i=1}^{c}\sqrt{\frac{n_i}{n}}\mathbf{Y}_i\right)^T\left(\sum_{i=1}^{c}\sqrt{\frac{n_i}{n}}\mathbf{Y}_i\right) \tag{3.20}$$

由引理3.1和假设条件

$$\mathbf{Y}_i\xrightarrow{\mathrm{L}}N_r(\theta, I_1-\sqrt{\mathbf{P}}\sqrt{\mathbf{P}}^T) \qquad 1\leqslant i\leqslant c$$

LR的极限分布与变量$R=\sum_{i=1}^{c}\mathbf{Z}_i{}^T\mathbf{Z}-\left(\sum_{i=1}^{c}\sqrt{\alpha}_i\mathbf{Z}_i\right)^T$

$\cdot\left(\sum\sqrt{\alpha}_i\mathbf{Z}_1\right)$的分布相同, 其中: $\mathbf{Z}_1, \cdots, \mathbf{Z}_c$相互独立且.

$$\mathbf{Z}_i\sim N(\theta, I_r-\sqrt{\mathbf{P}}\sqrt{\mathbf{P}}^T)$$

取c阶正交阵P^1, 其第一列为$(\sqrt{\alpha_1}, \sqrt{\alpha_2}, \cdots, \sqrt{\alpha_c})$.

令

$$\mathbf{V}=(\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_c)^T=P\mathbf{Z}\equiv P(\mathbf{Z}_1, \cdots, \mathbf{Z}_c)^T$$

则$\mathbf{V}_1, \cdots\mathbf{V}_c$相互独立, $\mathrm{Var}\mathbf{V}_i=I-P^1\sqrt{\mathbf{P}}\sqrt{\mathbf{P}}P\equiv\Sigma$, $\Sigma^2=\Sigma$.

则

$$\sum_{i=1}^{c}\mathbf{Z}_i{}^T\mathbf{Z}-\left(\sum_{i=1}^{c}\sqrt{\alpha}_i\mathbf{Z}_i\right)^T\left(\sum_{i=1}^{c}\sqrt{\alpha}_i\mathbf{Z}_i\right)$$

$$=\sum_{i=2}^{c}\mathbf{V}_i{}^T\mathbf{V}_i$$

由引理3.2

$$\mathbf{V}_i{}^T\mathbf{V}_i\sim\chi^2_{r-1}$$

(3.20)式其V_i的相互独立性 知$\sum_{i=2}^{c} V_i{}^T V_i \sim \chi^{2(c-1)(r-1)}$

定理得证.

推论3.1 若定理3.4的假设条件成立，则

$$X^2 = n\sum_{i=1}^{c}\sum_{j=1}^{r}(X_{ij}-n_iX_{\cdot i})^2/n_iX_{\cdot i} \qquad (3.21)$$

的极限分布是$\chi^{2}_{(r-1)(c-1)}$ •

证明

$$X^2 = n\sum_{j=1}^{r}\sum_{i=1}^{c}\{(X_{ij}-np_i)^2+(n_ip_i-\frac{n_iX_{\cdot i}}{n})^2+$$

$$+2(X_{ij}-np_i)(n_ip_i-\frac{n_iX_{\cdot i}}{n})\}/n_iX_{\cdot i}$$

$$n\sum_{j=1}^{r}\sum_{i=1}^{c}(n_ip_i-\frac{n_iX_{\cdot i}}{n})^2/n_iX_{\cdot i}$$

$$=n\sum_{j=1}^{r}\sum_{i=1}^{c}n_i(p_i-\frac{X_{\cdot i}}{n})^2/X_{\cdot i}=\sum_{j=1}^{r}\frac{1}{X_{\cdot i}}\cdot(X_{\cdot i}$$

$$-np_i)^2$$

$$\overset{c}{=}\sum_{i=1}^{r}(X_{\cdot i}-np_i)^2/np_i \qquad (因为\frac{X_{\cdot i}}{n}\to p_i\ a,\ s).$$

$$2n\sum_{j=1}^{r}\sum_{i=1}^{c}(X_{ij}-n_ip_i)(n_ip_i-\frac{n_iX_{\cdot i}}{n})/n_iX_{\cdot i}$$

$$=2n\sum_{j=1}^{r}\sum_{i=1}^{c}(X_{ij}-n_ip_i)(p_i-\frac{X_{\cdot i}}{n})/X_{\cdot i}$$

$$=2n\sum_{j=1}^{r}(p_i-\frac{X_{\cdot i}}{n})(X_{\cdot i}-np_i)/X_{\cdot i}$$

$$=-2\sum_{i=1}^{r}(X_{\cdot i}-np_i)^2/X_{\cdot i}$$

$$\overset{e}{=}-2\sum_{j=1}^{r}(X_{\cdot i}-np_i)^2/np_i$$

因此有

$$X^2=\sum_{i=1}^{c}\sum_{i=1}^{r}n(X_{ij}-n_ip_i)^2/n_iX_{ij}-$$

$$\sum_{i=1}^{r}(X_{\cdot i}-np_i)^2/np_i$$