非参数统计第二次作业

列联表

华中科技大学

数学与统计学院

统计 1701 班

尹恒

U201710027

2020 年 5 月 10 日

目录

1 摘要

由于今年的特殊疫情，本文选取与医疗相关的三个股票板块，并进行板块的简要分析。在每个版块内各取二十支股票计算半年收益率。以此为基础，按照优良中三个级别分别进行概率一致性检验、按总中位数进行中位数检验，并计算 *Cramér* 关联系数、*Pearson* 关联系数以及 *Pearson* 均方关联系数，最后进行实验结果分析。本文还涉及到概率一致性检验中检验统计量适用于 2×2 列联表的推论。

2 运用 M-H 检验解释辛普森悖论

2.1 辛普森谬论

辛普森悖论 (Simpson's Paradox) 亦有人译为辛普森诡论，为英国统计学家 E.H. 辛普森 (E.H.Simpson) 于 1951 年提出的悖论，即在某个条件下的两组数据，分别讨论时都会满足某种性质，可是一旦合并考虑，却可能导致相反的结论。

当人们尝试探究两种变量 (比如新生录取率与性别) 是否具有相关性的时候，会分别对之进行分组研究。然而，在分组比较中都占优势的一方，在总评中有时反而是失势的一方。该现象于 20 世纪初就有人讨论，但一直到 1951 年，E.H. 辛普森在他发表的论文中阐述此一现象后，该现象才算正式被描述解释。后来就以他的名字命名此悖论，即辛普森悖论。此悖论的最终原因和选择偏差、幸存者偏差、以及柏克森悖论一样，是源自对撞因子。

为了避免辛普森悖论的出现，就需要斟酌各分组的权重，并乘以一定的系数去消除以分组数据基数差异而造成的影响。同时，我们必需清楚了解情况，以综合考虑是否存在造成此悖论的潜在因素。

2.2 M-H 检验

2.2.1 实验原理

有时需要将几个 2×2 列联表合成一个做整体分析。当一个整体试验包括几个在不同环境中操作的小试验时，在零假设下的共问的概率随着环境的不同而不同，并且每一个小试检都有自己的 2×2 列联表，这时常常需要进行这种处理。因为得到每个列联表的环境不同，所以这几个表不能合成单一的一个 2 x 2 列联表。

数据 将数据综合以后放入几个 2×2 列联表中，每个列联表的行列总和都是非随机的. 假设表的数目 $k\geq2$，并且第 i 个表具有如下形式:

假定条件

1. 每个观测只归入到一个单元中。
2. 行列总和确定且不随机。
3. 几个列联表是由独立的实验得到的。

	列 1	列 2	
行 1	x_i	$r_i - x_i$	r_i
行 2	$c_i - x_i$	$N_i - r_i - c_i + x_i$	$N_i - ri$
	c_i	$N_i - c_i$	N_i

检验统计量

$$T_4 = \frac{\sum x_i - \sum \frac{r_i c_i}{N_i}}{\sqrt{\sum \frac{r_i c_i (N_i - r_i)(N_i - c_i)}{N_i^2 (N_i - 1)}}}$$

T_4 近似服从标准正态分布，即 $T_4 \sim N(0,1)$

假设检验:

在第 i 个列联表中，p_{1i} 是被归入第一行第一列中的观测的概率，p_{2i} 是第二行第一列相应的概率

A.(双边检验)

$$H_0 : p_{1i} = p_{2i}, \quad \forall i = 1, 2, \ldots, k$$

$$H_1 : p_{1i} > p_{2i}, \text{ 对某个i 或} p_{1i} < p_{2i}$$

拒绝域 $\left\{T_4 \leq z_{\frac{\alpha}{2}}\right\} \cup \left\{T_4 \geq z_{1-\frac{\alpha}{2}}\right\}$。

B.(左边检验)

$$H_0 : p_{1i} \geq p_{2i}, \quad \forall i = 1, 2, \ldots, k$$

$$H_1 : p_{1i} \leq p_{2i}, \quad \forall i = 1, 2, \ldots, k \text{ 且对某个} i \text{ 有} p_{1i} < p_{2i}$$

拒绝域：$\{T_4 \leq z_\alpha\}$

C.(右边检验)

$$H_0 : p_{1i} \leq p_{2i}, \quad \forall i = 1, 2, \ldots, k$$

$$H_1 : p_{1i} \geq p_{2i}, \quad \forall i = 1, 2, \ldots, k \text{ 且对某个} i \text{ 有} p_{1i} > p_{2i}$$

拒绝域：$\{T_4 \geq z_\alpha\}$

注：若行或列随机，这种检验仍然有效，T_4 换为如下 T_5 更准确

$$T_5 = \frac{\sum x_i - \sum \frac{r_i c_i}{N_i}}{\sqrt{\sum \frac{r_i c_i (N_i - r_i)(N_i - c_i)}{N_i^3}}} \sim N(0,1)$$

2.2.2 检验实例

现有两个车间生产螺母和螺钉的合格率如下:

	不合格数	合格数	总计	不合格率
车间 A	2300	25000	27300	8.4%
车间 B	2000	25000	27000	7.4%
总计	4300	50000	54300	

表 1: 整体统计数据

如果不考虑将螺栓细分为螺钉和螺母，则两个车间的不合格率存在显著差异，且 B 车间的不合格率更低一些.

	车间	不合格数	合格数	总数	不合格率
螺母	车间 A	2000	20000	22000	9.09%
	车间 B	600	5000	5600	10.71%
螺钉	车间 A	300	5000	5300	5.66%
	车间 B	1400	20000	21400	10.71%

表 2: 分组统计数据

以螺钉和螺母作为层，则看到无论是哪一种产品，都是车间 A 的不合格率更低。

作出假设：

H_0 : 车间 A 不合格率高

H_1 : 车间 A 不合格率较低

用 M-H 检验

$H_0 : p_{1i} \geq p_{2i}, \quad \forall i = 1,2,\ldots,k$

$H_1 : p_{1i} \leq p_{2i}, \quad \forall i = 1,2,\ldots,k$ 且对某个i 有$p_{1i} < p_{2i}$

计算检验统计量 T_4 :

$$T_4 = \frac{(2000+300) - (\frac{22000\times2600}{27600} + \frac{5300\times1700}{26700})}{\sqrt{\frac{22000\times5600\times2600\times25000}{27600^2\times27599} + \frac{5300\times21400\times25000\times1700}{26700^2\times26699}}} = -4.3648 < z_{0.95}$$

故拒绝原假设，车间 A 的不合格率较低，从表一整体数据得到车间 A 的不合格率高。从统计学的角度来说我们的实验结果是没有错误的，说明我们的直觉出现了偏差，因此这种现象被称为谬论，从数据上看，由于车间 A 关于螺母和螺钉的不合格数差异非常大，导致整体数据发生变化，除此之外，应该有其他因素：比如每个工厂的机器不同，导致生产不同物品的能力不同，导致不合格率会有定向的偏差。

想要避免辛普森谬论的出现，我们就应该在分析问题是注意分组的权重的影响，同时考虑是否有现实生活中的潜在因素进行综合分析，因为这种因素一般不容易从数据中直接观测到，没有接触到原始数据，所以要多加分析。

3 证明题 2

思考题: 证明: 如果 $r=2, c=2$, 那么:

$$T=\sum_{i=1}^{n}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}},\ \text{其中}E_{ij}=\frac{n_i C_j}{N}$$

等价于

$$\mathrm{T}=\frac{N(O_{11}O_{22}-O_{12}O_{21})^2}{n_1 n_2 C_1 C_2}$$

证明:

$$\begin{aligned}
T&=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\\
&=\frac{\left(O_{11}-\frac{n_1C_1}{N}\right)^2}{\frac{n_1C_1}{N}}+\frac{\left(O_{12}-\frac{n_1C_2}{N}\right)^2}{\frac{n_1C_2}{N}}+\frac{\left(O_{21}-\frac{n_2C_1}{N}\right)^2}{\frac{n_2C_1}{N}}+\frac{\left(O_{12}-\frac{n_1C_2}{N}\right)^2}{\frac{n_1C_2}{N}}\\
&=\frac{(O_{11}N-n_1C_1)^2}{Nn_1C_1}+\frac{(O_{12}N-n_1C_2)^2}{Nn_1C_2}+\frac{(O_{21}N-n_2C_1)^2}{Nn_2C_1}+\frac{(O_{22}N-n_2C_2)^2}{Nn_2C_2}\\
&=\frac{(O_{11}(n_1+n_2)-n_1C_1)^2}{Nn_1C_1}+\frac{(O_{12}(n_1+n_2)-n_1C_2)^2}{Nn_1C_2}+\frac{(O_{21}(n_1+n_2)-n_2C_1)^2}{Nn_2C_1}+\frac{(O_{22}(n_1+n_2)-n_2C_2)^2}{Nn_2C_2}\\
&=\frac{(O_{11}n_2-O_{21}n_1)^2}{Nn_1C_1}+\frac{(O_{12}n_2-O_{22}n_1)^2}{Nn_1C_2}+\frac{(O_{21}n_1-O_{11}n_2)^2}{Nn_2C_1}+\frac{(O_{22}n_1-O_{12}n_2)^2}{Nn_2C_2}\\
&=\frac{C_2(n_2O_{11}-n_1O_{21})^2+C_1(n_1O_{22}-n_2O_{12})^2}{n_1n_2C_1C_2}
\end{aligned}$$

对于上式的分子:

$$\begin{aligned}
\text{分子}&=(N-C_1)(n_2O_{11}-n_1O_{21})^2+C_1(n_1O_{22}-n_2O_{12})^2\\
&=(N-C_1)((N-n_1)O_{11}-n_1O_{21})^2+C_1(n_1O_{22}-(N-n_1)O_{12})^2\\
&=(N-C_1)(O_{11}N-n_1C_1)^2+C_1(n_1C_2-O_{12}N)^2\\
&=(N-C_1)(O_{11}N-n_1C_1)^2+C_1(n_1(N-C_1)-O_{12}N)^2\\
&=(N-C_1)(O_{11}N-n_1C_1)^2+C_1(O_{11}N-n_1C_1)^2\\
&=N(O_{11}N-n_1C_1)^2\\
&=N(O_{11}O_{11}+O_{11}O_{12}+O_{11}O_{21}+O_{11}O_{22}-(O_{11}+O_{12})(O_{11}+O_{21}))^2\\
&=N(O_{11}O_{22}-O_{12}O_{21})^2
\end{aligned}$$

即

$$\mathrm{T}=\frac{N(O_{11}O_{22}-O_{12}O_{21})^2}{n_1n_2C_1C_2}$$

原命题得证。

4 概率差异的 χ^2 检验 ($r \times c$ 情形)

4.1 实验原理

数据

共有 r 个总体，从每一个总体中抽取一个随机变量。第 i 个样本的观测数为 n_i，每个样本的观测可归入 c 类中的一类，O_{ij} 为样本 i 的观测归入第 j 类的数目。

	类 1	类 2	$\cdots$	类 c	总和
总体 1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
总体 2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
总体 r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	n_r
	C_1	C_2	$\cdots$	C_c	N

假定条件

1. 每个样本都是一个随机样本。
2. 不同样本的输出结果是相互独立的。
3. 每个观测只能归入 c 类中的一类。

检验统计量 给定检验统计量 T 为：

$$T = \sum_{i=1}^{n}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}} - N, \text{ 其中} E_{ij} = \frac{n_i C_j}{N}$$

这时如果 H_0 为真，O_{ij} 代表格 (i,j) 的观测数，E_{ij} 代表格 (i,j) 期望的观测数。

零分布 T 的零分布是渐近自由度为 $(r-1)(c-1)$ 的 χ^2 分布. 即：

$$T \sim \chi^2((r-1)(c-1))$$

假设 记 p_{ij} 为随机取到第 i 个总体划分为第 j 个类的概率，$i = 1,2,\cdots,r, j = 1,2,\cdots c.$

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{rj}, \forall j = 1,2,\ldots,c$$

$$H_1 : \text{等式不全成立}$$

拒绝域:$\left\{T \geq \chi^2_{1-\alpha}((r-1,c-1))\right\}$

4.2 检验过程

在英为财情上选择和医疗有关的三个板块，每个板块随机选取 20 支股票，记录数据并进行概率一致性检验。

计算所有股票最近半年的收益率如下表：

长春高新	33.31	迈瑞医疗	40.9	恒瑞医药	6.26
复星医药	26.43	乐普医疗	15.21	药明康德	10.48
上海医药	0.71	健帆生物	54.82	智飞生物	59.4
华东医药	-18.17	鱼跃医疗	70.77	云南白药	1
以岭药业	131.54	大博医疗	40.63	康泰生物	52.64
九州岛通	26.71	欧普康视	39.15	片仔癀	27.65
华润三九	-8.3	南微医学科技	13.79	华兰生物	43.76
步长制药	6.74	国药股份	14.33	沃森生物	24.75
天士力	-9.14	山东药玻	36.25	新和成	17.33
信立泰	-6.12	凯利泰	77.49	安图生物	39.66
吉林敖东	-6.78	蓝帆医疗	32.78	上海莱士	14.15
浙江医药	33.93	中国医药	9.81	白云山	-13.23
海正药业	42.28	济民制药	-14.96	凯莱英	45.74
恩华药业	15.22	心脉医疗	30.66	康龙化成	24.33
康恩贝	-16.1	Allmed Medical	54.68	华熙生物	13.74
东诚药业	-1.61	英科医疗	239.48	华大基因	53.19
金达威	2.92	开立医疗	22.3	健友股份	39.44
中恒集团	3.68	三诺生物	16.54	贝达药业	43.68
贵州百灵	-5.4	万东医疗	40.89	丽珠集团	18.04
海王生物	12.06	理邦仪器	83.8	天坛生物	21.98

根据收益率划分等级：若收益率低于 15%, 则记为中，若收益率高于 15% 低于 50%，则记为良，若收益率大于 50%，则记为优. 分别计算三个板块的等级统计数据如下:

	优	良	中	总和
医疗设备与用品	6	10	4	20
主要药物	1	6	13	20
生物技术与药物	3	11	6	20
总计	10	27	23	60

表 3: 等级统计数据表

分别对三个板块的等级做饼状图如下：

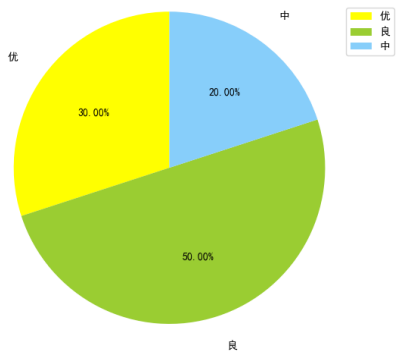

图 1: 医疗设备与用药

由图表可以看出三个板块在三个收益率等级的数量还是存在明显可见的差异，由此我们可

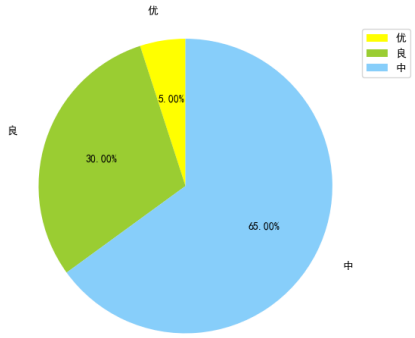

图 2: 主要药物

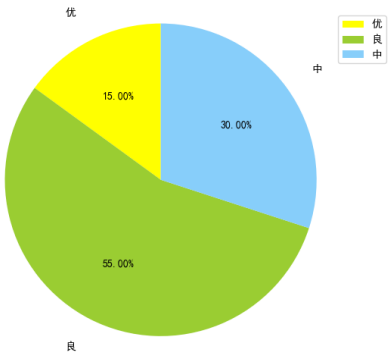

图 3: 生物技术与药物

以先预估概率一致性检验的结果为概率不一致。

作出假设：

H_0 : 不用板块中股票半年收益率为优、良、中的概率相等

H_1 : 存在两个不同的板块的股票半年收益率为优、良、中的概率不同

将观测数据代入得：

$$T = \sum_{i=1}^{n}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}} - N = 11.1816 > \chi^2_{0.95}(4) = 9.488$$

所以拒绝原假设，即三个板块的半年收益率有显著的等级差异。与预期估计一致。

5 中位数检验

用平均收益率做代表，检验三个板块收益率中位数是否相同。

5.1 实验原理

数据

从 C 个总体中各抽取容量为 n_i 的随机样本，则可确定联合样本的中位数，我们称为总中位

数。O_{1i} 为样本 i 超过总中位数的观测个数，O_{2i} 为样本 $i \leqslant$ 总中位数的观测个数。
可作如下列联表：

样本	1	2	$\cdots$	c	总和
$>$ 总中位数	O_{11}	O_{12}	$\cdots$	O_{1c}	a
$\leqslant$ 总中位数	O_{21}	O_{22}	$\cdots$	O_{2c}	b
	n_1	n_2	$\cdots$	n_c	N

假定条件

1. 每一样本都是随机的。

2. 样本之间相互独立

3. 度量尺度至少是顺序的

4. 若所有总体有相同的中位数，则对所有总体而言，一个观测超过总中位数的概率相同，记为 p。

检验统计量

$$T = \sum_{i=1}^{c} \frac{(O_{1i} - O_{2i})^2}{n_i}$$

由于 T 的精确分布很难求得，采用大样本逼近 $T \sim \chi^2(c-1)$

假设检验

$$H_0 : cc \text{ 个总体中有相同的中位数}$$

$$H_1 : \text{至少有两个总体的中位数不同}$$

拒绝域：$\left\{T > \chi^2_{1-\alpha}(c-1)\right\}$

多重比较

若零假设被拒绝，可对 2×2 列联表重复地使用中位数检验，对总体间进行逐行多重比较。

5.2 检验过程

用 excel 处理数据后根据数据绘制列联表：

	医疗设备与用品	主要药物	生物技术与药物	总和
$>$ 总中位数	13	6	11	30
$\leqslant$ 总中位数	7	14	9	30
	20	20	20	60

做出假设：

$$H_0 : \text{三个板块的半年股票收益率有相同的中位数}$$

H_1 : 三个板块的半年股票收益率至少有两个总体的中位数不同

计算检验统计量观测值：

$$T=\sum_{i=1}^{c}\frac{(O_{1i}-O_{2i})^2}{n_i}=5.2<5.991$$

故不能拒绝零假设，即认为三个板块近半年的股票收益率拥有相同的中位数。

由表可以看出三个板块超过和不超过中位数的数量总数相等，但是每个板块的差异较大，统计结果不能拒绝原假设，在实际过程中就要注意是否有其他因素如极端数据等对实验结果的影响过大。

6 计算 Cramer 系数和 R_2, R_3

6.1 *Cramér* 系数

$$T=\sum_{i=1}^{n}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}}-N,\ \text{其中}E_{ij}=\frac{n_iC_j}{N}$$

$$R_1=\frac{T}{T_{\max}}=\frac{T}{N(q-1)}$$

$$Cramér\text{系数}=\sqrt{\frac{T}{N(q-1)}}=\sqrt{\frac{11.1816}{60\times(3-1)}}=0.3053$$

6.2 Pearson 关联系数

$$R_2=\sqrt{\frac{T}{T+N}}=\sqrt{\frac{11.1816}{11.1816+60}}=0.3963$$

6.3 Pearson 均方关联系数

$$R_3=\frac{T}{N}=\frac{11.1816}{60}=0.1864$$

7 总结

本次作业涵盖了几乎所有的列联表相关内容，包括 $M-H$ 检验、概率一致性检验，中位数检验和相依性度量。从辛普森悖论这个很有趣的例子入手，熟悉了各种检验方式和计算，对辛普森悖论的原理，结果有了更清楚的认识。通过对数据的分析也让我更熟悉了 python 和 excel，以及对 latex 更熟练，排版等更高效。