第 12 章 分类数据分析

补充教材

12.6 中位数卡方检验

二分类独立性卡方检验可以用于：检验「独立两组样本」是否有相同的中位数(或分布)。这种检验方法称作「中位数(卡方) 检验」(median test)。

首先将两组样本合起来，计算「共同中位数」。计算各组样本数据，大于「共同中位数」的数目以及小于「共同中位数」的数目。

独立样本	样本 I	样本 II	总和
大于共同中位数的样本量	a	b	$a+b$
小于共同中位数的样本量	c	d	$c+d$
总和	$a+c$	$b+d$	$n=a+b+c+d$

计算 $k^* = \frac{a^2}{(a+b)(a+c)/n} + \frac{b^2}{(a+b)(b+d)/n} + \frac{c^2}{(c+d)(a+c)/n} + \frac{d^2}{(c+d)(b+d)/n} - n$

经过代数运算后 $k^* = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$

若 $k^* \geq \chi^2_{\alpha,1}$，则拒绝 H_0。

12.9 列联系数

列联系数(cotingency coefficient)表达分类数据列联表资料的关联强度，有下列四个系数。列联系数是基于卡方统计量:

$$\chi^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(Y_{ij}-e_{ij})^2}{e_{ij}} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{Y_{ij}^2}{e_{ij}} - n$$

1. ϕ 系数

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

ϕ 系数的数值在 [0,1] 区间，如果 $\phi = 0$ ，则两个属性是独立。列联系数衡量两个属性关系的强度，无法衡量关系的正负方向。

R 语言 R> DescTools::Phi(table) # 请见 12.11 节

2. *Phi* 系数

$$Phi = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

属性	0	1	
0	a	b	$a+b$
1	c	d	$c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

Phi 系数的数值在 [-1,1] 区间，如果 $\phi = 0$ ，则两个属性是独立。$\phi = -1$，则两个属性负的关系。$\phi = +1$ ，则两个属性正的关系。当列联表是 2×2 矩阵，$\phi^2 = (Phi)^2$

R 语言 R> psych::phi(t) # t 是 1×4 向量 或 2×2 矩阵 ，

请见 C:/大话统计学 网络资源/R-code/R_Chap12

3. 皮尔逊列联系数(Pearson's contingency coefficient)

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

皮尔逊列联系数 C 的数值在[0,1] 区间，如果 $\phi = 0$ ，则两个属性是独立。

R 语言 R> DescTools::ContCoef(table)

请见 C:/大话统计学 网络资源/R-code/R_Chap12

4. Cramer V 系数

$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}} \qquad k = \min\{a,b\}$$

Cramer V 系数的数值在 $[0, \sqrt{M/(M-1)}]$ 区间，M=min{a,b}，a 与 b 分别是列联表行与列的数目。如果 $V \leqslant 0.3$，则两个属性是弱相关。如果 $0.3 < V \leqslant 0.7$，则两个属性是中相关。如果 $V > 0.7$，则两个属性是强相关。

R 语言 R> DescTools::CramerV(table)

请见 C:/大话统计学 网络资源/R-code/R_Chap12

12.2 多项分布卡方检验

例题 12.1：遗传学家相信：两只淡褐色的马(palomino)生下来的小马，有四分之一是黑色，有四分之二是淡褐色，有四分之一是白色。现在随机抽样 96 只父母都是淡褐色的小马，其中有 21 只黑色的，52 只淡褐色，23 只白色。这个样本数据能否支持遗传学家的看法。

检验 ：$\begin{cases} H_0 : p_1 = 0.25, p_2 = 0.5, p_k = 0.25 \\ H_1 : \text{以上至少有一为不等式} \end{cases}$

如果检验的显著性水平是 0.05，问检验的结果如何？

解答：$e_1 = np_1^0 = 96(0.25) = 24$ $e_2 = np_2^0 = 96(0.5) = 48$ $e_3 = np_3^0 = 96(0.25) = 24$

y_i	e_i	$y_i - e_i$	$(y_i - e_i)^2$	$\frac{(y_i - e_i)^2}{e_i}$
21	24	-3	9	0.375
52	48	4	16	0.333
23	24	-1	1	0.045
Σ				0.750

因为 $k^* = 0.750 < \chi^2_{0.05,2} = 5.991$，所以接受 H_0

例题 12.7： 同例题 13.7，实验室研究两种生产方法，抽样产品的重量如下：

方法 A：73，67，72，46，83，75，62，90，95

方法 B：71，47，68，87，77，92，65，86，79，57

检验这两种方法的产品重量分布相同。显著性水平 0.05，问检验的结果如何？

解答：两种方法合并中位数是 73，因为在方法 A，所以其样本量减 1。

独立样本	方法 A	方法 B	总和
大于共同中位数的样本量	4	5	9
小于共同中位数的样本量	4	5	9
总和	8	10	18

$$k^* = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{18(20-20)^2}{(8)(10)(9)(9)} = 0$$

因为 $0 = k^* < \chi^2_{0.05,1} = 3.84$，所以接受 H_0。

12.7 两总体独立样本比例检验

二分类联立事件表，如果每个分类都分成两类：

		B 类 B1	B 类 B2	总和
A 类	A1	a	b	$a+b$
A 类	A2	c	d	$c+d$
总和		$a+c$	$b+d$	$n = a+b+c+d$

同构型卡方检验是，利用抽样的样本数据，检验：

$$\begin{cases} H_0 : \text{分類 } \mathbf{B} \text{ 的兩組樣本對於分類 } \mathbf{A} \text{ 有相同比例} \\ H_1 : \text{分類 } \mathbf{B} \text{ 的兩組樣本對於分類 } \mathbf{A} \text{ 沒有相同比例} \end{cases}$$

计算 $k^* = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$

若 $k^* \geq \chi^2_{\alpha,1}$，则拒绝 H_0。

这个问题和第 9.8 节双总体比例检验的问题相同，其检验结果是否一致？

令 B 分类为双总体：B1 总体中 A1 的比例为 p_1，B2 总体中 A1 的比例为 p_2

检验：$H_0: p_1 = p_2$

根据样本数据：$\hat{p}_1 = \frac{a}{a+c}$，$\hat{p}_2 = \frac{b}{b+d}$，$n_1 = a+c$，$n_2 = b+d$

$$z^* = (\hat{p}_1 - \hat{p}_2) \Big/ \sqrt{\left(\frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}\right)\left(1 - \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

经过代数运算 $(z^*)^2 = k^* = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$，$(z_{\alpha/2})^2 = \chi^2_{\alpha,1}$

例题 12.8： 例题 9.8 两条生产线 A,B 抽样及不良品数目：

	生产线 A	生产线 B	总和
不良品	6	3	9
良品	94	97	191
总和	100	100	200

如果检验的显著性水平是 0.05，问检验的结果如何？

解答：$k^* = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{200(6\times 97 - 3\times 94)^2}{(100)(100)(9)(191)} = 1.05 < \chi_{0.05,1} = 3.84$， 接受 H_0。

12.8 两总体成对样本比例检验

McNemar 检验是，检验「成对」(相依)样本，是否有相同比例。名目(分类)尺度的成对样本，和区间尺度的成对样本相同，第一个分类(总体)变量的样本和第二个分类(总体)变量的样本，是成对相依的，譬如时间上的前后或其他相依关系。

成对样本		甲总体 A	甲总体 B	总和
乙总体	A	a	b	$a+b$
	B	c	d	$c+d$
总和		$a+c$	$b+d$	$n = a+b+c+d$

注意：两组成对样本对应相同分类 。

$$\begin{cases} H_0: \text{前後兩組成對樣本對於 } A, B \text{ 兩類有相同比例} \\ H_1: \text{前後兩組成對樣本對於 } A, B \text{ 兩類沒有相同比例} \end{cases}$$

检验步骤：

计算 $\chi^* = \frac{(b-c)^2}{b+c}$ 或修正为 $\chi^* = \frac{(|b-c|-1)^2}{b+c}$

若 $\chi^* \geq \chi^2_{\alpha,1}$，则拒绝 H_0。

例题 12.9： 有 100 人同时应征甲公司与乙公司，其中有 48 人被两家公司录取，有 12 人只被甲公司录取，有 5 人只被乙公司录取，有 35 人两家公司都拒收。检验甲公司与乙公司录取率是否相等。检验的显著性水平是 0.05。

成对样本 (同一人应征)		乙公司 录取	乙公司 拒收	总数
甲公司	录取	48	12	60

	拒收	5	35	40
总数		53	47	100

解答：p_1，p_2 分别为甲公司和乙公司的录取率，检验 $H_0: p_1 = p_2$

$$\chi^* = \frac{(b-c)^2}{b+c} = \frac{(12-5)^2}{12+5} = 2.88 \quad 或 \quad \chi^* = \frac{(|b-c|-1)^2}{b+c} = \frac{(|12-5|-1)^2}{12+5} = 2.12$$

$\chi^* = 2.88 < \chi^2_{\alpha,1} = 3.84$ 或 $\chi^* = 2.12 < \chi^2_{\alpha,1} = 3.84$，所以接受 H_0。

习题

1. 遗传学家相信，某花种子开出红花，白花，黄花的概率分别为：0.25，0.5，0.25。现在 抽样 564 个种子，开出 141 个红花，291 个白花，132 个黄花。利用卡方检验，上述假设是否成立，显著性水平 0.05。

2. 一颗骰子，掷 1000 次，出现点数和次数如下：1 点：158，2 点：172，3 点：164，4 点：181，5 点：160，6 点：165。利用卡方检验，检验这颗骰子是否公平($p_i = \frac{1}{6}$)，显著性水平 0.05。

3. 假设某城市每天停电的次数是：泊松分布平均数 4.2。下列是 150 天的每天停电次数：

每天停电次数	0	1	2	3	4	5	6	7	8	9	10	11
天数	0	5	22	23	32	22	19	13	6	4	4	0

利用卡方检验，上述假设是否成立，显著性水平 0.05。

4. 100 个真空管测试后，有 41 个寿命是小于 30 小时，有 31 个寿命是在 30 小时到 60 小时，有 13 个寿命是在 60 小时到 90 小时，有 15 个寿命是大于 90 小时。如果假设真空管的寿命是指数分布平均数 50 小时。利用卡方检验，上述假设是否成立，显著性水平 0.05。

5. 研究儿童用左手或右手之习惯，抽取 408 位儿童，并依儿童用左手或右手之习惯与其双亲是否有用左手之习惯来分类，得知下列的数据：

	儿童用右手	儿童用左手	总计
双亲皆用右手	307	48	355
双亲至少有一用左手	47	6	53
总计	354	54	408

取 α=0.05，试检验「儿童用左手」与「父母之一有用左手」无关。

6. 随机抽取 1250 位成人，询问是否爱看电视的暴力节目。所得的数据如下：

	回答是	回答否	回答不确定
男性	378	237	26
女性	438	146	25

以 α = 0.05，检验性别与爱看暴力节目有否显著的差异？

7. 问卷调查全国军人对面米食的喜好情形。随机抽取 435 位军人，并依祖籍的区域分类，所得的数据如下：

地区	偏好米食人数	回答人数
南部	65	118
北部	59	135
中部	48	90
西部	43	92

以 $\alpha = 0.05$，偏好米食的习惯，与地区是否有显著的差异？

8. 拟检查某一随机数生成器(random number generator)的质量，取输出结果中的 500 个整数(0 至 9)，记录下各个整数出现的次数。如果随机数生成器是有随机性，则整数 0,1,...,9 出现之概率相等。此随机数生成器是否有任何的偏差(质量不良)？试以卡方检验回答 (α=0.05)。

整数	0	1	2	3	4	5	6	7	8	9	总计
次数	43	58	51	59	39	56	45	37	60	52	500

9. 从台北市民抽出 100 位公民，先调查「是否支持某政党」，然后给他们观看该党宣传影片，再调查「是否支持该政党」，得到数据如下：

	观看前	观看后
支持	62	49
不支持	38	51

(1) α = 0.05，检验「观看宣传影片」是否影响支持该政党比例。

(2) 如果上述观看前、观看后人数改为同一人回答。检验结果是否相同。

成对样本 (同一人回答)		观看后 支持	观看后 不支持	总数
观看前	支持	41	21	62
	不支持	8	30	38
总数		49	51	100

10. 大学生给一篇文章，测验记忆的能力，分数是 1 到 99 分，中位数是 50 分。现在一班抽样 15 个学生，有 10 个学生分数是 50 分以上，有 5 个学生分数是 50 分以下。以α=0.05，检验这一班学生记忆力超过平均水平。

11. 统计学的期末成绩：A,B,C,D,F，学生主修和成绩的人数数据如下：

		学生主修 文科	工科	理科
成绩	A	8	15	13
	B	14	19	15
	C	15	4	7
	D,F	3	1	4

以 $\alpha = 0.05$，检验：学生主修和统计学成绩是否独立。

12. 珍奥斯汀(Jane Austen，1775-1817）是一个英国小说作家，作品有《理性与感性》(Sense and Sensibility)、《傲慢与偏见》(Pride and Prejudice)、《爱玛》(Emma)等，还有一部小说 Sanditon 还没有完成就去世，后来有人完成后半部，现在比较其撰写风格，数据如下：

出现次数	理性与感性	爱玛 Emma	Sanditon 前	Sanditon 后
such 后有 a	14	16	8	2

such 后没有 a	133	180	93	81
and 后有 I	12	14	12	1
and 后没有 I	241	285	139	153
on 后有 the	11	6	8	17
on 后有 the	259	265	221	204

(1) α = 0.05，检验这四本小说的撰写风格是否相同。

(2) 检验《理性与感性》和《爱玛 Emma》的撰写风格是否相同。

(3) 检验《理性与感性》和《Sanditon 前半部》的撰写风格是否相同。

(4) 检验《Sanditon 前半部》和《Sanditon 后半部》的撰写风格是否相同。

这个习题说明：统计学也可以应用到文学。

13. 心理学家测验白领阶级和蓝领阶级的工作态度，四个蓝领阶级的分数是：23，18，22，21。五个白领阶级的分数是：23，28，25，24，26。检验这两个阶级的工作态度没有差异，显著性水平 0.05。

14. 下列数据是连续 25 个产品的质量水平：

100,110,122,132,99,96,88,75,45,211,154,143,161,142,99,111,105,133,142,

150,153,121,126,117,155

检验这个样本数据是正态分布，平均数 124，标准偏差 33

15. 一摇奖机内有标示 0 到 9 的小球十个，连续摇奖 100 次，得各数字出现之次数分布为：

数字 x	0	1	2	3	4	5	6	7	8	9
次数	6	5	5	12	14	16	7	5	15	15

试以显著性水平 $\alpha = 0.05$ 检验各小球出现的概率是否可能均等？

16. 某大型连锁超级市场企划人员想知道目前本县中各超市所贩卖的五种品牌(M, N, K, O, L)果汁，了解这五种品牌果汁的销售是否相同，某日总共有 230 位顾客购买了果汁，其中有 40 位购买 M 牌果汁、50 位购买 N 牌果汁、35 位购买 K 牌果汁、45 位购买 O 牌果汁、60 位购买 L 牌果汁，请根据这个样本回答以下的问题。

(1) 若此大型连锁超级市场企划人员认为这五种品牌(M, N, K, O, L)果汁一样畅销，请用 5%的显著性水平检验?

(2) 若此大型连锁超级市场企划人员认为这五种品牌(M, N, K, O, L)果汁一样畅销，请计算此次检验之 p 值。

(3) 若此大型连锁超级市场企划人员认为这五种品牌果汁的销售比例为(M=15%, N=20%, K=10%, O=25%, L=30%)，请用 5%的显著性水平检验?

(4) 若此大型连锁超级市场企划人员认为这五种品牌果汁销售比例为(M=15%, N=20%, K=10%, O=25%, L=30%)，请计算此次检验之 p 值。

17. 去年家电市场中冰箱的市场由三大厂牌分食，S 牌占了 45%、P 牌占了 35%、K 牌占了 15%、其他占了 5%，K 牌冰箱在今年改变策略大举广告想提升市占率，为了解今年冰箱市场这三种品牌的销售是否与去年相同，抽样了某连锁家电经销商今年的销售总共卖掉 1200 台冰箱，其中有 480 台为 S 牌、380 台为 S 牌、290 台为 S 牌、50 台为其他，请根据这个样本回答以下的问题。

(1) 若此大型连锁超级市场企划人员认为三大厂牌(S, P, K)冰箱的市场今年不变，请用 5%的显著性水平检验?

(2) 若此大型连锁超级市场企划人员认为这三大厂牌(S, P, K)冰箱的市场今年不变，请计算此次检验之 p 值。

18. 某西餐厅中午推出四种特餐(烤鸡、卤牛肉、炸猪排、清蒸鱼)，西餐厅的经理以他过去的经验猜测四种特餐的分布如下：烤鸡占 45%、卤牛肉占 20%、炸猪排占 10%、清蒸鱼占 25%，为了解西餐厅的经理的猜测是否正确，抽样了过去一周 100 位顾客的点餐，其中有 39 位点烤鸡、25 位点卤牛肉、15 位点炸猪排、21 位点清蒸鱼，请根据这个样本回答以下的问题。

(1) 请用 1%的显著性水平检验此西餐厅经理的猜测是否正确?

(2) 请计算此次此西餐厅经理的猜测检验之 p 值。

19. 某大型电子制造商的人事部门想知道公司员工一周 5 天的出席状况，抽样 200 位过去一周曾请假的员工的缺席日，这些缺席日的数据分布如下：

一周每天	星期一	星期二	星期三	星期四	星期五
每日缺席员工数	19	42	57	47	35

人事部门原先猜测的分布为：星期一占 10%、星期二占 25%、星期三占 30%、星期四占 20%、星期五占 15%，请根据这个样本回答以下的问题。

(1) 请用 5%的显著性水平检验此大型电子制造商人事部门的猜测是否正确?

(2) 请计算此次此此大型电子制造商人事部门的猜测检验之 p 值。

(3) 若抽样 100 位过去一周曾请假的员工的缺席日，这些缺席日的数据分布如下：

一周每天	星期一	星期二	星期三	星期四	星期五
每日缺席员工数	10	21	29	22	18

请用 5%的显著性水平检验此大型电子制造商人事部门的猜测是否正确? 请计算此次此此大型电子制造商人事部门的猜测检验之 p 值。

(4) 若抽样 50 位过去一周曾请假的员工的缺席日，这些缺席日的数据分布如下：

一周每天	星期一	星期二	星期三	星期四	星期五
每日缺席员工数	3	10	15	13	9

请用 5%的显著性水平检验此大型电子制造商人事部门的猜测是否正确? 请计算此次此此大型电子制造商人事部门的猜测检验之 p 值。

20. 某消费性商品经销商的人事部门想知道公司员工的年龄与工作兴趣是否有关，抽样 200 位员工对工作的兴趣，这些年龄与工作兴趣的数据分布如下：

年龄	高度工作兴趣	工作兴趣普通	完全无工作兴趣
30 岁以下	31	24	13
30 岁至 50 岁	42	30	4
50 岁以上	32	21	3

请根据这个样本回答以下的问题。

(1) 请用 10%的显著性水平检验此消费性商品经销商公司员工的年龄与工作兴趣是否有关?

(2) 请计算此消费性商品经销商公司员工的年龄与工作兴趣检验之 p 值。

21. 某保险公司分析专员想知道性别(女性或男性)对于保险的看法是否相同，抽样 1200 位大学生询问其可接受的保险种类，这些性别与可接受的保险种类的数据分布如下：

性别	限期储蓄险	寿险	无
女性	100	80	325

男性	160	60	475

请根据这个样本回答以下的问题。

(1)　请用 5%的显著性水平检验性别与可接受的保险种类是否有关?

(2)　请计算性别与可接受的保险种类检验之 p 值。

22. 某运动商品广告经销商想知道年龄别(成人或小孩)与喜欢的球类是否有关，抽样多位成人与小孩询问其最喜欢的球类，这些年龄别与最喜欢的球类的数据分布如下：

年龄别	棒球	篮球	足球	高尔夫球	网球
成人	24	17	30	18	22
小孩	21	20	22	12	28

请根据这个样本回答以下的问题。

(1)　请用 5%的显著性水平检验年龄别(成人或小孩)与喜欢的球类是否有关?

(2)　请计算年龄别(成人或小孩)与喜欢的球类检验之 p 值。

23. 某二手车经销商专卖丰田二手车，其旗下有三位销售员，他们过去 3 个月销售的业绩与车种数据分布如下：

销售员	丰田 Camry	丰田 Luxus	丰田 Yaris
王忠成	7	2	6
李书形	11	4	8
陈文华	8	5	3

请根据这个样本回答以下的问题。

(1)　请用 5%的显著性水平检验销售员与销售车种的业绩是否有关?

(2)　请计算销售员与销售车种的业绩检验之 p 值。

24. 以下数据是 40 位员工的年龄，从某计算机游戏经销商的员工中抽样出来。

26	21	25	20	21	29	26	23	22	24
24	30	23	32	26	24	32	16	36	26
21	31	26	23	32	35	40	30	14	26
46	27	33	25	27	21	26	18	29	36

请根据这个样本回答以下的问题。

(1)　请用 5%的显著性水平检验此计算机游戏经销商员工的年龄是否为正态分布?

(2)　请计算此计算机游戏经销商员工的年龄是否为正态分布检验之 p 值。

25. 假设从某县中抽样出来 150 位居民，询问其去年整年的收入，再根据所算出的平均数与标准偏差将这 150 个数据标准化，填入以下的区间中统计次数：

区间 Intervals	频数 Frequency
$Z \leqq -1.5$	15
$-1.5 < Z \leqq -0.5$	32
$-0.5 < Z \leqq 0.5$	65
$0.5 < Z \leqq 1.5$	25
$Z > 1.5$	13

请根据这个样本回答以下的问题。

(1) 请用 5%的显著性水平检验此县居民整年的收入是否为正态分布?

(2) 请计算此县居民整年的收入是否为正态分布检验之 p 值。