

第九章 χ^2 检验

第一节 四格表资料的 χ^2 检验

一、 χ^2 检验的原理

一句话抓住本质

χ^2 检验不是在检验“频数”，

而是在检验：

👉 现实世界，是否偏离了“原假设下应该看到的世界”。

1. χ^2 检验在回答什么问题？（Purpose）

在四格表中，最典型的问题是：

两组的“率 / 构成比”，是否相同？

比如：

- 治疗组 vs 对照组的有效率
- 男 vs 女的患病率
- 暴露 vs 未暴露的发病率

统计假设写成一句话就是：

$$H_0: \pi_1 = \pi_2$$

👉 χ^2 检验 = “率的差异检验”

2. 核心直觉：什么叫“理论频数”？（非常关键）

先别算公式，先想一件事：

如果 H_0 真的成立,
那每一个格子里 “应该有多少人” ?

这个 “应该有多少” ,
就是 **理论频数 (theoretical frequency, T)** 。

理论频数不是拍脑袋来的

它来自一个极其朴素的逻辑:

行比例 × 列比例

于是得到统一公式:

$$T_{ij} = \frac{n_{i+} n_{+j}}{n}$$

✦ 直觉翻译:

- 第 i 行有多少人?
 - 第 j 列有多少人?
 - 如果 “行和列彼此独立” ,
那它们的交集应该是多少?
-

3. χ^2 统计量到底在算什么? (灵魂)

$$\chi^2 = \sum \frac{(A - T)^2}{T}$$

你只要记住一句话:

“实际看到的”
和
“在 H_0 成立时应该看到的” ,
偏离得有多严重?

为什么是这个形式？

- $A - T$: 偏离量
- 平方: 不让正负抵消
- 除以 T : **标准化**
 - 同样差 5 人
 - 在 $T = 10$ 和 $T = 100$ 的意义完全不同

👉 χ^2 是 “标准化后的偏离总量”

4. 为什么四格表的自由度是 1? (非常爱考)

四格表: $2 \text{ 行} \times 2 \text{ 列}$

在 **行合计、列合计固定** 的前提下:

- 四个格子
- 只有 **1 个可以自由变**

🔴 所以:

$$v = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$$

👉 这就是为什么四格表 χ^2 检验一定是 $v = 1$

5. χ^2 值和 P 值的关系 (不要机械)

当 H_0 成立时:

$$\chi^2 \sim \chi^2(1)$$

直觉记忆法:

- χ^2 小
 - 👉 实际 \approx 理论
 - 👉 P 大 \rightarrow 不拒绝 H_0

- χ^2 大
 - ☞ 实际 \neq 理论
 - ☞ P 小 \rightarrow 拒绝 H_0

✦ 判断规则（你已经很熟）：

- 若 $\chi^2 \geq \chi_{\alpha,1}^2$
 - ☞ $P \leq \alpha \rightarrow$ 拒绝 H_0
- 否则 \rightarrow 不拒绝 H_0

6.把你给的例 9-1 用“一句话”说清楚

【例 9-1】为了解吲达帕胺片治疗原发性高血压的疗效,将 80 名高血压患者随机分为两组,试验组用吲达帕胺片加辅助治疗,对照组用安慰剂加辅助治疗,观察结果见表 9-1,试分析吲达帕胺片治疗原发性高血压的有效性。

表 9-1 两种疗法治疗原发性高血压的疗效

组别	有效	无效	合计	有效率/%
对照组	22 (a)	18 (b)	40 ($a+b$)	55.00
试验组	31 (c)	9 (d)	40 ($c+d$)	77.50
合计	53 ($a+c$)	27 ($b+d$)	80 (n)	66.25

本例为两样本率比较的资料,表 9-1 中的数字 a 、 b 、 c 、 d 表示该表的四个基本数据,进而可以计算出各行列周边的合计数和两组的有效率,称为四格表 (fourfold table) 资料。可以用 χ^2 检验推断两个总体率之间有无差别。

(李康,贺佳主编, 2024, p. 89) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

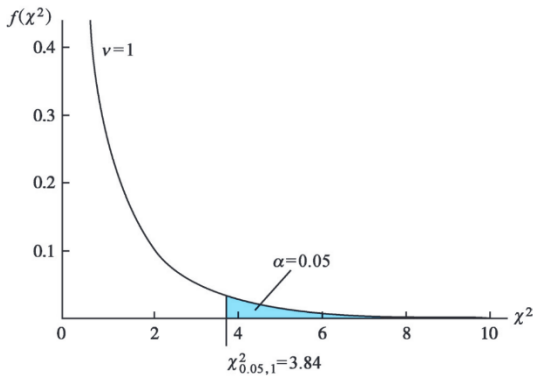


图 9-1 自由度 $\nu=1$ 的 χ^2 分布曲线图

(李康,贺佳主编, 2024, p. 90) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

现以例 9 - 1 为例说明 χ^2 检验的步骤：

(1) 建立检验假设并确定检验水准 $H_0: \pi_1 = \pi_2$, 即试验组与对照组的总体有效率相等 $H_1: \pi_1 \neq \pi_2$, 即试验组与对照组的总体有效率不等 $\alpha = 0.05$

(2) 计算检验统计量 按公式 (9 - 2) 计算 T_{11} , 然后利用四格表的各行列的合计数计算 T_{12} 、 T_{21} 和 T_{22} , 即

$$\begin{aligned} T_{11} &= (40 \times 53)/80 = 26.5, T_{12} = 40 - 26.5 = 13.5 \\ T_{21} &= 53 - 26.5 = 26.5, T_{22} = 27 - 13.5 = 13.5 \end{aligned}$$

按公式 (9 - 1) 计算 χ^2 值

$$\chi^2 = \frac{(22 - 26.5)^2}{26.5} + \frac{(18 - 13.5)^2}{13.5} + \frac{(31 - 26.5)^2}{26.5} + \frac{(9 - 13.5)^2}{13.5} = 4.528$$

(3) 确定 P 值, 作出推断结论 以 $\nu = 1$ 查 χ^2 分布界值表 (附表 7) , 得 $P < 0.05$ 。按 $\alpha = 0.05$ 水准, 拒绝 H_0 , 接受 H_1 , 可以认为两组治疗原发性高血压的总体有效率不等, 即可以认为吲达帕胺片治疗原发性高血压优于对照组。

这个例子在做什么？

- 假设：

两组治疗原发性高血压的**总体有效率相同**

- 用合并有效率，算出：

如果真的相同，每组“应该”有多少有效 / 无效

然后发现：

实际观察到的有效 / 无效分布，
已经偏离到“随机波动解释不了”的程度。

结果：

- $\chi^2 = 4.528$
- 自由度 $v = 1$
- $P < 0.05$

👉 拒绝 H_0

👉 两组总体有效率不等

👉 治疗组优于对照组

8.考试级一句话总结

χ^2 检验通过比较实际频数与在原假设成立时的理论频数之间的偏离程度，判断分类变量之间是否存在统计学关联。对于四格表资料，在原假设成立时， χ^2 统计量近似服从自由度为 1 的 χ^2 分布。

二、四格表 χ^2 检验的专用公式（直觉优先版）

一句话抓住本质

四格表专用公式不是“另一种检验”，
而是：在条件满足时，对通用 χ^2 的一次代数简化。

1.什么时候“必须”用这个专用公式？（Usage）

教材给出的条件非常关键，不是装饰：

- 总例数： $n \geq 40$

- **所有理论频数**: $T_{ij} \geq 5$

✦ 含义是:

χ^2 分布的**大样本近似**已经足够好,
可以安全地用代数公式代替逐格计算。

👉 不满足这些条件:

- 要么用 **连续性校正 χ^2**
- 要么直接用 **Fisher 精确检验**

2.这个公式到底在“比什么”? (核心直觉)

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

先别背, 先看结构。

1 为什么核心是 $ad-bc$?

四格表:

结局 + 结局 -

组 1 a b

组 2 c d

- ad : **同向配对**
- bc : **交叉配对**

✦ 如果两组率相等 (H_0 成立) :

$$ad \approx bc$$

👉 所以:

- $ad - bc$
= 两组“关联强度”的核心量
= 偏离独立性的原始信号
-

2 为什么要平方？

和所有检验一样：

- 不让正负抵消
 - 关注偏离“大小”而不是方向
-

3 为什么要乘以 n ？

样本越大，同样的比例差异，越可信

这是**信息量加权**。

4 分母在干什么？

$$(a + b)(c + d)(a + c)(b + d)$$

它把：

- 行合计
- 列合计

全部纳入，作用只有一个：

标准化偏离量，使其在 H_0 下服从 χ^2 分布

👉 本质上，它和

$$\sum (A - T)^2 / T$$

是在做**同一件事**。

3.为什么“算出来一模一样”？（关键认知）

你已经看到了：

- 通用公式 $\rightarrow \chi^2 = 4.528$
- 专用公式 $\rightarrow \chi^2 = 4.528$

这不是巧合。

✦ **结论一定要记：**

在满足使用条件时，
四格表专用公式与通用公式完全等价。

不同的只是：

- 计算路径
- 不是统计含义

4.什么时候“千万别乱用”？（很重要）

以下情况**不要直接用这个公式**：

- ✗ 有格子的理论频数 < 5
- ✗ 总例数太小
- ✗ 数据极端不平衡

- 👉 后果不是“略不准确”，
- 👉 而是 **P 值系统性失真**。

5.和前面 χ^2 原理的关系（一行打通）

通用公式：

从“理论世界”出发，逐格比较

专用公式：

在四格表结构下，

把“所有偏离”压缩成一个代数量

👉 思想完全一致，只是走捷径。

6.考试级一句话总结

对于两样本率比较的四格表资料，在总例数不小于 40 且各格理论频数不小于 5 的条件下，可采用四格表 χ^2 检验的专用公式计算检验统计量，其结果与通用 χ^2 公式一致。

三、四格表资料 χ^2 检验的校正公式

一句话抓住本质

连续性校正不是“提高显著性”，

而是：防止你因为样本离散性而“过度显著”。

1.为什么普通 χ^2 会“偏激进”？ (Why correction is needed)

根本原因只有一个：

- χ^2 分布是连续分布
- 四格表的频数是离散整数

👉 用连续分布去逼近离散数据，

在小自由度 (尤其 $v=1$) 时会出问题。

✦ 具体后果是：

算出来的 χ^2 偏大 \rightarrow P 值偏小 \rightarrow 更容易拒绝 H_0

也就是说：

- 更容易“显著”
 - 但这是数学近似带来的假象
-

2. Yates 校正到底在做什么？（核心直觉）

Yates 在 1934 年提出了一个非常“工程师式”的修正：

既然连续分布会夸大偏离，
那我就把“偏离量”稍微削小一点。

怎么削？

👉 每一个偏离，先减去 0.5 再平方

3. 通用公式的校正，在干嘛？（不用推）

原来：

$$\chi^2 = \sum \frac{(A - T)^2}{T}$$

校正后：

$$\chi_c^2 = \sum \frac{(|A - T| - 0.5)^2}{T}$$

🔴 直觉翻译：

- $|A - T|$ ：真实偏离
- -0.5 ：给离散性的“缓冲带”
- 平方后再标准化

👉 校正一定会让 χ^2 变小

👉 P 值变大，更保守

4. 四格表专用公式的校正（结构一眼懂）

原来（未校正）：

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

校正后：

$$\chi_c^2 = \frac{(|ad - bc| - n/2)^2 n}{(a + b)(c + d)(a + c)(b + d)}$$

直觉解释（非常重要）：

- 原来用的是 $ad - bc$
- 现在用的是

$$|ad - bc| - n/2$$

👉 把“关联强度”人为压小一点

5. 校正的真实作用：一句话说清

Yates 校正 = 给“自由度为 1 的 χ^2 ”踩刹车

- 防止：
 - 小样本
 - 边缘频数不均
 - 偶然波动
- 导致：
 - 虚假的显著性

✚ 所以它**不是为了更敏感**，
而是为了**不乱下结论**。

6.什么时候用哪一个？（必须背的决策规则）

你给的三条，其实就是**标准决策树**，我帮你翻成“操作语言”。

✅ 情况 1

$n \geq 40$ ，且所有 $T \geq 5$

👉 近似已经很好

👉 直接用普通 χ^2

- 通用公式
 - 或四格表专用公式
-

⚠️ 情况 2

$n \geq 40$ ，但有 $1 \leq T < 5$

👉 连续近似开始不稳

👉 必须用 Yates 校正 χ^2

❌ 情况 3

$n < 40$ 或 $T < 1$

👉 χ^2 近似彻底失效

👉 用 Fisher 确切概率法

7.一个非常重要但常被忽略的事实

Yates 校正往往“过于保守”。

现实中常见情况是：

- 未校正 χ^2 ：显著
- 校正 χ^2 ：不显著

这不是算错，而是：

证据本来就不够强，只是被近似放大了。

8.考试级一句话总结（直接可写）

对于自由度为 1 的四格表资料，由于 χ^2 分布为连续分布而实际频数为离散型变量，采用普通 χ^2 检验可能导致 P 值偏小。Yates 提出的连续性校正通过修正偏离量，使检验结果更加保守，从而提高结论的可靠性。

给你一个永远不会混乱的心智锚点

普通 χ^2 ：踩油门

Yates 校正 χ^2 ：踩刹车

Fisher 精确检验：不开车，直接算概率

【例 9-2】某医师欲比较胞磷胆碱与神经节苷脂治疗脑血管疾病的疗效,将 58 例脑血管疾病患者随机分为两组,结果见表 9-2。问两种药物治疗脑血管疾病的有效率是否不同?

表 9-2 两种药物治疗脑血管疾病有效率的比较

药物分组	有效	无效	合计	有效率/%
胞磷胆碱组	25 (23.66)	3 (4.34)	28	89.29
神经节苷脂组	24 (25.34)	6 (4.66)	30	80.00
合计	49	9	58	84.48

注:括号内数字为理论频数。

(1) 建立检验假设并确定检验水准

$H_0: \pi_1 = \pi_2$, 即两种药物治疗脑血管疾病的总体有效率相等

$H_1: \pi_1 \neq \pi_2$, 即两种药物治疗脑血管疾病的总体有效率不等

$\alpha = 0.05$

(2) 计算检验统计量

按照公式 (9-2) 计算各观察值的理论频数, 记于表 9-2 的括号中。本例 $n=58$, 但有 2 个格子的理论频数分别 4.34 和 4.66, 均符合 $1 \leq T < 5$, 需用四格表资料 χ^2 检验的校正公式 (9-4) 或公式 (9-5)。本例用公式 (9-5) 计算校正 χ^2 值

$$\chi_c^2 = \frac{(|25 \times 6 - 3 \times 24| - 58 / 2)^2 \times 58}{28 \times 30 \times 49 \times 9} = 0.376$$

(3) 确定 P 值, 作出推断结论

以 $\nu=1$ 查 χ^2 分布界值表 (附表 7) 得 $P > 0.05$ 。按 $\alpha = 0.05$ 水准, 不拒绝 H_0 , 尚不能认为两种药物治疗脑血管疾病的有效率不等。

(李康, 贺佳主编, 2024, p. 91) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

四、四格表资料的 Fisher 确切概率法

一句话抓住本质

Fisher 检验不问“偏离大不大”，

而是直接问：

👉 在原假设成立时，这种结果到底有多不可能？

1. 为什么一定要有 Fisher? (Why it must exist)

回顾你前面的路径，其实非常清晰：

- χ^2 检验：
👉 用连续分布近似离散数据
- Yates 校正：
👉 给这种近似踩刹车
- 但当样本再小一点时：

近似本身就不成立了

✚ 所以在以下情况：

- $n < 40$
- 或某个格子的理论频数 $T < 1$

👉 **任何 χ^2 近似（包括校正）都不可靠**

这时，唯一理智的做法是：

放弃近似，直接算“真实概率”

2. Fisher 检验到底在干什么？（核心直觉）

Fisher 的想法非常“物理学家式”：

**如果周边合计数已经固定，
那四格表内部其实只有有限种可能排法。**

于是问题变成：

**在所有“可能的排法”中，
像我们现在看到的这种表，
出现的概率有多小？**

3. 它依赖的世界观：超几何分布（Hypergeometric）

你不用背分布，只要理解一句话：

Fisher 检验假设：
行和列的总数是给定的，
只是在“有限人群里重新分配”。

这正是 **超几何分布** 的应用场景。

4.确切概率公式在算什么？（别被阶乘吓到）

$$P_i = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

你只需要知道三点：

1 这是一个“组合概率”

- 不是近似
- 不是统计量
- 是精确概率

2 周边合计数被当作常量

- 行合计、列合计都不变
- 只是在格子内部“重新分人”

3 所有可能表的概率加起来等于 1

$$\sum P_i = 1$$

👉 这就是“确切”的含义。

5.单侧 vs 双侧：Fisher 真正容易混的地方

核心原则：

- **单侧检验**
 - 👉 累加 “比当前表更极端、且方向一致” 的概率
- **双侧检验**
 - 👉 累加 “概率 \leq 当前表” 的所有表

📌 注意：

**Fisher 的 “双侧”
不是简单乘 2**

这是它和 t / χ^2 的本质不同点之一。

6. Fisher 和 χ^2 的根本差异（一行打通）

χ^2 :

- 用偏离量
- 依赖近似
- 适合中大样本







Fisher:


- 直接算概率
- 不靠近似
- **小样本也绝对可靠**

👉 Fisher 永远不会 “假显著” ,
但代价是：
计算量大、结果更保守。

7. 什么时候用 Fisher?（最终决策规则）

你现在可以把四格表的选择逻辑**彻底封箱**：

-  $n \geq 40$, 所有 $T \geq 5$
 普通 χ^2 / 专用公式
-  $n \geq 40$, 但有 $1 \leq T < 5$
 Yates 校正 χ^2
-  $n < 40$ 或 $T < 1$
 **Fisher 确切概率法**

 这不是经验，是**理论底线**。

8. 考试级一句话总结（直接可写）

Fisher 确切概率法是在四格表周边合计数固定的条件下，基于超几何分布直接计算各可能表型的概率，用以检验分类变量之间是否存在关联，适用于样本量较小或理论频数过低的情况。

【例 9-3】某研究者为研究乙肝免疫球蛋白预防大鼠胎儿宫内感染 HBV 的效果,将 17 例 HBsAg 阳性大鼠随机分为预防注射组和非预防组,观察两组所产出的新生大鼠 HBV 感染情况,结果见表 9-3。问两组新生大鼠的 HBV 总体感染率有无差别?

表 9-3 两组新生大鼠 HBV 感染率的比较

组别	阳性	阴性	例数	感染率/%
非预防组	7	2	9	77.78
预防注射组	2	6	8	25.00
合计	9	8	17	52.94

本例 $n=17 < 40$, 不满足 χ^2 检验的应用条件, 宜用 Fisher 确切概率法, 其假设检验过程如下:

(1) 建立检验假设并确定检验水准

$H_0: \pi_1 = \pi_2$, 即两组新生大鼠 HBV 的总体感染率相等

$H_1: \pi_1 \neq \pi_2$, 即两组新生大鼠 HBV 的总体感染率不等

$\alpha = 0.05$

(2) 计算概率

在四格表周边合计数不变的条件下, 表 9-3 内 4 个实际频数变动的组合数共有“周边合计数中最小数+1”个, 即 $8+1=9$ 个, 根据公式 (9-6) 计算各种组合的四格表概率, 结果见表 9-4。例如实际观察到的四格表资料的概率为

$$P^* = \frac{9! \times 8! \times 9! \times 8!}{7! \times 2! \times 2! \times 6! \times 17!} = 0.041\ 464$$

(李康,贺佳主编, 2024, p. 92) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

表 9-4 各种组合的四格表计算的概率

四格表序号	阳性	阴性	$a - T_a$	P	四格表序号	阳性	阴性	$a - T_a$	P
1	1	8	-3.76	0.000 370	6	6	3	1.24	0.193 501
	8	0				3	5		
2	2	7	-2.76	0.011 847	7*	7	2	2.24*	0.041 464*
	7	1				2	6		
3	3	6	-1.76	0.096 750	8	8	1	3.24	0.002 962
	6	2				1	7		
4	4	5	-0.76	0.290 251	9	9	0	4.24	0.000 041
	5	3				0	8		
5	5	4	0.24	0.362 814					
	4	4							

注：* 为实际数据的四格表。

(3) 确定累计概率 P 值,作出推断结论

双侧检验:在四格表周边合计数不变的条件下, a 值的理论频数为 $T_{11}=T_a=(9 \times 9)/17=4.76$;在实

(李康,贺佳主编, 2024, p. 92) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

际观察频数 $a=7$ 时, $|a - T_a|=|7 - 4.76|=2.24$ 。观察上述 9 个 2×2 表,若拒绝 H_0 , P 值的计算应包括 $|a - T_a| \geq 2.24$ 的四格表的概率之和。双侧累计概率 P 值为

$$\begin{aligned}
 P &= P(1) + P(2) + P(7) + P(8) + P(9) \\
 &= 0.000\ 370 + 0.011\ 847 + 0.041\ 464 + 0.002\ 962 + 0.000\ 041 \\
 &= 0.056\ 684
 \end{aligned}$$

根据所得 P 值,在 $\alpha=0.05$ 检验水准下,不拒绝 H_0 ,尚不能认为预防注射组与非预防组的新生大鼠 HBV 的总体感染率不等。

(李康,贺佳主编, 2024, p. 93) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

第二节 配对四格表资料的 χ^2 检验

一句话抓住本质

McNemar 检验只问一件事：

👉 在“同一批对象”上，两种方法“改判方向”是否对称？

一、什么时候“只能”用 McNemar? (Purpose)

这是**配对设计**，不是独立样本。

典型场景

- 同一标本，用 **A / B 两种检测方法**
- 同一病人，用 **治疗前 / 治疗后**
- 同一对象，被 **两种诊断标准**判定

🔴 关键特征：

- 观察单位一一对应
- 每个对象都有“两次结果”
- 结果是**二分类**（是/否，阳/阴）

👉 一旦是配对，就不能用普通四格表 χ^2 。

二、四格表里，哪一部分才“有信息”？（核心直觉）

配对四格表的四个格子：

- a: A+, B+（一致）
- d: A-, B-（一致）
- b: A+, B-（不一致）
- c: A-, B+（不一致）

关键洞察（非常重要）：

是否有差别，
不取决于“都判对多少”，
而取决于“谁更常把结果改成阳性”。

也就是说：

- a, d: 没有区分信息
- 只有 b, c 才反映两法差异

👉 McNemar 只看不一致对 (discordant pairs) 。

三、原假设在说什么？ (Hypothesis)

当两种方法没有差别时：

$$H_0: B = C (\pi_1 = \pi_2)$$

翻译成人话：

A 判阳、B 判阴
和
A 判阴、B 判阳
出现的概率应该一样。

四、检验统计量在干什么？ (别被公式骗了)

未校正 ($b+c \geq 40$)

$$\chi^2 = \frac{(b-c)^2}{b+c}, v=1$$

直觉解释：

不一致的总次数是 $b + c$,
在这些“分歧”中,
偏向哪一边更多?

- $b = c \rightarrow$ 没偏向 $\rightarrow \chi^2 = 0$
- 差得越多 $\rightarrow \chi^2$ 越大

连续性校正 ($b+c < 40$)

$$\chi_c^2 = \frac{(|b - c| - 1)^2}{b + c}, v = 1$$

✦ 含义和 Yates 校正一致:

当不一致对本来就少时,
别因为离散性而“虚假显著”。

五、为什么自由度永远是 1? (容易忽略)

在 McNemar 检验中:

- 真正参与检验的只有 b, c
- 且在 $b + c$ 固定时, 只剩 **1 个自由变化量**

👉 自由度固定为 $v = 1$ 。

六、什么时候用哪个公式? (操作规则)

- $b + c \geq 40$
👉 用未校正公式
- $b + c < 40$
👉 用连续性校正公式

🔴 注意：

这里的“40”不是总样本量 n ，
而是 **不一致对的数量** $b + c$ 。

七、一个非常重要的“现实提醒”（教材这段很关键）

McNemar 可能“显著，但没啥用”。

为什么？

- 它**完全忽略**：
 - 总样本量 n
 - 一致对 a, d

所以当：

- n 很大
- a, d 极大（两法一致率极高）

👉 即使 $b \neq c$ 达到统计学显著，

👉 **实际差异可能微不足道。**

🔴 **统计显著 \neq 实际重要**

这一节是教你第一次正面面对这件事。

八、和前面所有 χ^2 方法的关系（一次打通）

独立四格表 χ^2 ：

两组不同对象，比较“构成比”

Fisher：

小样本独立四格表，算精确概率

McNemar:
同一对象两次判定，只看“改判方向”

九、考试级一句话总结（直接可写）

McNemar 检验适用于配对设计的四格表计数资料，通过比较两种方法在不一致对中的分布是否对称，检验两种处理或检测方法的总体率是否存在差异，其检验统计量在原假设成立时近似服从自由度为 1 的 χ^2 分布。

给你一个“不会用错”的终极锚点

是否配对，是第一判断。
配对 → **McNemar**
不配对 → **普通 χ^2 / Fisher**

【例 9-4】现有 198 份痰标本，每份标本分别用 A 和 B 两种培养基培养结核菌，结果见表 9-5。问 A 和 B 两种培养基的阳性培养率是否不等？

表 9-5 两种培养基培养结核菌结果

A 培养基	B 培养基		合计
	+	-	
+	48 (a)	24 (b)	72
-	20 (c)	106 (d)	126
合计	68	130	198

(李康,贺佳主编, 2024, p. 93) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

本例的检验步骤如下：

(1) 建立检验假设并确定检验水准

$H_0: B=C$, 即两种培养基的总体阳性培养率相等

$H_1: B \neq C$, 即两种培养基的总体阳性培养率不相等

$\alpha = 0.05$

(2) 计算检验统计量

本例 $b+c > 40$, 用公式 (9-7) 计算得

$$\chi^2 = \frac{(24-20)^2}{24+20} = 0.364, \quad \nu = 1$$

(3) 确定 P 值, 作出推断结论

查 χ^2 分布界值表 (附表 7) 得 $P > 0.05$, 按 $\alpha = 0.05$ 水准, 不拒绝 H_0 , 尚不能认为两种培养基的阳性培养率不同。

需要注意: 配对设计的四格表资料只能用配对 χ^2 检验, 而不能随意转化为两组独立样本的 χ^2 检验。

(李康, 贺佳主编, 2024, p. 94) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

第三节 $R \times C$ 列联表资料的 χ^2 检验

一、 $R \times C$ 列联表的 χ^2 检验 (直觉优先版)

一句话抓住本质

$R \times C$ χ^2 检验仍然只做一件事:

👉 实际分布, 是否偏离“行列相互独立”这一原假设。

1. 它在检验什么? (Purpose)

当分类变量 **不再是 2×2** , 而是:

- R 行 \times C 列
- 多个类别 \times 多个类别

我们关心的是:

两个分类变量之间是否有关联？

对应的原假设是：

H_0 :行变量与列变量相互独立

✦ 本质仍然是“构成比是否相同 / 是否独立”

只是类别数更多了。

2.为什么原理和四格表完全一样？（Core intuition）

不管是 2×2 ，还是 $R \times C$ ：

- 都有 **实际频数** A_{ij}
- 都有 **理论频数** T_{ij}

而理论频数的世界观始终不变：

$$T_{ij} = \frac{n_{i+}n_{+j}}{n}$$

如果行列独立，

那“这一行的比例 \times 这一列的比例”

就应该决定这个格子里有多少人。

3. χ^2 的直觉定义没有变

χ^2 = 所有格子里

“实际 - 理论” 的标准化偏离之和

只是现在：

- 格子从 4 个
 - 变成了 $R \times C$ 个
-

4.为什么要用“专用公式”？（别被公式吓到）

你给出的公式是：

$$\chi^2 = n(\sum \frac{A_{ij}^2}{n_{i+}n_{+j}} - 1)$$

这不是新思想，只是代数压缩

它等价于：

$$\sum \frac{(A_{ij} - T_{ij})^2}{T_{ij}}$$

✦ 意义只有一个：

不用逐格算 T_{ij} ，
直接用行合计、列合计和实际频数。

👉 省算力，不改逻辑。

5.自由度为什么是 $(R-1)(C-1)$ ？（一定要懂）

自由度来源于两个约束：

- 行合计固定 → 少 R 个自由度
- 列合计固定 → 少 C 个自由度
- 总合计固定 → 再少 1 个

最终剩下：

$$v = (R - 1)(C - 1)$$

✦ 记忆锚点：

每多一行、多一列，
就多一层“可以偏离独立”的空间。

6. χ^2 分布图想告诉你什么？（别忽略）

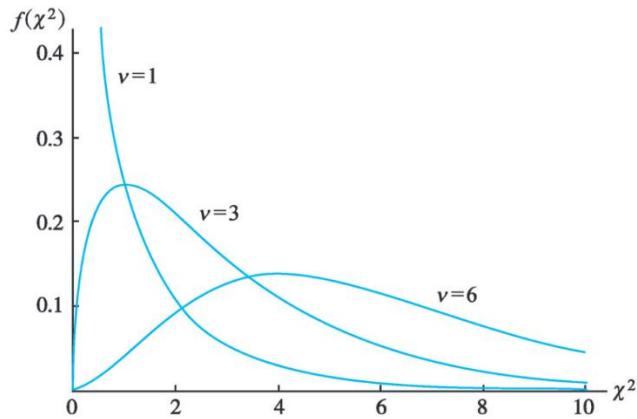


图 9-2 不同自由度的 χ^2 分布曲线图

(李康,贺佳主编, 2024, p. 94) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

图 9-2 画不同自由度的 χ^2 分布，其实在强调一句话：

**自由度越大，
 χ^2 分布越向右展开，
同样的 χ^2 值越“不显著”。**

👉 所以：

- 不能跨自由度直接比 χ^2 值大小
- 必须结合 v 和 P 值

7. 什么时候用 $R \times C \chi^2$?（和前面一秒区分）

- ☒ 两个分类变量
- ☒ 多水平 (>2)
- ☒ 各格理论频数不过小（通常 ≥ 5 ）

👉 $R \times C \chi^2$

如果：

- 有很多小频数
👉 考虑 **合并类别** 或 **Fisher 扩展法**
-

8.和前面所有方法的关系（一次封箱）

四格表 χ^2 ：

R=2, C=2 的特例

R×C χ^2 ：

四格表的自然推广

McNemar：

配对数据，不能用

Fisher：

小样本独立数据，算精确概率

9.考试级一句话总结（直接可写）

R×C 列联表 χ^2 检验通过比较各格实际频数与在行列独立假设下的理论频数之间的偏离程度，检验两个分类变量之间是否存在统计学关联，其检验统计量在原假设成立时近似服从自由度为 $(R - 1)(C - 1)$ 的 χ^2 分布。

给你一个“不会乱用”的终极判断句

只要是：多分类 × 多分类，
独立样本，问“有没有关联”
——第一反应就是 R×C χ^2 。

【例 9-5】某医院用三种方案治疗急性无黄疸型病毒性肝炎 254 例,观察结果见表 9-6,问三种方法治疗急性肝炎的有效率是否不同?

表 9-6 三种方案治疗急性肝炎的效果

组别	例数	有效	无效	有效率/%
西药组	100	51	49	51.00
中药组	80	35	45	43.75
中西药结合组	74	59	15	79.73
合计	254	145	109	57.09

(李康,贺佳主编, 2024, p. 94) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

本例为三个样本率的比较,是 3×2 列联表资料,具体的检验过程如下:

(1) 建立检验假设并确定检验水准

H_0 : 三种治疗方案的总体有效率相等

H_1 : 三种治疗方案的总体有效率不全相等

$\alpha = 0.05$

(2) 计算检验统计量

按公式 (9-9) 计算 χ^2 值

$$\begin{aligned}
 \chi^2 &= 254 \times \left(\frac{51^2}{100 \times 145} + \frac{49^2}{100 \times 109} + \frac{35^2}{80 \times 145} + \frac{45^2}{80 \times 109} + \frac{59^2}{74 \times 145} + \frac{15^2}{74 \times 109} - 1 \right) \\
 &= 254 \times (0.1794 + 0.2203 + 0.1056 + 0.2322 + 0.3244 + 0.0279 - 1) \\
 &\approx 22.809 \\
 \nu &= (3 - 1) \times (2 - 1) = 2
 \end{aligned}$$

(3) 确定 P 值,作出推断结论

查 χ^2 界值表(附表 7)得 $P < 0.05$, 在 $\alpha = 0.05$ 的检验水准下,拒绝 H_0 , 接受 H_1 , 可以认为三种疗法的有效率有差别。

(李康,贺佳主编, 2024, p. 95) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

【例 9-6】某研究人员收集了亚洲、欧洲和北美洲人的 A、B、AB、O 血型资料,结果见表 9-7,问不同地区人群的血型分布(构成比)是否不同?

表 9-7 世界三个不同地区血型样本的频数分布

地区	例数	A	B	AB	O
亚洲	1 080	321	369	95	295
欧洲	517	258	43	22	194
北美洲	995	408	106	37	444
合计	2 592	987	518	154	933

本例为三个样本构成比的比较,是 3×4 列联表资料。检验过程如下:

(1) 建立检验假设并确定检验水准

H_0 : 不同地区人群血型分布总体构成比相同

H_1 : 不同地区人群血型分布总体构成比不全相同

$\alpha = 0.05$

(2) 计算检验统计量

按公式(9-9)计算 χ^2 值为

$$\chi^2 = 2\,592 \times \left(\frac{321^2}{1\,080 \times 987} + \frac{369^2}{1\,080 \times 518} + \cdots + \frac{444^2}{995 \times 933} - 1 \right) = 297.375$$
$$\nu = (3 - 1) \times (4 - 1) = 6$$

(3) 确定 P 值,作出推断结论

查附表 7 得 $\chi_{0.05,6}^2 = 12.59$, $\chi^2 > 12.59$, $P < 0.05$, 在 $\alpha = 0.05$ 检验水准下,拒绝 H_0 , 可以认为三个不同地区的人群血型分布不全相同。

(李康,贺佳主编, 2024, p. 95) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

二、多个样本率间的多重比较

一句话抓住本质

总体 χ^2 只回答“有没有差别”,

多重比较才回答“差别发生在哪”,

而 Bonferroni 是为“别乱判显著”付出的代价。

1.为什么 χ^2 显著之后还不够? (Purpose)

当你对 k 个样本率 做整体 χ^2 检验并拒绝 H_0 :

$$H_0: \pi_1 = \pi_2 = \cdots = \pi_k$$

你真正知道的只有:

至少有一个总体率不同。

但你不知道:

- 是哪两个不同
- 是不是只有一个“异类”
- 还是很多对都不同

👉 必须做两两比较 (post-hoc) 。

2.为什么不能直接做很多次四格表 χ^2 ? (核心风险)

逻辑和前面 t 检验 / ANOVA 完全一样:

- 每做一次 χ^2 比较
- 都有一次犯 I 类错误的风险

当比较次数一多:

总体犯 I 类错误的概率会迅速膨胀。

✦ 结论和态度要非常明确:

**多个样本率的两两比较,
不能直接反复用普通 χ^2 。**

3.解决思路: χ^2 分割 + Bonferroni (核心思想)

教材这节的关键词是：

χ^2 分割法 + Bonferroni 校正

核心直觉只有一句话：

把“总体比较”拆成多个“两两比较”，
同时把每一次检验的 α 压低，
以保证“整个实验”的 α 不变。

👉 这是一个保守但安全的策略。

4. Bonferroni 在这里到底干了什么？ (Very important)

Bonferroni 的思想你已经见过很多次，本质不变：

α 是“总预算”，
比较次数越多，
单次能用的 α 就越少。

5. 两种最常见的应用场景（一定要分清）

情况 1：k 个实验组，两两都要比

问题形式：

“任意两个样本率是否不同？”

- 两两比较次数：

$$m = \binom{k}{2} = \frac{k(k-1)}{2}$$

- 调整后的检验水准：

$$\alpha' = \frac{\alpha}{\binom{k}{2}}$$

✦ 特点:

- 比较次数多
 - α' 很小
 - **非常保守**
 - 组数不宜太多
-

情况 2：多个实验组 vs 同一个对照组

问题形式:

“各实验组是否不同于对照组?”

- 比较次数:

$$m = k - 1$$

- 调整后的检验水准:

$$\alpha' = \frac{\alpha}{k - 1}$$

✦ 特点:

- 不比较实验组彼此
 - 功效明显高于“全两两比较”
 - 和 **Dunnett 思想**在本质上是一致的
-

6.每一次“两两比较”具体怎么做? (别混)

在 χ^2 分割法中:

- **每一对样本率**
- 仍然构成一个 **四格表**
- 使用:
 - 普通 χ^2
 - 或 Yates 校正 χ^2

- 或 Fisher (小样本)

✦ 唯一变化的是:

判断显著性时, 用的是 α' , 不是 α

7. 这是一种“什么风格”的方法? (一定要有判断)

χ^2 分割 + Bonferroni:

- 思想简单
- 逻辑透明
- **非常保守**
- 容易“错过真实差异”

所以教材特别强调:

比较的组数不宜过多。

8. 和前面方法的关系, 一句话打通

ANOVA + 多重比较 (均数)

↓

χ^2 总体验证 + χ^2 分割 (率)

逻辑完全平行, 只是对象不同:

- 一个是 **均数**
 - 一个是 **率 / 构成比**
-

9. 考试级一句话总结 (直接可写)

当多个样本率的总体 χ^2 检验拒绝原假设后，为明确不同总体率之间的具体差异，可采用 χ^2 分割法进行两两比较，并通过 Bonferroni 法调整检验水准，以控制多重比较中 I 类错误的发生概率。

给你一个“不会乱用”的终极判断句

先问：是不是只和对照组比？

是 $\rightarrow \alpha' = \alpha / (k - 1)$

否 $\rightarrow \alpha' = \alpha / \binom{k}{2}$

【例 9-7】对例 9-5 中的三种治疗方法进行两两比较，以推断是否任意两种疗法治疗急性无黄疸型病毒性肝炎的有效率均有差别？

(李康, 贺佳主编, 2024, p. 96) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

本研究目的为三个实验组间的两两比较,检验水准 α' 用公式(9-10)估计得

$$\alpha' = \frac{0.05}{3 \times (3-1)/2} = \frac{0.05}{3} = 0.017$$

(1) 建立检验假设并确定检验水准

$H_0: \pi_A = \pi_B$, 即任两对比组的总体有效率相等

$H_1: \pi_A \neq \pi_B$, 即任两对比组的总体有效率不等

$\alpha = 0.05$

(2) 计算检验统计量

本研究目的为三个实验组间的两两比较,将其分为三个 2×2 表并分别计算出 χ^2 值,结果见表 9-8。

(3) 确定 P 值,作出推断结论

查附表 7,按 $\alpha' = 0.017$ 检验水准,得到自由度 $\nu = 1$ 的 χ^2 界值在 5.02~6.63 之间,可采用线性插值法得到 $\chi^2_{0.017,1} \approx 5.88$,任意两组比较的 P 值见表 9-8。结果显示:西药组与中药组相比,不拒绝 H_0 ,尚不能认为两组间有效率不同;而中西药结合组与其他两组相比,均拒绝 H_0 ,接受 H_1 ,可认为中西药结合组与其他两组相比有效率不同。结论:中西药结合治疗急性无黄疸型病毒性肝炎具有更好的疗效。

表 9-8 三种疗法治疗肝炎有效率的两两比较

对比组	有效	无效	合计	χ^2	P
西药组	51	49	100	0.936 < 5.88	> 0.017
中药组	35	45	80		
合计	86	94	180		
中药组	35	45	80	20.926 > 5.88	< 0.017
中西药结合组	59	15	74		
合计	94	60	154		
西药组	51	49	100	15.096 > 5.88	< 0.017
中西药结合组	59	15	74		
合计	110	64	174		

S

(李康,贺佳主编, 2024, p. 96) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

【例 9-8】以例 9-5 中的西药治疗组为对照组,中药治疗组与中西药结合为试验组,试分析两试验组与对照组的总体有效率有无差别?

本研究目的为各试验组与同一对照组的比较,检验水准 α' 用式 (9-11) 估计得

$$\alpha' = \frac{0.05}{3-1} = 0.025$$

(1) 建立检验假设并确定检验水准

$H_0: \pi_T = \pi_C$, 即各试验组与对照组的总体有效率相等

$H_1: \pi_T \neq \pi_C$, 即各试验组与对照组的总体有效率不等

$\alpha = 0.05$

(2) 计算检验统计量

本例为各试验组与同一对照组的比较,计算 χ^2 统计量的结果见表 9-8。

(3) 确定 P 值,作出推断结论

查附表 7, $\chi_{0.025,1}^2 = 5.02$, 按 $\alpha' = 0.025$ 水准,中药组与西药组相比较, $0.936 < 5.02$, 则 $P > 0.025$, 不拒绝 H_0 , 两组疗效无显著差异; 中西药结合组与西药组相比, $15.096 > 5.02$, 则 $P < 0.025$, 拒绝 H_0 , 两组疗效的差异具有统计学意义, 说明中西药结合治疗急性无黄疸型病毒性肝炎的疗法优于单纯用西药的疗法。

(李康, 贺佳主编, 2024, p. 97) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

三、 $R \times C$ 列联表 χ^2 检验的注意事项 (直觉优先版)

一句话总览

$R \times C \chi^2$ 好用, 但前提是:

数据够大、问题问得对、方法没用错。

1 理论频数条件不是形式, 而是底线

标准经验规则 (一定要记)

- 任一格理论频数 $T \geq 1$
- $1 \leq T < 5$ 的格子数 \leq 总格子数的 $1/5$

✦ 这不是“严格数学定理”, 而是保证 χ^2 近似可信的经验下限。

如果不满足，怎么办？（三条路）

① 增加样本量

- 最干净
- 但现实中往往最难

② 合并或删除类别（基于专业判断）

- 例如把“极少见分型”合并
- ⚠️ 必须有**专业合理性**，不能纯为过检验

③ 改用 Fisher 确切概率法 ($R \times C$ 扩展)

- 当数据稀疏、样本小
 - 不要硬上 χ^2
-

2 总体显著 \neq 任意两两显著

当 $R \times C$ χ^2 检验拒绝 H_0 时，你**只能得到一句话**：

各总体率“总体上”存在差别。

但你**不能说**：

- 任意两组都不同
- 哪两组不同

👉 **下一步必须做：多个样本率的多重比较**

(χ^2 分割 + Bonferroni，上一节你已经完全掌握)

3 χ^2 检验不看“顺序”，只看“分类”

这是一个非常容易被忽略的本质点。

χ^2 的世界观是：

类别是“并列的、无序的”。

所以：

- $R \times C \chi^2$
 - 👉 只比较 构成比是否不同
 - 不利用类别之间的顺序信息
-

什么时候不该用 χ^2 ? (很关键)

如果你的分类是**有序的**，例如：

- 痊愈 / 显效 / 有效 / 无效
- 轻 / 中 / 重
- 优 / 良 / 中 / 差

而你的研究目的其实是：

比较“疗效高低”“程度变化”

👉 χ^2 会浪费顺序信息

👉 不推荐

📌 正确选择：

- 非参数秩和检验 (如秩和、趋势检验)
 - 这是下一章的内容
-

💡 关联性检验： χ^2 的“另一种用法”

最后这一点是在帮你**统一世界观**。

对于交叉分类数据：

- 是否吸烟 × 是否饮酒
- 是否暴露 × 是否发病
- 性别 × 疾病类型

📌 本质问题是：

两个分类变量是否独立？

重要结论（一定要记）

不论是“多样本率比较”，
还是“两个分类变量的关联性检验”，
 $R \times C \chi^2$ 的计算公式是完全一样的。

- 原理：
👉 独立条件下概率的乘法原则
 - 表现形式：
👉 实际频数 vs 理论频数
-

🧠 把这一节压缩成 4 句“判断口诀”

- 1 小频数 → 先别算 χ^2
 - 2 总体显著 \neq 两两显著
 - 3 有顺序 → χ^2 不合适
 - 4 交叉分类 → 仍然是 χ^2 在干活
-

📌 考试级一句话总结（可直接写）

在 $R \times C$ 列联表 χ^2 检验中，应注意理论频数条件的限制、总体检验与两两比较的区别以及 χ^2 检验对分类顺序不敏感的特点。当数据稀疏或分析目的涉及有序分类时，应考虑采用其他统计方法。

最后给你一个方法选择的终极锚点

χ^2 检验解决的是：

“分类 \times 分类，有没有关联 / 构成比是否不同”。

一旦问题变成“程度、趋势、顺序”，

χ^2 就该退场。

[BIBLIOGRAPHY] Please click Zotero - Refresh in Word/LibreOffice to update all fields