卡方检验

秦雪英
北京大学公共卫生学院
流行病与卫生统计学系

案例讨论

北京大学 PEKING UNIVERSITY

- 实例1:
- 有研究者对女性乳腺癌与首次生育年龄的关联感兴趣，并开展了一项病例对照研究。研究者从选定的一些医院中选了乳腺癌病例，并从与病例同时住院但未患乳腺癌的年龄相当的女性中随机选择对照。所有被调查女性都被问及她们的生育史和首次生育年龄。研究者将至少有过一次生育经历的研究对象按照首次生育年龄是否大于29岁分为两组：（1）首次生育年龄≤29岁；（2）首次生育年龄>29岁。结果发现：3220 名患有乳腺癌的女性中有 683 名（21.2%）其首次生育年龄>29岁，与此同时，10245 名未患乳腺癌的女性中有 1498 名（14.6%）其首次生育年龄>29岁。请问，乳腺癌患者和非患者其首次生育年龄的构成是否有差异?
- 请回答：

1. 这个研究的资料（或变量）属于什么类型的资料?
2. 请选择合适的假设检验方法?

数据来源： *Rosner, Bernard (Bernard A.). Fundamentals of Biostatistics. Boston :Brooks/Cole, Cengage Learning, 2011.*
MacMahon, B., Cole, P., Lin, T. M., Lowe, C. R., Mirra, A. P., Ravnihar, B., Salber, E. J., Valaoras, V. G., & Yuasa, S. (1970). Age at first birth and breast cancer risk. Bulletin of the World Health Organization, 43, 209–221.

复习

- 变量类型：**计量（定量的）**和**计数（定性的或分类的）**
  - 计量变量：用数字进行测量，可以直接进行加减乘除运算 （又分为离散型数值变量和连续型数值变量），一般带有度量衡单位。
  - 计数变量：对研究特征进行分类，不可以直接进行加减乘除运算（又分为无序和有序的分类变量），一般无度量衡单位。

  - 3220 名患有乳腺癌的女性中有 **683 名（21.2%）**其首次生育年龄>29岁，与此同时，10245 名未患乳腺癌的女性中有 **1498 名（14.6%）**其首次生育年龄>29岁。

✓ **计数资料**

✓ **构成比：本例中：事件（年龄分组）的结果只有两种可能性，因此服从二项分布，这里的构成比又称为 binomial proportions**

复习

- 迄今已学的假设检验的方法
  - t检验
  - 方差分析
  - 秩和检验

- 这些检验方法的前提假设？它们是否适用于本资料类型？
  - 计量资料
  - 比较的是均数或中位数
  - 假设的总体分布
    - t检验：正态分布
    - 方差分析：正态分布
    - 秩和检验：偏态分布或总体分布未知

案例讨论

- 假设：乳腺癌患者的首次生育年龄>29岁的总体构成比为π_1，非乳腺癌患者的首次生育年龄>29岁的总体构成比为π_2。
- H_0: $\pi_1 = \pi_2$
- H_1: $\pi_1 \neq \pi_2$

- 对于两个独立样本的总体率（或构成比）的假设检验：
  - 大样本情形下，利用二项分布正态近似的原理（中心极限定理），可以采用z检验（或u检验），比较两个样本的总体率的差值是否偏离0。
  - 列联表的卡方检验

卡方检验概述

北京大学 PEKING UNIVERSITY

- 分类资料的统计推断，在实际工作中往往更多地应用卡方（χ^2）检验（chi-square test）。
- χ^2检验是以χ^2分布为理论依据，用途很广的假设检验方法，包括：
    - 两个总体率或构成比之间有无差别
    - 多个总体率或构成比之间有无差别
    - 两个分类变量之间有无关联性
    - 频数分布的拟合优度的χ^2检验

主要内容

- χ^2检验的基本思想
- 独立四格表资料的χ^2检验
- 配对四格表资料的χ^2检验
- 行×列表资料的χ^2检验
- 独立性检验
- 频数分布拟合优度的χ^2检验

主要内容

- χ^2检验的基本思想
- 独立四格表资料的χ^2检验
- 配对四格表资料的χ^2检验
- 行×列表资料的χ^2检验
- 独立性检验
- 频数分布拟合优度的χ^2检验

重点掌握

一般了解

χ^2检验的基本思想

- 历史
- 在十九世纪，统计分析方法主要被用于生物数据分析。当时主流意见认为正态分布普遍适用于此类数据，而Pearson在他1900年的论文中就针对了他们的研究数据作出了指正。
- 直到十九世纪末期， Pearson指出了部分数据具有明显的偏态，正态分布并不是普遍适用。为了更好地对这些观察数据进行建模， Pearson在1893年至1916年发表的系列文章中提出了一个包含正态分布以及众多偏态分布的连续概率分布族—Pearson distribution。同时，他指出数据统计分析的步骤应该是在从Pearson分布族中选取合适的分布来进行建模后，使用拟合优度检验技术来评价模型和实验数据间的拟合优度。
- 1900年，Pearson发表了著名的关于χ^2检验的文章，该文章被认为是现代统计学的基石之一。在该文章中，Pearson研究了拟合优度检验。

https://en.wikipedia.org/wiki/Chi-squared_test

Karl Pearson
FRS

Pearson in 1912

Born	Carl Pearson 27 March 1857 Islington, London, England
Died	27 April 1936 (aged 79) Coldharbour, Surrey, England
Alma mater	King's College, Cambridge University of Heidelberg
Known for	Principal component analysis Pearson distribution Pearson's chi-squared test Pearson's r Phi coefficient Chi-square distribution Contingency table Histogram Kurtosis Mode Random walk *The Grammar of Science*

χ^2检验的基本思想

- 例1: 在一项病例对照研究中，3220 名患有乳腺癌的女性中有 683 名（21.2%）其首次生育年龄>29岁，与此同时，10245 名未患乳腺癌的女性中有 1498 名（14.6%）其首次生育年龄>29岁。请问，乳腺癌患者和非患者其首次生育年龄的构成是否有差异?

- 两个变量：是否患乳腺癌，首次生育年龄是否>29岁。可将研究对象分为四组。

表1：乳腺癌病例组和对照组的首次生育年龄构成比

	首次生育年龄		合计
	>29岁	≤29岁	
乳腺癌病例	**683**	**2537**	3220
乳腺癌对照	**1498**	**8747**	10245
合计	2181	11284	13465

χ^2检验的基本思想

- 行列表，又称列联表（contingency table）或交叉表（cross tabulation table）。
- 由分组(组别)变量和结果(结局)变量的R（row）和C（column）个分类组合形成。
- 有$R \times C$个格子（cell）。
- 行数R=2，列数C=2的列联表又称四格表（two by two table/ 2×2 table）。

	首次生育年龄		
	>29岁	≤29岁	合计
乳腺癌病例	683	2537	3220
乳腺癌对照	1498	8747	10245
合计	2181	11284	13465

行合计
row margins

cells，单元格

列合计
column margins

观察频数或实际频数
（observed frequency）

χ^2检验的基本思想

	首次生育年龄		合计
	>29岁	≤29岁	
乳腺癌病例	**683**	**2537**	3220
乳腺癌对照	**1498**	**8747**	10245
合计	2181	11284	13465

- 若要比较乳腺癌患者和非患者其首次生育年龄的构成是否有差异，还需要引入**理论频数（或期望频数）**的概念。
- 理论频数（expected frequency）：是根据H_0: $\pi_1 = \pi_2$确定的。无效假设下，乳腺癌病例组和对照组的首次生育年龄的总体构成比相等，最理想的总体构成比应等于合计的构成比。
- **首次生育年龄>29岁的合计构成比**：$2181/13465 = 16.2\%$
- H_0成立的条件下，
  - **乳腺癌病例的首次生育年龄>29岁的构成比**为16.2%，那么**乳腺癌病例首次生育年龄>29岁的理论频数**：$3220 \times 16.2\% = 521.6$
  - **乳腺癌对照的首次生育年龄>29岁的构成比**为16.2%，那么**乳腺癌对照首次生育年龄>29岁的理论频数**：$10245 \times 16.2\% = 2698.4$

χ^2检验的基本思想

实际频数

	首次生育年龄 >29岁	首次生育年龄 ≤29岁	合计
乳腺癌病例	**x_1**	**n_1-x_1**	n_1
乳腺癌对照	**x_2**	**n_2-x_2**	n_2
合计	x_1+x_2	$(n_1+n_2)-(x_1+x_2)$	n_1+n_2

- 首次生育年龄>29岁的合计构成比: $(x_1+x_2)/(n_1+n_2)$
- H_0成立的条件下，乳腺癌病例首次生育年龄>29岁的理论频数（或称期望频数）：$n_1 \times (x_1+x_2)/(n_1+n_2)$ [**行合计与列合计之积除以总频数**]

理论频数

	首次生育年龄 >29岁	首次生育年龄 ≤29岁	合计
乳腺癌病例	$n_1 \times (x_1+x_2)/(n_1+n_2)$	$n_1 \times [(n_1+n_2)-(x_1+x_2)]/(n_1+n_2)$	n_1
乳腺癌对照	$n_2 \times (x_1+x_2)/(n_1+n_2)$	$n_2 \times [(n_1+n_2)-(x_1+x_2)]/(n_1+n_2)$	n_2
合计	x_1+x_2	$(n_1+n_2)-(x_1+x_2)$	n_1+n_2

χ^2检验的基本思想

实际频数

首次生育年龄	>29岁	≤29岁	合计
乳腺癌病例	683	2537	3220
乳腺癌对照	1498	8747	10245
合计	2181	11284	13465

理论频数

首次生育年龄	>29岁	≤29岁	合计
乳腺癌病例	521.6	2698.4	3220
乳腺癌对照	1659.4	8585.6	10245
合计	2181	11284	13465

实际*频率**

首次生育年龄	>29岁	≤29岁	合计
乳腺癌病例	21.2%	78.8%	100.0%
乳腺癌对照	14.6%	85.4%	100.0%
合计	16.2%	83.8%	100.0%

理论*频率**

首次生育年龄	>29岁	≤29岁	合计
乳腺癌病例	16.2%	83.8%	100.0%
乳腺癌对照	16.2%	83.8%	100.0%
合计	16.2%	83.8%	100.0%

* 注意：这里的频率是指行的频率，此外还可以计算列的频率，和每个格子的频率。

χ^2检验的基本思想

北京大學 PEKING UNIVERSITY

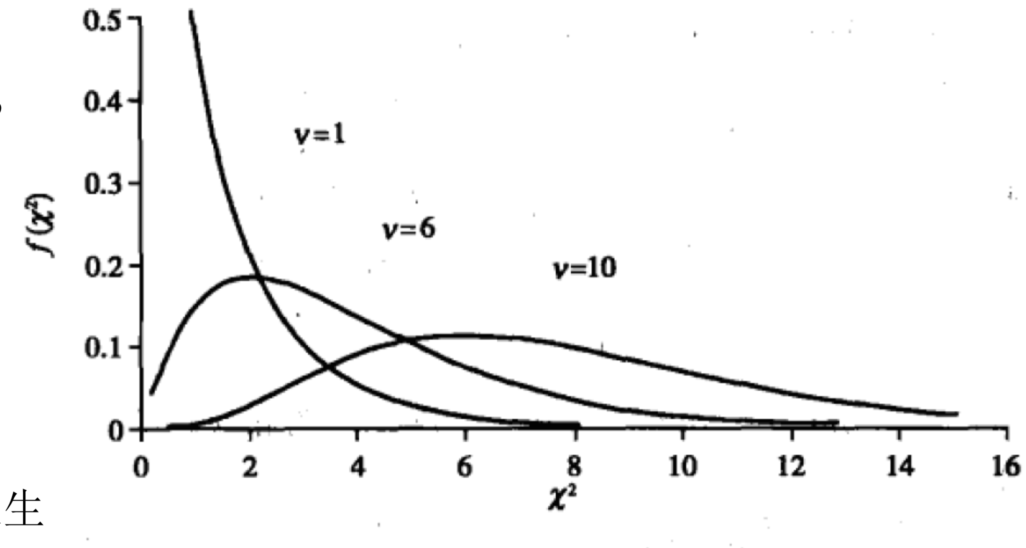
- 比较乳腺癌患者和非患者的首次生育年龄的构成比是否有差异，则比较四格表里的理论频数和实际频数：
- H_0成立前提下，实际频数应与对应的理论频数越接近越好，如果越接近，则接受H_0，拒绝H_1。如果差别越大，则拒绝H_0。
- 如何判断差异多大就可以拒绝H_0呢?
- 统计学家证明，在H_0成立前提下，$\sum\frac{(O-E)^2}{E}$服从自由度为$(R-1)(C-1)$的χ^2分布，即：

$$\chi^2=\sum\frac{(O-E)^2}{E}, \nu=(R-1)(C-1)$$

- 其中，O为实际频数，E为理论频数。 ν是自由度。如果得到的χ^2值超过自由度下χ^2分布的显著性水平对应的χ^2临界值，则拒绝H_0。

χ^2分布

- χ^2分布
  - χ^2分布是一种连续型分布：只有一个参数，即自由度ν。
  - χ^2分布的形状依赖于自由度的大小，当$\nu \leq 2$时，曲线呈L形，随着ν增加，曲线逐渐趋于对称。当自由度为∞时，χ^2分布 趋于正态分布。
  - χ^2分布具有可加性。两个独立的随机变量X_1和X_2分别服从自由度为ν_1和ν_2的χ^2分布，那么它们的和（X_1+X_2）服从自由度$\nu_1+\nu_2$的χ^2分布。
  - 当自由度确定后，χ^2分布曲线下右侧尾部的面积为α时，相应的χ^2值记为$\chi^2_{\alpha,\nu}$或$\chi^2_{\alpha}(\nu)$，即χ^2分布的界值。
  - 自由度为1，$\chi^2_{0.05}(1) = 3.84$。它是自由度为1时的χ^2分布的一个临界值，在假设检验中常用。



不同自由度的χ^2分布曲线图

图来源：孙振球，徐勇勇主编。医学统计学（第三版），北京，人民卫生出版社，2010年。

χ^2分布与标准正态分布的关系

- 标准正态变量Z的平方服从自由度为1的χ^2分布。即：

  $Z^2=\chi^2(1)$

- 样本均数$\bar{Y}$标准化后的平方服从自由度为1的χ^2分布
- 设X_1,X_2分别是自由度为n_1和n_2的两个独立的χ^2随机变量，则$X=X_1+X_2$服从自由度为$n=n_1+n_2$的χ^2分布。
- 设$Y_1\ldots\ldots,Y_n$是n个均值为μ，标准差为σ的独立正态分布变量的样本，$\bar{Y}$是样本均数，则比值

$$\frac{\sum_{t=1}^{n}(Y_t-\bar{Y})^2}{\sigma^2}=\chi^2$$

  服从自由度为n−1的χ^2分布。

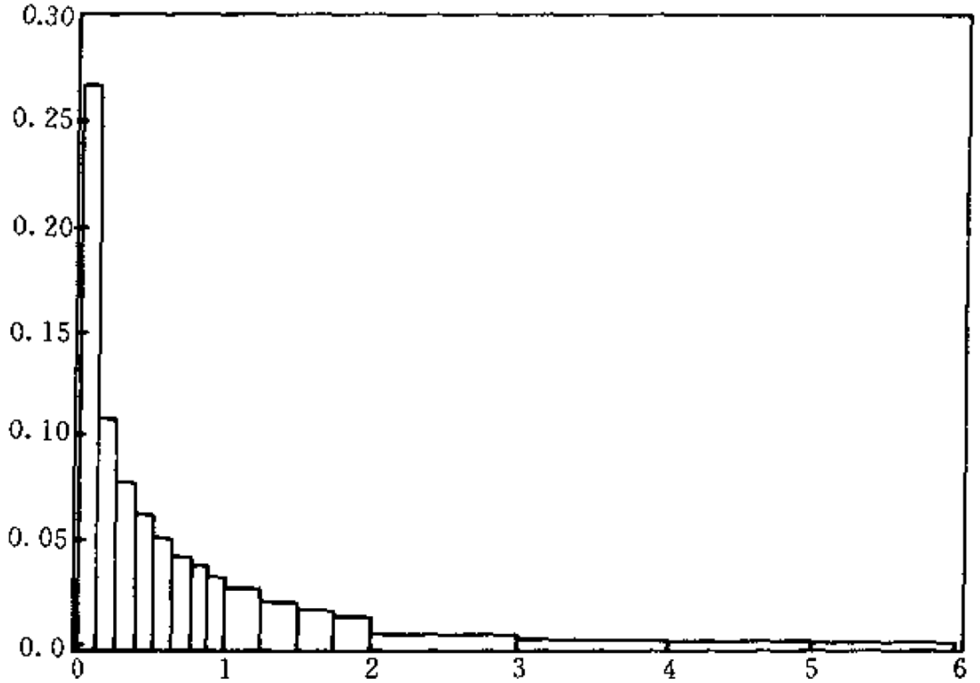


图 8.2 Z^2分布

来源：蒋庆琅原著，方积乾等译。实用统计分析方法，北京，北京医科大学、中国协和医科大学联合出版社，1998年。

独立四格表资料的χ^2检验

- 独立四格表资料：
  - 又称为成组设计的四格表资料
  - 是进行**两个独立样本**的总体率或总体构成比的比较的假设检验

  - 假设检验的基本步骤?

独立四格表资料的χ^2检验

1. 建立检验假设，确定检验水准
   - H_0: $\pi_1 = \pi_2 = \pi$，即乳腺癌病例组与对照组首次生育年龄的总体构成比相等
   - H_1: $\pi_1 \neq \pi_2$，即乳腺癌病例组与对照组首次生育年龄的总体构成比有差别
   - $\alpha = 0.05$
2. 计算检验统计量
   - $E_{11} = 521.6, E_{12} = 2698.4, E_{21} = 1659.4, E_{22} = 8585.6$
   - 根据$\chi^2 = \sum \frac{(O-E)^2}{E}$，计算得：
   - $\chi^2 = \frac{(683-521.6)^2}{521.6} + \frac{(2537-2698.4)^2}{2698.4} + \frac{(1498-1659.4)^2}{1659.4} + \frac{(8747-8585.6)^2}{8585.6} = 78.37$
   - $\nu = (R-1)(C-1)=1$
3. 确定P值，作出推断结论
   - $\nu = 1$，查附表χ^2界值表，得 $\chi^2_{0.05,1} = 3.84$，本例χ^2 =78.37，按$\alpha = 0.05$水准，拒绝H_0，接受H_1，认为乳腺癌病例组与对照组首次生育年龄的总体构成比有差别，即可认为乳腺癌患者首次生育年龄>29岁组的构成比高于对照组。

独立四格表资料的χ^2检验

- 几个需注意的问题：

1. 四格表资料的χ^2检验的自由度为1
2. 四格表资料χ^2检验的专有公式
3. $\chi^2=\sum\frac{(O-E)^2}{E}$分布为连续性分布，在用于离散型的频数分布时需要进行连续性校正。因此，该公式仅在二项分布近似正态分布的条件下成立，即所有格子的期望频数均应大于5。如果不满足怎么办?

独立四格表资料的χ^2检验

- 几个需注意的问题：

1. 四格表资料的χ^2检验的自由度为1

   - 行列合计固定的情况下，只要确定了一个格子的频数，就可以通过行列合计求出其他格子的频数，因此，可以自由取值的格子数为1，自由度为1。

独立四格表资料的χ^2检验

- 几个需注意的问题：

2. 四格表资料χ^2检验的专有公式

$$\chi^2 = \frac{(ad - bc)^2 \times n}{(a+b)(c+d)(a+c)(b+d)}, \quad \nu = 1$$

其中，a, b, c, d是四格表的实际频数。

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

两公式在四格表计算中完全等价。

独立四格表资料的χ^2检验

- 几个需注意的问题：

3. $\chi^2=\sum\frac{(O-E)^2}{E}$分布为连续性分布，在用于离散型的频数分布时需要进行连续性校正。因此，该公式仅在二项分布近似正态分布的条件下成立，即所有格子的期望频数均应大于5。如果不满足怎么办？

- 连续性校正，又称Yates校正：

$$\chi_C^2=\sum\frac{(|O-E|-0.5)^2}{E}$$

四格表专有公式的校正公式：$\chi_C^2=\frac{(|ad-bc|-n/2)^2\times n}{(a+b)(c+d)(a+c)(b+d)}$，$\nu=1$

独立四格表资料的χ^2检验

- 几个需注意的问题：

3. Yates校正

- 对于小样本资料使用连续性校正较好，此时未校正的χ^2值太大，导致拒绝零假设的机会过多，当样本含量大时，连续性校正对χ^2值的影响不大。
- 实际工作中，对于四格表资料，通常规定：
  - 当**n>=40，且所有的E>=5**时，用χ^2检验的基本公式或四格表的专用公式；若$P \approx \alpha$时，用四格表资料的Fisher精确概率法。
  - 当**n>=40，但有1<=E<5**时，用χ^2检验的校正公式或四格表的专用校正公式或改用四格表资料的Fisher精确概率法
  - 当**n<40或E<1**，用四格表资料的Fisher精确概率法

独立四格表资料的χ^2检验

- 例1中：
- $\chi^2 = \frac{(|683-521.6|-0.5)^2}{521.6} + \frac{(|2537-2698.4|-0.5)^2}{2698.4} + \frac{(|1498-1659.4|-0.5)^2}{1659.4} + \frac{(|8747-8585.6|-0.5)^2}{8585.6} = 77.88$

- 不同的软件在进行卡方检验的时候，采用的公式稍有区别。
- stata，spss大样本情况下用的是基本公式，而R的chisq.test直接给出的是校正公式的结果。

独立四格表资料的χ^2检验

- 例2: 某医师欲比较包磷胆碱与神经节苷脂治疗脑血管疾病的疗效，将78例脑血管疾病患者随机分为两组，结果见下表。问两组药物治疗脑血管疾病的有效率是否相等?

表2：两组药物治疗脑血管疾病的有效率的比较

组别	有效	无效	合计	有效率（%）
包磷胆碱组	46	6	52	88.46
神经节苷脂组	18	8（**4.67**）	26	69.23
合计	64	14	78	82.05

数据来源：孙振球，徐勇勇主编。医学统计学（第三版），北京，人民卫生出版社，2010年。

独立四格表资料的χ^2检验

1. 建立检验假设，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

n=78，但有一个格子的理论频数为4.67<5，需采用四格表资料卡方检验的校正公式，计算得 χ^2=3.14.

3. 确定P值，作出推断结论

本例自由度为1， $\chi^2=3.14 < 3.84$，$P>0.05$，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。本例若不校正卡方值，χ^2=4.35 ， $P < 0.05$，将得到相反的结论。

独立四格表资料的Fisher确切概率法

- 前述χ^2检验基于二项分布近似正态分布的前提，但当样本含量较少（如四格表资料总例数$n<40$，或有期望频数$E<1$），上述前提往往不能满足，需要采用Fisher确切检验（Fisher’s exact test）进行分析。
- 该法由R.A.Fisher 于1934年提出，其理论依据是超几何分布，并非χ^2检验的范畴。
- 该方法可直接计算概率，因此也叫Fisher确切概率检验（Fisher’s exact probability test）。
- 以四格表资料为例说明这一检验方法。

独立四格表资料的Fisher确切概率法的基本思想

- 在周边合计数保持不变的条件下，表格中的实际频数a、b、c、d可有多种组合，各种组合表对应的概率可用超几何分布公式计算。
- 在四格表周边合计数不变的条件下，共有“周边合计数中最小数 + 1”种组合。
- 先求出所有组合四格表的概率，然后将所有小于等于原样本四格表概率的所有四格表概率值相加，得到双侧检验P值。
- 原样本四格表以左（包括原样本）的所有四格表概率之和为左侧概率，原样本四格表以右（包括原样本）的所有四格表概率之和为右侧概率，若备择假设为$\pi_1 > \pi_2$,单侧概率为右侧概率，若备择假设为$\pi_1 < \pi_2$,单侧概率为左侧概率。
- 按超几何分布（hypergeometric distribution）的原理，四格表的概率计算公式为:

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

！：阶乘符号

独立四格表资料的Fisher确切概率法

- 例3: 某医师为研究乙肝免疫球蛋白预防胎儿宫内感染HBV的效果，将33例HBsAg阳性孕妇随机分为预防注射组和非预防组，结果见表3。问两组新生儿的HBV的总体感染率有无差别?

表3：两组新生儿HBV感染率的比较

组别	阳性	阴性	合计	感染率（%）
预防注射组	4	18	22	18.18
非预防组	5	6	11	45.45
合计	9	24	**33**	27.27

数据来源：孙振球，徐勇勇主编。医学统计学（第三版），北京，人民卫生出版社，2010年。

独立四格表资料的Fisher确切概率法

表3：两组新生儿HBV感染率的比较

组别	阳性	阴性	合计	感染率（%）
预防注射组	4	18	22	18.18
非预防组	5	6	11	45.45
合计	9	24	**33**	27.27

从cell(1, 1)的频数为0开始，由于四格表的自由度为1，在行列合计不变的情况下，其他三个格子的频数可以由cell(1, 1)确定。P_i为cell(1, 1) 从小到大出现的四格表的概率。

保持行合计和列合计不变的情况下，四格表的所有可能组合：

(1)

0	22
9	2

$P_1=0.00000143$ $P_2=0.00009412$ $P_3=0.00197656$

(4)

	19
	5

$P_4=0.01844785$

(5)

4	18
5	6

$P_5=0.08762728$

(6)

5	17
4	7

$P_6=0.22532729$

(7)

6	16
3	8

$P_7=0.31921366$

(8)

7	15
2	9

$P_8=0.24321040$

(9)

8	14
1	10

$P_9=0.09120390$

(10)

9	13
0	11

$P_{10}=0.01289752$

若遇到四格表a+b=c+d或a+c=b+d的情况时，四格表各组合的概率呈对称分布

(1)

0	22
9	2

$P_1=0.00000143$

(2)

1	21
8	3

$P_2=0.00009412$

(3)

2	20
7	4

$P_3=0.00197656$

(4)

3	19
6	5

$P_4=0.01844785$

(5)

4	18
5	6

$P_5=0.08762728$

(6)

5	17
4	7

$P_6=0.22532729$

(7)

6	16
3	8

$P_7=0.31921366$

(8)

7	15
2	9

$P_8=0.24321040$

(9)

8	14
1	10

$P_9=0.09120390$

(10)

9	13
0	11

$P_{10}=0.01289752$

Q：如何通过这些概率来进行假设检验呢?

A：取决于单侧检验还是双侧检验。

- 如果是双侧检验，概率P为所有小于等于原样本四格表概率的所有四格表概率之和。
- 如果是右侧检验，概率P为原样本四格表以右（包括原样本）的所有四格表概率之和。
- 如果是左侧检验，概率P为原样本四格表以左（包括原样本）的所有四格表概率之和。

组别	阳性	阴性	合计	感染率（%）
预防注射组	4	18	22	18.18
非预防组	5	6	11	45.45
合计	9	24	**33**	27.27

组合	四格表	P_i
(1)	0, 22 / 9, 2	P_1=0.00000143
(2)	1, 21 / 8, 3	P_2=0.00009412
(3)	2, 20 / 7, 4	P_3=0.00197656
(4)	3, 19 / 6, 5	P_4=0.01844785
(5)	4, 18 / 5, 6	P_5=0.08762728
(6)	5, 17 / 4, 7	P_6=0.22532729
(7)	6, 16 / 3, 8	P_7=0.31921366
(8)	7, 15 / 2, 9	P_8=0.24321040
(9)	8, 14 / 1, 10	P_9=0.09120390
(10)	9, 13 / 0, 11	P_{10}=0.01289752

- 本例n=33<40, 宜用四格表资料的Fisher 精确概率法
- 1. 建立检验假设，确定检验水准
  - H_0：$\pi_1 = \pi_2$，两组新生儿HBV的总体感染率相等
  - H_1：$\pi_1 \neq \pi_2$，两组新生儿HBV的总体感染率不等
  - α =0.05
- 2. 计算现有样本四格表的概率P^*及各组合四格表的概率P_i。此例为双侧检验，因此计算满足P_i<=P^*的所有四格表的累计概率。
  - $P = P_1 + P_2 + P_3 + P_4 + \boldsymbol{P_5} + P_{10}$=0.1210

左侧概率：$P_1 + P_2 + P_3 + P_4 + \boldsymbol{P_5}$

右侧概率：$\boldsymbol{P_5} + P_6 + P_7 + P_8 + P_9 + P_{10}$

若是单侧检验，取单侧概率值

- 3. 确定P值，作出推断结论
  - 双侧检验P值为0.1210，按α =0.05水准，不拒绝H_0，尚不能认为预防注射与非预防注射的新生儿HBV的感染率不等。

独立四格表资料的Fisher确切概率法

- 对于小样本资料，四格表χ^2检验公式不再适用，此时用Fisher确切概率法。
- 对于大样本资料，即可以直接用四格表χ^2检验公式（或校正公式）的情况，Fisher确切概率法和四格表χ^2检验的结果相近。

小结：独立四格表资料比较

- 用于两独立样本的总体率或者构成比的比较，或两个变量是否存在关联（后面内容）
- 检验基本步骤：

1. 建立检验假设：
   - H_0：两样本的总体率（或构成比）相等，或两变量不存在关联
   - H_1：两样本的总体率（或构成比）不等，或两变量存在关联
   - 建立检验水准α
2. 计算检验统计量（或直接计算累计概率P）
   - 以卡方检验为例：计算理论频数，然后根据$\chi^2=\sum\frac{(O-E)^2}{E}$或连续性校正公式计算$\chi^2$值。
3. 确定P值，作出推断结论。
   - 根据检验水准α，如果$\chi^2>\chi^2_{\alpha,1}$，则拒绝H_0，反之，接受H_0。P值为自由度为1的χ^2分布下得到的χ^2值右侧的曲线下面积。

小结：独立四格表资料比较

- 独立四格表资料χ^2检验公式的选择
  - 当**n>=40，且所有的E>=5**时，用χ^2检验的基本公式或四格表的专用公式；若$P \approx \alpha$时，用四格表资料的Fisher精确概率法。
  - 当**n>=40，但有1<=E<5**时，用χ^2检验的校正公式或四格表的专用校正公式或改用四格表资料的Fisher精确概率法
  - 当**n<40或E<1**，用四格表资料的Fisher精确概率法

主要内容

✓χ^2检验的基本思想

✓独立四格表资料的χ^2检验

- 配对四格表资料的χ^2检验
- 行×列表资料的χ^2检验
- 独立性检验
- 频数分布拟合优度的χ^2检验

配对四格表资料的χ^2检验（McNemar test）

- 配对设计，同定量资料中的配对设计
  - 研究对象按照某种特征配成对子，对内随机分配研究方案
  - 同一研究对象治疗前后的比较
  - 同一研究对象接受不同干预方法、诊断方法等的比较
  - 双胞胎设计
  - 同胞对设计
  - ……

- 共同特征：对内观察单位并不是相互独立的。

配对四格表资料的χ^2检验（McNemar test）

北京大学 PEKING UNIVERSITY

- 例4: 假设研究者想比较两种针对乳腺癌切除术后的化疗方案A和B的效果。两个治疗组在其他预后因素上应尽可能具有可比性。 为了实现这一研究目的，研究者开展了一项匹配设计的实验研究，即按照年龄和临床特征将研究对象配成对子，每对匹配的研究对象随机分配A方案和B方案。对研究对象随访 5 年，以是否生存作为结果变量。得到如下结果：

表4-1: 两种化疗方案对乳腺癌术后生存率影响的比较

方案	5年生存例数	5年内死亡例数	合计
A	526	95	621
B	515	106	621
合计	1041	201	**1242**

- 是否可以直接用四格表χ^2检验来进行两方案生存率的比较?
- 不能!
- 独立四格表χ^2检验进行两样本的总体率（或构成比）的比较时，要求两个样本是相互独立的。显然，本例中的两个样本不独立。

数据来源：*Rosner, Bernard (Bernard A.). Fundamentals of Biostatistics. Boston :Brooks/Cole, Cengage Learning, 2011.*

配对四格表资料的χ^2检验（McNemar test）

- 例4: 假设研究者想比较两种针对乳腺癌切除术后的化疗方案A和B的效果。两个治疗组在其他预后因素上应尽可能具有可比性。 为了实现这一研究目的，研究者开展了一项匹配设计的实验研究，即按照年龄和临床特征将研究对象配成对子，每对匹配的研究对象随机分配A方案和B方案。对研究对象随访 5 年，以生存作为结果变量。得到如下结果：

重新设计四格表：

表4-2: 两种化疗方案对乳腺癌术后生存率影响的比较

		方案B		合计
		5年生存	5年内死亡	
方案A	5年生存	510	16	526
	5年内死亡	5	90	95
	合计	515	106	**621**

配对四格表资料的χ^2检验（McNemar test）

北京大学 PEKING UNIVERSITY

表4-1: 两种化疗方案对乳腺癌术后生存率影响的比较

方案	5年生存例数	5年内死亡例数	合计
A	526	95	621
B	515	106	621
合计	1041	201	**1242**

观察和分析的单位：个体数

表4-2: 两种化疗方案对乳腺癌术后生存率影响的比较

		方案B		合计
		5年生存	5年内死亡	
方案A	5年生存	510	16	526
	5年内死亡	5	90	95
	合计	515	106	**621**

观察和分析的单位：对子数

注意：观察和分析的单位，并非必须是个体的人individual，可以是研究者感兴趣的任意事物，例如牙科学研究中的牙齿，基础医学中研究的细胞等。只在人群研究中，individual是最常见的研究单位。

配对四格表资料的χ^2检验（McNemar test）

表4-2: 两种化疗方案对乳腺癌术后生存率影响的比较

		方案B		合计
		5年生存	5年内死亡	
方案A	5年生存	510	16	526
	5年内死亡	5	90	95
	合计	515	106	**621**

- 510对研究对象中，方案A和方案B均生存；
- 90对研究对象中，方案A和方案B均死亡；
- 只有21对（16+5）研究对象，方案A和方案B的结局不同。
- 结局一致（concordant）的对子在比较方案A和方案B的差异时，不提供任何信息；只有结局不一致（discordant）的对子才提供两方案结局不同的信息。
- 共有两类结局不一致的对子：方案A好于方案B（16），方案B好于方案A（9）。

配对四格表资料的χ^2检验（McNemar test）

表4-2: 两种化疗方案对乳腺癌术后生存率影响的比较

		方案B		合计
		5年生存	5年内死亡	
方案A	5年生存	510	16	526
	5年内死亡	5	90	95
合计		515	106	**621**

- 假设不一致的对子中，方案A好于方案B的概率是p。
- 如果两个方案的治疗效果无差异，那么方案A好于方案B的对子数应该等于方案B好于方案A的对子数，即$p = 1/2$。因此配对设计计数资料比较的假设实际应为：
  - $H_0: p = 1/2$　　　　　$H_0: b = c$
  - $H_1: p \neq 1/2$　或者　$H_1: b \neq c$

配对四格表资料的χ^2检验（McNemar test）

北京大學 PEKING UNIVERSITY

- 假设配对设计的计数资料的比较中，不一致的对子数为n_D，其中一种类型的不一致对子数为n_A，另一种类型的不一致对子数为n_B，根据二项分布的均数和标准差的公式：
- 在无效假设成立的前提下：$E(n_A) = n_D/2, \mathrm{Var}(n_A) = n_D/4.$
- 根据大样本情形下（$n_D \geq 20$ 或$n_D/4 \geq 5$）二项分布近似正态分布的原理，因此有：

$$\chi^2 = \left(\frac{n_A - n_D/2}{\sqrt{n_D/4}}\right)^2 = \frac{(n_A - n_D/2)^2}{n_D/4} = \frac{(n_A - (n_A + n_B)/2)^2}{(n_A + n_B)/4} = \frac{((n_A - n_B)/2)^2}{(n_A + n_B)/4} = \frac{(n_A - n_B)^2}{(n_A + n_B)} = \frac{(b-c)^2}{(b+c)} \quad (n_D > 40\text{时成立})$$

或者按χ^2公式推导： $\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(b-(b+c)/2)^2}{(b+c)/2} + \frac{(c-(b+c)/2)^2}{(b+c)/2} = \frac{(b-c)^2}{b+c}$

校正公式： $\chi^2 = \left(\frac{|n_A - n_D/2| - 1/2}{\sqrt{n_D/4}}\right)^2 = \frac{(|n_A - n_B| - 1)^2}{(n_A + n_B)} = \frac{(|b-c|-1)^2}{(b+c)} \quad (n_D \geq 20\text{时成立})$

配对χ^2检验（又称McNemar检验）

配对四格表资料的χ^2检验（McNemar test）

北京大学 PEKING UNIVERSITY

- 1. 建立检验假设，确定检验水准
  - H_0： $b = c$, 即两种化疗方案的总体生存率相同
  - H_1： $b \neq c$, 即两种化疗方案的总体生存率不同
  - $\alpha = 0.05$
- 2. 计算检验统计量
  - $b + c = 21$, 因此用校正公式，得$\chi^2 = \frac{(|5-16|-1)^2}{5+16} = 4.76$
- 3. 确定*P*值，作出推断结论
  - 本例自由度为1, $\chi^2 = 4.76$，$P = 0.029$，按$\alpha = 0.05$水准，拒绝H_0, 接受H_1，可认为两种化疗方案的生存率不同，方案A的生存率更高。

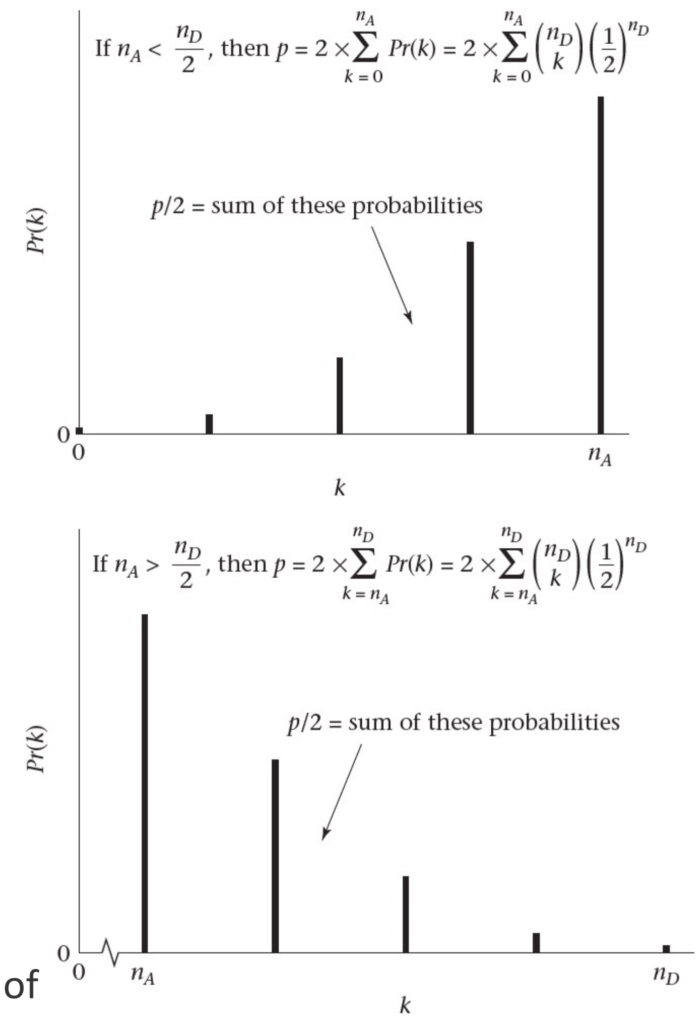
表4-2: 两种化疗方案对乳腺癌术后生存率影响的比较

		方案B		合计
		5年生存	5年内死亡	
方案A	5年生存	510	16	526
	5年内死亡	5	90	95
	合计	515	106	**621**

配对四格表资料的精确概率计算(一般性了解)

- 如果$n_D < 20$*，二项分布近似正态分布不成立，可以使用二项分布来计算精确概率。公式如下：
  - 如果$n_A < \frac{n_D}{2}$，$P = 2\times \sum_{k=0}^{n_A} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D}$
  - 如果$n_A > \frac{n_D}{2}$，$P = 2\times \sum_{k=n_A}^{n_D} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D}$
  - 如果$n_A = \frac{n_D}{2}$，$P = 1$
- 上述公式适用于n_D为任何值的情况，但对$n_D < 20$的情况尤其适用。

Computation of the *p*-value for McNemar's test—exact method



* 注：也有文献或教科书里写的是25。见Pembury Smith, M.Q.R., Ruxton, G.D. Effective use of the McNemar test. *Behav Ecol Sociobiol* **74**, 133 (2020).

小结：配对四格表资料的χ^2检验

- 一般步骤：

1. 整理配对四格表，找出不一致的对子总数 n_D 和其中一种类型的不一致对子数 n_A 。
2. 根据n_D大小选择正确的检验方法，以McNemar检验为例（ $n_D \geq 20$ 的情况）：
3. 计算检验统计量

$$\chi^2 = \frac{(n_A - n_D/2)^2}{n_D/4} \text{ 或 } \chi^2 = \left(\frac{|n_A - n_D/2| - 1/2}{\sqrt{n_D/4}}\right)^2$$

4. 对于双侧检验：如果$\chi^2 > \chi^2_{\alpha,1}$，则拒绝$H_0$，否则接受$H_0$。精确概率$P = Pr(\chi_1^2 > \chi^2)$

主要内容

✓ χ^2检验的基本思想

✓独立四格表资料的χ^2检验

✓配对四格表资料的χ^2检验

- 行×列表资料的χ^2检验
- 独立性检验
- 频数分布拟合优度的χ^2检验

行×列表资料的χ^2检验

北京大學 PEKING UNIVERSITY

- 四格表资料：每个变量只有两类。
- 行×列表资料：$R \times C$表 R: row, C: column
  - 行变量有R类；列变量有C类
- 例5（例1的延续）：研究者想进一步研究首次生育年龄与乳腺癌发病率的关系。他们开展了一项队列研究，并将首次生育年龄分为5组，结果如下表5。请问：不同生育年龄组的乳腺癌发病率是否相同？

表5：不同生育年龄组的乳腺癌发病率的比较

	首次生育年龄					合计
	<20	20-24	25-29	30-34	>=35	
病例组	320	1206	1011	463	220	3220
对照组	1422	4432	2893	1092	406	10245
合计	1742	5638	3904	1555	626	13465
病例所占百分比	0.184	0.214	0.259	0.298	0.351	0.239

数据来源： Rosner, Bernard (Bernard A.). Fundamentals of Biostatistics. Boston :Brooks/Cole, Cengage Learning, 2011. Reprinted with permission by WHO Bulletin, 43, 209–221, 1970.

行×列表资料的χ^2检验

- 行×列表资料的χ^2检验方法与四格表的χ^2检验相同
- 如果比较结果差异有统计学意义，则需要考虑进行两两比较
- 行×列表资料的χ^2检验可用基本公式，或由基本公式推导出下列专有公式：

$$\chi^2 = \sum \frac{(O-E)^2}{E}, v = (R-1)(C-1)$$

$$\chi^2 = n\left(\sum \frac{O^2}{n_R n_C} - 1\right), v = (R-1)(C-1)$$

行×列表资料的χ^2检验

北京大学 PEKING UNIVERSITY

实际频数

	首次生育年龄					
	<20	20-24	25-29	30-34	>=35	合计
病例组	320	1206	1011	463	220	3220
对照组	1422	4432	2893	1092	406	10245
合计	1742	5638	3904	1555	626	13465
病例所占百分比	0.184	0.214	0.259	0.298	0.351	0.239

理论频数

	首次生育年龄					
	<20	20-24	25-29	30-34	>=35	合计
病例组	**416.6**	**1348.3**	**933.6**	**371.9**	**149.7**	3220
对照组	**1325.4**	**4289.7**	**2970.4**	**1183.1**	**476.3**	10245
合计	1742	5638	3904	1555	626	13465
病例所占百分比	0.239	0.239	0.239	0.239	0.239	0.239

比较实际频数和理论频数，两个表格越接近，越有可能说明不同生育年龄组的乳腺癌发病率越一致。

行×列表资料的χ^2检验

- 1. 建立检验假设，确定检验水准

H_0：$\pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5$，即五个生育年龄组的乳腺癌总体发病率相等

H_1：$\pi_1, \pi_2, \pi_3, \pi_4, \pi_5$ 不等或不全相等，即五个生育年龄组的乳腺癌总体发病率不全相等

$\alpha = 0.05$

- 2. 计算检验统计量及自由度

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 130.338, \quad \nu = (5-1)(2-1) = 4$$

- 3. 确定P值，作出推断结论

$\chi^2 = 130.338$, 查 χ^2界值表，得$P<0.001$，按$\alpha = 0.05$水准，拒绝H_0，接受H_1，可认为五个生育年龄组的乳腺癌总体发病率不全相等。

行×列表资料的χ^2检验的基本步骤

北京大學 PEKING UNIVERSITY

- 用于多个独立样本的率（或构成比）的比较
- 基本步骤

1. 分析行×列表资料，计算每个格子的理论频数
2. 计算检验统计量及自由度

$$\chi^2 = \sum \frac{(O-E)^2}{E}, v = (R-1)(C-1)$$

3. 对于给定的检验水准α，
   - 如果$\chi^2 > \chi^2_{\alpha,(R-1)(C-1)}$，拒绝$H_0$; 如果$\chi^2 < \chi^2_{\alpha,(R-1)(C-1)}$，接受$H_0$
4. χ^2统计量对应的P值为$v = (R-1)(C-1)$自由度下χ^2分布计算所得的χ^2统计量右侧的曲线下面积。

行×列表资料的χ^2检验的注意事项

1. 使用行×列表资料的χ^2检验需满足以下两点：
   - 不宜有超过1/5的格子数的期望频数<5
   - 不宜有任意格子的期望频数<1
   - 一般来说，连续性校正公式不用于大于 2 × 2 的列联表，因为统计学家根据经验发现该校正无助于通过卡方分布逼近检验统计量。

如果出现上述情况，最好的解决方法有：增加样本量、考虑删去理论频数太小的行或列，或者将理论频率太小的行或列与性质相近的行或列合并等。

2. 多个样本率比较时，若拒绝无效假设，只能推断总体率之间总的来说有差别，但不能说明具体哪两个总体率之间有差别。如需进一步推断两两总体率的差别，需要做多个样本率的多重比较（Bonferroni correction）。

3. 对于有序的行×列表资料，不宜采用χ^2检验，因为χ^2检验与分类变量的顺序无关。

行×列表资料的分类及其检验方法的选择

1. 双向无序行×列表资料
   - 若研究目的为多个总体率（或构成比）的比较，可用行×列表资料的χ^2检验。
   - 若研究目的为两个变量之间是否有关联和关联的大小，可以用行×列表资料的 χ^2 检验及计算Pearson列联系数。
2. 单向有序行×列表资料
   - 若分组变量（如病情轻重）是有序的，而指标变量（如有效率）是无序的，研究目的是分析不同分组变量的指标变量的总体率或构成比的差异，可以用行×列表资料的χ^2检验。
   - 若分组变量（如治疗方法）是无序的，而指标变量（如得分）是有序的，研究目的是比较不同分组变量的指标变量的构成比，这时宜用非参数检验方法（符号秩检验）。
3. 双向有序属性相同的行×列表资料
   - 如两种方法对同一批次样品的测定结果。若研究目的是分析两变量的一致性，宜用一致性检验或kappa检验，或其他特殊模型
4. 双向有序属性不同的行×列表资料
   - 若研究目的是分析不同分组变量间指标变量的差异，可以视为单向有序行×列表资料，选用非参数检验。
   - 若研究目的是分析两个有序变量是否存在关联，可以用等级相关分析。
   - 若研究目的是分析两个有序变量之间是否存在线性变化趋势，可以用有序分组资料的线性趋势检验。

频数表资料和等级资料的两样本比较

- 例6: 39名吸烟工人和40名不吸烟工人的碳氧血红蛋白HbCO含量(%)见表6。问吸烟工人的HbCO含量是否高于不吸烟工人的HbCO含量?

表6：吸烟工人和不吸烟工人的HbCO含量比较

含量	吸烟工人	不吸烟工人	合计
（1）	（2）	（3）	（4）
很低	1	2	3
低	8	23	31
中	16	11	27
偏高	10	4	14
高	4	0	4
合计	39（n_1）	40（n_2）	79

采用两个独立样本比较的Wilcoxon秩和检验

数据来源：孙振球，徐勇勇主编。医学统计学（第三版），北京，人民卫生出版社，2010年。

频数表资料和等级资料的两样本比较

1. 建立检验假设，确定检验水准

- H_0：吸烟工人和不吸烟工人的HbCO含量总体分布位置相同
- H_1：吸烟工人的HbCO含量高于不吸烟工人的HbCO含量
- α =0.05

2.

含量	吸烟工人	不吸烟工人	合计	秩范围	平均秩	秩和	
						吸烟工人	不吸烟工人
（1）	（2）	（3）	（4）	（5）	（6）	（7）=(2)*(6)	（8）==(3)*(6)
很低	1	2	3	1～3	2	2	4
低	8	23	31	4～34	19	152	437
中	16	11	27	35～61	48	768	528
偏高	10	4	14	62～75	68.5	685	274
高	4	0	4	76～79	77.5	310	0
合计	39（n_1）	40（n_2）	79	--	--	1917 （T_1）	1243（T_2）

3.

- 查表，得P<0.005，按α =0.05水准，拒绝H_0，接受 H_1，认为吸烟工人的HbCO含量高于不吸烟工人。

频数表资料和等级资料的两样本比较

1. 建立检验假设，确定检验水准

- H_0：吸烟工人和不吸烟工人的HbCO含量总体分布位置相同
- H_1：吸烟工人的HbCO含量高于不吸烟工人的HbCO含量
- α =0.05

2. 计算检验统计量

- 先确定各等级的合计人数、秩范围和平均秩，再计算两样本各等级的秩和。
- 本例 T=1917，用下式计算u值，得u=3.702

$$u=\frac{T-n_1(N+1)/2}{\sqrt{\frac{n_1 n_2(N+1)}{12}\left(1-\frac{\sum(t_j^3-t_j)}{N^3-N}\right)}}$$

3. 确定P值，作出推断结论

- 查表，得P<0.005，按α =0.05水准，拒绝H_0，接受 H_1，认为吸烟工人的HbCO含量高于不吸烟工人。

频数表资料和等级资料的多样本比较

北京大学 PEKING UNIVERSITY

- 例7: 四种疾病患者痰液内嗜酸性粒细胞的检查结果见表7。请问，四种疾病患者痰液内的嗜酸性粒细胞有无差别?

多个独立样本比较的Kruskal–Wallis H检验

表7: 四种疾病患者痰液内的嗜酸性粒细胞比较

白细胞	支气管扩张	肺水肿	肺癌	病毒性呼吸道感染	合计
（1）	（2）	（3）	（4）	（5）	（6）
-	0	3	5	3	11
+	2	5	7	5	19
++	9	5	3	3	20
+++	6	2	2	0	10

数据来源：孙振球，徐勇勇主编。医学统计学（第三版），北京，人民卫生出版社，2010年。

频数表资料和等级资料的多样本比较

- 1. 建立检验假设，确定检验水准
  - H_0：四种疾病患者痰液内嗜酸粒细
  - H_1：四种疾病患者痰液内嗜酸粒细
  - α =0.05
- 2. 计算检验统计量
  - 先确定各等级的合计人数、秩范围
    的秩和。本例$R_1 = 0×6 + 2×21 +$
    算其他样本的秩和，按照下述公式

$$H=\frac{12}{N(N+1)}\left(\sum\frac{R_i^2}{n_i}\right)-3(N+1)$$

白细胞	支气管扩张	肺水肿	肺癌	病毒性呼吸道感染	合计	秩范围	平均秩
（1）	（2）	（3）	（4）	（5）	（6）	（7）	（8）
-	0	3	5	3	11	1～11	6
+	2	5	7	5	19	12～30	21
++	9	5	3	3	20	31～50	40.5
+++	6	2	2	0	10	51～60	55.5
R_i	739.5	436.5	409.5	244.5	--	--	--
n_i	17	15	17	11	60	--	--
$\bar{R}_i$	43.50	29.10	24.09	22.23	--	--	--

- 3. 确定P值，作出推断结论
  - 自由度为4−1=3，查卡方界值表，得P<0.005，按α =0.05水准，拒绝H_0，接受H_1，认为四种疾病患者痰液内的嗜酸性粒细胞有差别。

频数表资料和等级资料的多样本比较

- 1. 建立检验假设，确定检验水准
  - H_0：四种疾病患者痰液内嗜酸粒细胞总体分布位置相同
  - H_1：四种疾病患者痰液内嗜酸粒细胞总体分布位置不同
  - α =0.05
- 2. 计算检验统计量
  - 先确定各等级的合计人数、秩范围和平均秩，再计算多个样本各等级的秩和。本例$R_1 = 0\times6 + 2\times21 + 9\times40.5 + 6\times55.5 = 739.5$.依次计算其他样本的秩和，按照下述公式，计算得$H_c = 154.52$

$$H=\frac{12}{N(N+1)}\left(\sum\frac{R_i^2}{n_i}\right)-3(N+1) \qquad H_C=H/C, \quad C=1-\sum(t_j^3-t_j)/(N^3-N)$$

- 3. 确定P值，作出推断结论
  - 自由度为4–1=3，查卡方界值表，得$P<0.005$，按α =0.05水准，拒绝H_0 ，接受H_1 ，认为四种疾病患者痰液内的嗜酸性粒细胞有差别。

主要内容

✓χ^2检验的基本思想

✓独立四格表资料的χ^2检验

✓配对四格表资料的χ^2检验

✓行×列表资料的χ^2检验

- 独立性检验
- 频数分布拟合优度的χ^2检验

独立性检验（关联分析）

北京大学 PEKING UNIVERSITY

- Pearson's chi-squared test is used to assess three types of comparison: **goodness of fit, homogeneity, and independence**.
  - A test of goodness of fit establishes **whether an observed frequency distribution differs from a theoretical distribution**.
  - The "**test of homogeneity**" is a way of determining whether two or more sub-groups of a population share the same distribution of a single categorical variable.
  - The "**test of independence**" is a way of determining whether two categorical variables are associated with one another in the population.
  - "**homogeneity and independence sound the same**": **the difference is a matter of design**. In the test of independence, observational units are collected at random from a population and two categorical variables are observed for each unit. In the test of homogeneity, the data are collected by randomly sampling from each sub-group separately. The null hypothesis is that each sub-group shares the same distribution of another categorical variable. **The difference between these two tests is subtle yet important.**
  - Note that in the test of independence, two variables are observed for each observational unit. In the goodness-of-fit test there is only one observed variable.

https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

什么是independence?

- 概率probability
  - 包括三个组成部分：随机试验、试验的可能结果、关注的事件（结果）。
  - 概率：事件A发生的概率等于事件A的发生数与所有可能结果的总数之比即：$P(A) = n(A)/N$。

- 复合事件A和B：表示A和B都发生。
  - 用A×B或$A \cap B$表示，复合事件AB的概率为：$P(AB) = n(AB)/N$

	A	$\bar{A}$	合计
B	$P(AB)$	$P(\bar{A}B)$	$P(B)$
$\bar{B}$	$P(A\bar{B})$	$P(\bar{A}\bar{B})$	$P(\bar{B})$
合计	$P(A)$	$P(\bar{A})$	1

$$P(A) = (a + c)/N$$
$$P(B) = (a + b)/N$$
$$P(\bar{A}) = (b + d)/N$$
$$P(\bar{B}) = (c + d)/N$$
$$P(AB) = a/N$$

蒋庆琅原著，方积乾等译。实用统计分析方法，北京，北京医科大学、中国协和医科大学联合出版社，1998年。

什么是independence?

	A	$\bar{A}$	合计
B	$P(AB)$	$P(\bar{A}B)$	$P(B)$
$\bar{B}$	$P(A\bar{B})$	$P(\bar{A}\bar{B})$	$P(\bar{B})$
合计	$P(A)$	$P(\bar{A})$	1

- 复合事件A和B：条件概率 conditional probability
  - 在一个事件发生的条件下，另一事件发生的概率称为条件概率
  - 给定A事件发生的条件下，事件B发生的概率用$P(B \mid A)$表示，｜代表given的意思，即给定。竖线右侧是条件，公式为：$P(B \mid A) = P(AB)/P(A)$
  - 条件概率$P(B \mid A)$与$P(A \mid B)$是不同的
  - 乘法定理：$P(AB) = P(B \mid A)\,P(A)$
  - 因为事件AB等价于事件BA，$P(AB) = P(A \mid B)\,P(B)$

独立事件

	A	$\bar{A}$	合计
B	P(AB)	$P(\bar{A}B)$	$P(B)$
$\bar{B}$	$P(A\bar{B})$	$P(\bar{A}\bar{B})$	$P(\bar{B})$
合计	$P(A)$	$P(\bar{A})$	1

- 假若事件B的概率与在给定A条件下的事件B的条件概率相等，那么事件B独立于事件A，或

$$P(B \mid A) = P(B)$$

- 意味着事件B的发生不受事件A的影响。如果B独立于A，那么B也独立于$\bar{A}$，即：

$$P(B \mid A) = P(B) = P(B \mid \bar{A})$$

- **为确定事件B是否独立于事件A，只需计算$P(B \mid A)$与$P(B \mid \bar{A})$是否相等，若两个值相等，则B独立于A。**

- 相互独立：当且仅当“A独立于B时B独立于A”，称两事件A和B相互独立。

独立事件的乘法定理

- 两事件的乘法定理如下：
  - 已知：

$$P(AB) = P(B \mid A)\, P(A)$$
$$P(AB) = P(A \mid B)\, P(B)$$

  - 当A和B相互独立时， $P(B \mid A) = P(B)$及$P(A \mid B) = P(A)$，则有：

$$\boldsymbol{P(AB)} = P(B \mid A)\, P(A) = \boldsymbol{P(B)P(A)}$$

实际频率（数） 理论频率（数）

Pearson's chi-squared test: the "**test of independence**" is a way of determining whether two categorical variables are associated with one another in the population.

A和B不独立意味着A和B存在关联

行×列表分类资料的关联性分析

北京大学 PEKING UNIVERSITY

例8: 从某地随机抽样，获得一个5801人的随机样本的血型资料。其中，ABO血型和MN血型结果如下表。问两种血型系统之间是否存在关联?

表8: 某地一个随机样本的血型数据

ABO血型	MN血型			合计
	M	N	MN	
O	431	490	902	1823
A	388	410	800	1598
B	495	587	950	2032
AB	137	179	32	348
合计	1451	1666	2684	5801

数据来源：孙振球，徐勇勇主编。医学统计学（第三版），北京，人民卫生出版社，2010年。

行×列表分类资料的关联性分析

例8: 从某地随机抽样，获得一个5801人的随机样本的血型资料。其中的ABO血型和MN血型结果如下表。问两种血型系统之间是否存在关联?

ABO血型	MN血型			合计
	M	N	MN	
O	431	490	902	1823
A	388	410	800	1598
B	495	587	950	2032
AB	137	179	32	348
合计	1451	1666	2684	5801

1. 建立检验假设，确定检验水准
   - H_0：两种血型系统间无关联
   - H_1：两种血型系统间有关联
   - $\alpha = 0.05$
2. 计算检验统计量及自由度
   - $\chi^2 = \sum\frac{(O-E)^2}{E} = 213.16$，$\nu = (4-1)(3-1) = 6$
3. 确定*P*值，作出推断结论
   - $\chi^2 = 213.16$, 查 χ^2界值表，得$P<0.005$，按$\alpha = 0.05$水准，拒绝H_0，接受H_1，可认为两种血型系统间有关联。

行×列表分类资料的关联性指标

- 关联性分析的主要内容：
  - 分析横标目和纵标目所代表的两个分类变量之间有无关联?
    - 行×列表资料的χ^2检验
  - 关系的密切程度如何?
    - Pearson 列联系数C：$\mathrm{C}=\sqrt{\frac{\chi^2}{n+\chi^2}}$　　χ^2是行×列表资料的χ^2值，n为样本含量。
    - C 取值范围0–1。0代表完全独立，1代表完全相关。越接近0，关系越不密切，越接近1，关联越密切。

行×列表分类资料的关联性分析

北京大学 PEKING UNIVERSITY

例8: 从某地随机抽样，获得一个5801人的随机样本的血型资料。其中的ABO血型和MN血型结果如下表。问两种血型系统之间是否存在关联?

$\chi^2 = 213.16$, 查 χ^2界值表，得P<0.005，按$\alpha = 0.05$水准，拒绝H_0，接受H_1，可认为两种血型系统间有关联。

进一步计算Pearson 列联系数，得$C = \sqrt{\frac{\chi^2}{n+\chi^2}}$=0.1883，可以看出两种血型系统虽有关联，关联有统计学意义，但关系并不密切。

ABO血型	MN血型			合计
	M	N	MN	
O	431	490	902	1823
A	388	410	800	1598
B	495	587	950	2032
AB	137	179	32	348
合计	1451	1666	2684	5801

四格表中可以计算的其他关联指标

- Phi (只适用于四格表)

$$\phi = \frac{n_{11}n_{22} - n_{21}n_{12}}{\sqrt{(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})}} = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

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

- Cramer’s V (更general的Phi)

$$Cramer's\ V = \sqrt{\frac{\chi^2}{[\min(I-1, J-1)]n}}$$

Φ ， Cramer’s V，及列联系数：都是愈接近于0，愈说明两个分类变量没有关系；愈接近于1，说明关系愈密切。

Φ系数只适用于四格表；非四格表资料可采用Cramer’s V系数和列联系数

Cramer’s V系数最为常用

率的比较 vs. 独立性检验

- 例1：比较的是乳腺癌病例和对照组人群的首次生育年龄构成比的不同。
  - 有两个独立人群，乳腺癌病例组和对照组；要比较的是两组人群生育年龄构成比的差异。
  - test for homogeneity of binomial proportions
  - 列合计或者行合计是固定的数值，例如病例人数和对照人数。而每行中的生育年龄构成比是一个服从二项分布的随机变量
- 例8: 比较两种血型的分布是否存在关联。
  - test for the independence of two characteristics in the same sample
  - 人群中随机选择的总样本，有两个特征变量（两种血型），在计算率或构成比的时候，无论哪一个变量均可以做分母。两个变量的行合计和列合计均是固定的。既可以比较ABO血型中的MN血型的构成比，也可以比较MN血型中ABO血型的构成比。

- 两种检验的步骤是完全相同的，差异只是取决于研究设计。

独立性检验（关联分析）

- Pearson's chi-squared test is used to assess three types of comparison: **goodness of fit, homogeneity, and independence**.
  - A test of goodness of fit establishes **whether an observed frequency distribution differs from a theoretical distribution**.
  - The "**test of homogeneity**" is a way of determining whether two or more sub-groups of a population share the same distribution of a single categorical variable.
  - The "**test of independence**" is a way of determining whether two categorical variables are associated with one another in the population.
  - "**homogeneity and independence sound the same**": **the difference is a matter of design**. In the test of independence, observational units are collected at random from a population and two categorical variables are observed for each unit. In the test of homogeneity, the data are collected by randomly sampling from each sub-group separately. The null hypothesis is that each sub-group shares the same distribution of another categorical variable. **The difference between these two tests is subtle yet important.**
  - Note that in the test of independence, two variables are observed for each observational unit. In the goodness-of-fit test there is only one observed variable.

率的比较和关联性分析中特别要注意的问题

北京大學 PEKING UNIVERSITY

- 混杂因素（辛普森悖论）
  - “Simpson’s paradox” occurs when a statistical association is created by combining samples which themselves show no evidence of association
  - 率的比较或关联性分析都是基于没有混杂因素（第三方因子）存在的假设，如果存在混杂因素，下结论应慎重。需要对混杂因素进行控制后再下结论。
  - 例9: 比较甲乙两种疗法治疗某病的治愈率，结果见表9:

表9: 甲乙两种疗法治疗某病的治愈率

病型	甲疗法			乙疗法		
	病人数	治愈数	治愈率（%）	病人数	治愈数	治愈率（%）
普通型	300	180	$\boxed{60.0}$	100	65	$\boxed{65.0}$
重型	100	35	$\boxed{35.0}$	300	125	$\boxed{41.7}$
合计	400	215	$\boxed{53.8}$	400	190	$\boxed{47.5}$

数据来源：孙振球，徐勇勇主编。医学统计学（第三版），北京，人民卫生出版社，2010年。

主要内容

✓χ^2检验的基本思想

✓独立四格表资料的χ^2检验

✓配对四格表资料的χ^2检验

✓行×列表资料的χ^2检验

✓独立性检验

- 频数分布拟合优度的χ^2检验

频数分布拟合优度的χ^2检验

- 频数分布拟合优度（goodness-of-fit）检验，是χ^2检验方法的另一常见应用，用于判断实际观察频数的分布是否服从某一理论期望频数分布。
- 如：判断某一变量观察值是否分布均匀, 判断某一变量观察值是否服从正态分布、二项分布、Poisson分布等理论分布
- 是单变量的χ^2检验

频数分布拟合优度的χ^2检验

例10: 某美容院对某年顾客抱怨原因进行分析，结果见下表，问

构成比是否有差异?

如果没有差异，
四种原因应各
占25%

表10：某年某医学美容院的顾客抱怨情况

抱怨原因	人数	构成比（%）
正常生理反应	87	20.47
顾客自身问题	195	45.88
管理问题	116	27.29
技师技术问题	27	6.35
合计	425	100.00

频数分布拟合优度的χ^2检验

1. 建立检验假设，确定检验水准

H_0：顾客抱怨原因的总体构成比相同

H_1：顾客抱怨原因的总体构成比不同

$\alpha = 0.05$

2. 计算检验统计量

如H_0成立，顾客抱怨原因构成比相同，均为25%，$E = n/R = 425/4 = 106.25$，按χ^2检验的基本公式，计算：

$$\chi^2 = \sum\frac{(O-E)^2}{E} = \frac{(87-106.25)^2}{106.25} + \frac{(195-106.25)^2}{106.25} + \frac{(116-106.25)^2}{106.25} + \frac{(27-106.25)^2}{106.25} = 137.63, \nu = 4-1 = 3$$

3. 确定P值，作出推断结论

查χ^2界值表，$\chi^2 = 137.63 > \chi^2_{0.05,3} = 7.81$，$P < 0.05$，按$\alpha$ =0.05水准，拒绝H_0，接受H_1，认为顾客各种抱怨原因的构成比有差异。

频数分布拟合优度的χ^2检验

- 例11: 假设从某社区中随机选取研究对象测量其舒张压（DBP）值，血压值的频数分布表见表10。请问血压值分布是否来自一个正态分布的总体。假设正态总体的均数和标准差用该样本的均数和标准差来表示，即$\bar{x} = 80.68, s = 12.00$。

表10: 某社区人群舒张压频数分布表

DBP分组 （mm Hg）	观察频数
< 50	57
$50 \leq DBP < 60$	330
$60 \leq DBP < 70$	2132
$70 \leq DBP < 80$	4584
$80 \leq DBP < 90$	4604
$90 \leq DBP < 100$	2119
$100 \leq DBP < 110$	659
≥ 110	251
合计	14736

通过标准正态转换，每个区间的期望频数计算公式：

$$14736\times\left(\Phi\left(\frac{(b-\mu)}{\sigma}\right)-\Phi\left(\frac{(a-\mu)}{\sigma}\right)\right)=$$

$$14736\times\left(\Phi\left(\frac{(60-80.68)}{12}\right)-\Phi\left(\frac{(50-80.68)}{12}\right)\right)=$$

$$14736\times(\Phi(-1.723)-\Phi(-2.577))=$$

$$14736\times(0.0424-0.0053)=547.1$$

数据来源：Hypertension Detection and Follow-up Program Cooperative Group. (1977). Blood pressure studies in 14 communities—A two-stage screen for hypertension. JAMA, 237(22), 2385–2391.

频数分布拟合优度的χ^2检验

- 例11: 假设从某社区中随机选取研究对象测量其舒张压（DBP）值，血压值的频数分布表见表10。请问血压值分布是否来自一个正态分布的总体。假设正态总体的均数和标准差用该样本的均数和标准差来表示，即$\bar{x} = 80.68, s = 12.00$。

表10: 某社区人群舒张压频数分布表

DBP分组 （mm Hg）	观察频数	期望频数
< 50	57	**77.9**
$50 \le DBP < 60$	330	**547.1**
$60 \le DBP < 70$	2132	**2126.7**
$70 \le DBP < 80$	4584	**4283.3**
$80 \le DBP < 90$	4604	**4478.5**
$90 \le DBP < 100$	2119	**2431.1**
$100 \le DBP < 110$	659	**684.1**
≥ 110	251	**107.2**
合计	14736	**14736**

数据来源：Hypertension Detection and Follow-up Program Cooperative Group. (1977). Blood pressure studies in 14 communities—A two-stage screen for hypertension. JAMA, 237(22), 2385–2391.

频数分布拟合优度的χ^2检验

- 1. 建立检验假设，确定检验水准

  H_0：舒张压的分布符合正态分布

  H_1：舒张压的分布不符合正态分布

  $\alpha = 0.05$

- 2. 计算检验统计量

  按χ^2检验的基本公式，计算：

  $\chi^2 = \sum\frac{(O-E)^2}{E} = \frac{(57-77.9)^2}{77.9} + \cdots + \frac{(251-107.2)^2}{107.2} = 350.2, \nu = g - 1 - k = 8 - 1 - 2 = 5$ (g: 组数，k：计算期望频数用到的参数个数)

- 3. 确定P值，作出推断结论

  查χ^2界值表，$\chi^2 = 350.2 > \chi^2_{0.05,5} = 11.071$，$P < 0.05$，按$\alpha = 0.05$水准，拒绝$H_0$，接受$H_1$，认为舒张压的分布不符合正态分布（N(80.68, 12)）。

主要内容

- ✓χ^2检验的基本思想
- ✓独立四格表资料的χ^2检验
- ✓配对四格表资料的χ^2检验
- ✓行×列表资料的χ^2检验
- ✓独立性检验
- ✓频数分布拟合优度的χ^2检验

本节课小结

- 卡方检验的应用
  - 总体率（或构成比）的比较
  - 独立性检验（关联分析）
  - 频数分布的拟合优度检验 （单变量）

- 四格表资料
  - 两个样本所在总体的率的比较
  - 两个样本所在总体的构成比的比较
  - 两个二分类变量的关联性检验

- 行×列表资料
  - 多个样本所在总体的率的比较
  - 多个样本所在总体的构成比的比较
  - 双向无序分类资料的关联性检验

χ^2检验基本公式

- χ^2检验的基本公式：$\chi^2 = \sum\frac{(O-E)^2}{E}, \nu = (R-1)(C-1)$
- χ^2检验的连续性校正公式 ：$\chi_C^2 = \sum\frac{(|O-E|-0.5)^2}{E}, \nu = (R-1)(C-1)$

- 由基本公式可以推导一些专用公式，手算时使用：
  - 四格表χ^2检验专用公式：$\chi^2 = \frac{(ad-bc)^2 \times n}{(a+b)(c+d)(a+c)(b+d)}, \quad \nu = 1$
  - 四格表专用公式的连续性校正公式：$\chi_C^2 = \frac{(|ad-bc|-\mathrm{n}/2)^2 \times n}{(a+b)(c+d)(a+c)(b+d)}, \quad \nu = 1$
  - 行×列表资料的χ^2检验公式：$\chi^2 = n\left(\sum\frac{O^2}{n_R n_C} - 1\right), \nu = (R-1)(C-1)$

χ^2检验的决策树

1. 是否是独立样本?

- 是，跳到2；否， McNemar test（四格表）

2. 是否所有格子的期望频数均大于等于5?

- 是，跳到3；否，Fisher精确概率法

3. 是否是四格表?

- 是，四格表的χ^2检验；否，行×列表资料的χ^2检验

参考用书

- 孙振球，徐勇勇主编。医学统计学（第三版），北京，人民卫生出版社，2010年。
- 蒋庆琅原著，方积乾等译。实用统计分析方法，北京，北京医科大学、中国协和医科大学联合出版社，1998年。
- Rosner, Bernard (Bernard A.). Fundamentals of Biostatistics. Boston :Brooks/Cole, Cengage Learning, 2011.
- 王燕，康晓平主编。卫生统计学教程，北京，北京大学医学出版社，2010。
- 李晓松主编。卫生统计学（第八版），北京，人民卫生出版社，2017年。

谢 谢