



北京大学  
PEKING UNIVERSITY

# 卡方检验实习

秦雪英

- 分类资料的比较中，常用的统计检验方法及其主要应用
- 卡方检验的基本思想
- 卡方检验的基本公式（理论频数 vs 实际频数）
- 自由度为1时，右侧曲线下面积为5%时的卡方临界值
- 不同研究设计的卡方检验方法的选择
- 不同样本量情况下，独立样本的四格表资料的卡方检验方法的选择





北京大学  
PEKING UNIVERSITY

随堂小练习：10道选择题



<https://ks.wjx.top/vm/YtAgAOD.aspx#>



1. 下列关于卡方分布特征的叙述，错误的是（ ）？

- A、卡方分布的形状依赖于自由度的大小
- B、卡方分布曲线是一簇曲线
- C、以 0 为中心，左右对称(正确答案)
- D、只有一个参数，即自由度
- E、自由度为无穷大时，卡方分布趋于正态分布



2. 横轴上, 自由度为 1 的卡方分布曲线下, 从 0 到\_\_\_\_\_的面积为 95%。

A、 1.96

B、 -1.96

C、 3.84(正确答案)

D、 2.58

E、 1.64



3. 为比较某新药与常规药治疗高血压的疗效是否有差别，将 80 名 50-60 岁的高血压患者按年龄相近配成 40 对。将对子中的两名患者随机分为两组，一组给予新药，另一组给予常规药物。经过一个疗程的治疗后，测量其血压下降是否超过 4mmHg。为实现本研究目的，并结合题目所给信息，请问以下哪个假设检验方法最为合适？

- A、独立样本四格表资料的卡方检验
- B、配对四格表资料的卡方检验(正确答案)
- C、成组 t 检验
- D、配对设计差值的符号秩和检验
- E、成组设计两样本比较的秩和检验



4. 关于独立四格表资料的卡方检验的检验假设, 说法正确的是\_\_\_\_\_。

- A、无效假设是两样本的率相同
- B、无效假设是两样本的分布相同
- C、备择假设是两总体的率相同
- D、备择假设是两总体的率不同(正确答案)
- E、无效假设是两总体的均数相同



5. 以下检验方法中，不属于非参数检验方法的是\_\_\_\_\_。

A、McNemar 检验

B、成组设计四格表资料的卡方检验

C、差分析(正确答案)

D、H 检验

E、符号秩检验





6. 对于  $R \times C$  行列表资料的卡方检验，其自由度的计算公式为\_\_\_\_\_。

A、  $R-1$

B、  $C-1$

C、  $R+C-1$

D、  $R \times C - 1$

E、  $(R-1)(C-1)$ (正确答案)



7. 为了评价两种不同的戒烟项目的效果，研究者将吸烟者随机分配干预方案 A 或 B。参加 A 项目的 6 名吸烟者中，有 1 人戒烟了，而 5 人未戒烟；参与 B 项目的 6 名吸烟者中，有 4 人戒烟了，而 2 人未戒烟。若想评价两种干预方案的戒烟效果，以下哪种假设检验最合适？

- A、McNemar 检验
- B、独立样本四格表资料的卡方检验
- C、完全随机设计的方差分析
- D、Fisher 确切概率法(正确答案)
- E、两独立样本比较的符号秩和检验



8. 将 36 名贫血患者随机等分成两组，分别用两种药物治疗，随访一段时间后，用“效果很差、无效、效果一般、效果较好、效果很好”来评价疗效，欲比较两组疗效有无差异，宜采用\_\_\_\_\_。

- A、McNemar 检验
- B、行×列表资料的卡方检验
- C、四格表资料卡方检验
- D、两独立样本比较的 Wilcoxon 秩和检验(正确答案)
- E、两独立样本比较的 t 检验



9. 将 36 名贫血患者随机等分成两组，分别用两种药物治疗，欲比较两药物治疗贫血的总体有效率的差异是否有统计学意义，宜采用\_\_\_\_\_。

- A、McNemar 检验
- B、四格表资料的 Fisher 确切概率法(正确答案)
- C、独立样本四格表资料卡方检验的基本公式
- D、独立样本四格表资料卡方检验的校正公式
- E、行×列表资料的卡方检验





10. 用两种方法检查已确诊的乳腺癌患者 120 名，甲法检出率为 60%，乙法检出率为 50%，两法一致的检出率为 35%。欲比较这两种方法的总体检出率是否有不同，宜采用\_\_\_\_\_。

- A、McNemar 检验(正确答案)
- B、四格表资料的 Fisher 确切概率法
- C、独立样本四格表资料卡方检验的基本公式
- D、独立样本四格表资料卡方检验的校正公式
- E、行×列表资料的卡方检验

## 案例讨论1



北京大学  
PEKING UNIVERSITY

- 例1：一项研究调查了使用口服避孕药与子宫内膜癌发生之间的关联。研究人员发现，在 117 名子宫内膜癌患者中，有16人在其生命中的某个时间使用口服避孕药，而 395 名对照者中有18人使用过这种药物。请问，子宫内膜癌病例和对照组的口服避孕药的服药比例是否存在差异？

## 例1 思路



北京大学  
PEKING UNIVERSITY

	服用口服避孕药	未服用口服避孕药	合计	服药比例 (%)
子宫内膜癌病例	16	101	117	13.68
子宫内膜癌对照	18	377	395	4.56
合计	34	478	512	6.64

- 两个独立样本的总体构成比的比较
- 独立样本（成组设计）四格表卡方检验
- 整理四格表，计算理论频数
- 一般公式和校正公式均可，大样本情形下，两者结果接近。

## 案例讨论2



北京大学  
PEKING UNIVERSITY

- 例2：用两种剂量的电离辐射照射小白鼠，第一种剂量照射18只，10天内死亡3只；第二种剂量照射31只，同期内死亡9只。问两种剂量对小白鼠的致死作用是否相同？



## 例2 思路



北京大学  
PEKING UNIVERSITY

	小鼠死亡例数	小鼠未死亡例数	合计	死亡率 (%)
第一种剂量	3	15	18	
第二种剂量	9	22	31	
合计	12	37	49	

- 两个独立样本的总体构成比的比较
- 独立样本（成组设计）四格表卡方检验
- 整理四格表，计算理论频数
- 一个格子的理论频数在 $[1, 5)$ 之间，宜用校正公式

## 案例讨论3



北京大学  
PEKING UNIVERSITY

- 例3：以下三个四格表来自两独立样本的四格表数据，它们的观察频数 $a$ ,  $b$ ,  $c$ ,  $d$  分别为：

1.  $a=5$ ,  $b=10$ ,  $c=11$ ,  $d=7$

2.  $a=2$ ,  $b=3$ ,  $c=7$ ,  $d=10$

3.  $a=1$ ,  $b=99$ ,  $c=10$ ,  $d=90$

<b>a</b>	<b>b</b>
<b>c</b>	<b>d</b>

- 问题：
- 四格表卡方检验与Fisher精确概率法的区别？
- 如何对两种检验方法进行选择？
- 请对上述三个四格表的数据做假设检验，并比较两个检验方法结果的差异。

## 案例讨论4



北京大学  
PEKING UNIVERSITY

例4：用下述资料做父母与孩子的饮酒状况相互独立的假设检验。

孩子	父母		合计
	饮酒	不饮酒	
饮酒	100	100	200
不饮酒	80	320	400
合计	180	420	600

## 例4 思路



北京大学  
PEKING UNIVERSITY

		父母	
孩子	饮酒	不饮酒	合计
饮酒	100	100	200
不饮酒	80	320	400
合计	180	420	600

- 配对设计
- 配对四格表的卡方检验
- 整理四格表，找出不一致的对子数
- 配对四格表卡方检验的一般公式或校正公式均可。



## 案例讨论5



北京大学  
PEKING UNIVERSITY

- 例5：在两对基因中，一对决定豌豆的颜色（黄色和绿色），另一对决定豌豆的形状（圆的和皱的，圆的为显性）。假如每对基因的作用是独立的，孟德尔的双因子杂合子的表型比例为：黄色且圆形：黄色且皱形：绿色且圆形：绿色且皱形=9:3:3:1。使用实验结果（315, 101, 108, 32）检验孟德尔的假设（ $\alpha = 0.05$ ）。
- 思路：频数分布的拟合优度检验

## 案例讨论6



北京大学  
PEKING UNIVERSITY

- 例6：色素性视网膜炎是一种通过不同遗传模式表现出来的疾病。已记录的病例具有显性、隐性和与性连锁三种遗传模式。有学者推测，该遗传模式可能与种族相关，调查结果如下：在调查的125名A血统病例中，46名患者为性连锁遗传，25名为隐性遗传，54名为显性遗传。在110名B血统病例中，1名为性连锁遗传，99名为隐性遗传，10 名为显性遗传。请问，这些数据是否表明遗传模式和血统之间存在关联？

## 例6 思路



北京大学  
PEKING UNIVERSITY

- 独立性检验或关联分析（两个变量是否存在关联）
- 行×列表的卡方检验（无序多分类， $2 \times 3$  or  $3 \times 2$ ）
- 计算理论频数
- 卡方检验的一般公式

- 使用R的MASS程序包中自带的数据库birthwt。分析低出生体重的影响因素。
- 查看数据库，了解变量的名称、内容和属性。
  - I. 分析低出生体重组与正常出生体重组（low）的子宫易激惹症（ui）的发生率有无差别。
  - II. 分析低出生体重组与正常出生体重组（low）的母亲孕期吸烟率（smoke）是否存在差别。
  - III. 分析低出生体重组与正常出生体重组（low）的母亲种族（race）的构成比是否存在差别。



## 案例讨论8



北京大学  
PEKING UNIVERSITY

- 研究者在某地区绝经后妇女中开展了一项匹配的病例对照研究，调查雌激素使用与子宫内膜癌的关联关系。子宫内膜癌病例和对照组按照年龄和城市居住时间进行匹配，结果整理为数据库demo of practice 8.xlsx。请问：子宫内膜癌病例组和对照组的雌激素使用情况是否存在不同（或雌激素使用是否与子宫内膜癌存在关联）。

- 题1：研究者欲在某大学的本科毕业生人群中开展体育锻炼相关的危险因素调查，其中研究者询问了这些学生在毕业前一年的居住场所，结果见下表：

	校内宿舍	校内公寓	校外公寓	家里	合计
从不锻炼	35	77	115	40	267
偶尔锻炼	33	68	31	9	141
经常锻炼	29	45	19	7	100

- 请问：不同锻炼水平的毕业生人群的居住场所构成比是否存在差异？

- 题2：为确定某药物对疾病的治疗效果，研究者记录107名研究对象的治疗前后的疾病诊断结果（阳性或者阴性），其中27名研究对象治疗前和治疗后的诊断结果均为阳性；50名研究对象治疗前的诊断结果为阳性、治疗后的诊断结果为阴性；12名研究对象治疗前的诊断结果为阴性，治疗后的诊断结果为阳性；18名研究对象治疗前和治疗后的诊断结果均为阴性。
  1. 请问，这是一种什么类型的研究设计？结局指标是什么类型的变量？
  2. 请整理数据，并选择合适的假设检验方法对药物治疗疾病的效果进行评价。



Table 3. Arguments for the R stat package function dhyper

	Differential Expression	NO Differential Expression	Total
IN Transcription Elongation	$x$	$m - x$	$m$
NOT IN Transcription Elongation	$k - x$	$n - (k - x)$	$n$
Total	$k$	$(m + n - k)$	$m + n$

With this knowledge in hand, it is trivial to reproduce the p-values described previously.

```
# Initialize variables
m <- 15      # Genes IN GO term
n <- 15      # Genes NOT IN GO term
k <- 15      # Gene hits, that is, differentially expressed
x <- c(0:15) # Genes both IN GO term and differentially expressed 'hits'

# Use the dhyper built-in function for hypergeometric density
probabilities <- dhyper(x, m, n, k, log = FALSE)
probabilities
```