统计检验之——卡方检验（Chi-Square Tests）

北京宏志中学　　徐德前　　100013

一. 问题情境

5 月 31 日是世界无烟日.有关医学研究表明，许多疾病，例如：心脏病、癌症、脑血管病、慢性阻塞性肺病等都与吸烟有关，吸烟已成为继高血压之后的第二号全球杀手.这些疾病与吸烟有关的结论是怎样得出的呢？我们看一下问题：

某医疗机构为了了解呼吸道疾病与吸烟是否有关，进行了一次抽样调查，共调查了 515 个成年人，其中吸烟者 220 人，不吸烟者 295 人．调查结果是：吸烟的 220 人中有 37 人患呼吸道疾病（简称患病），183 人未患呼吸道疾病（简称未患病）；不吸烟的 295 人中有 21 人患病，274 人未患病．

问题：根据这些数据能否断定“患呼吸道疾病与吸烟有关”？

二．探索活动

为了研究问题，将上述数据用下表来表示：

	患病	未患病	合计
吸烟	37	183	220
不吸烟	21	274	295
合计	58	457	515

估计吸烟者与不吸烟者患病的可能性差异：

有无关系——直观判断

	患　病	未患病	合　计（n）
吸　烟	17%	83%	100%（220）
不吸烟	7%	93%	100%（295）

在吸烟的人中，有 $\frac{37}{220}\approx 17\%$ 的人患病，在不吸烟的人中，有 $\frac{21}{295}\approx 7\%$ 的人患病．

问题：由上述结论能否得出患病与吸烟有关？把握有多大？

三．建构数学

1．独立性检验：

（1）假设 H_0：患病与吸烟没有关系．若将表中“观测值”用字母表示，则得下表：

	患病	未患病	合计
吸烟	a	b	$a+b$
不吸烟	c	d	$c+d$
合计	$a+c$	$b+d$	$a+b+c+d$

（近似的判断方法：设$n=a+b+c+d$，如果H_0成立，则在吸烟的人中患病的比例与不吸烟的人中患病的比例应差不多，由此可得$\frac{a}{a+b}\approx\frac{c}{c+d}$，即$a(c+d)\approx c(a+b)\Rightarrow ad-bc\approx 0$，因此，$|ad-bc|$越小，患病与吸烟之间的关系越弱，否则，关系越强.）

设$n=a+b+c+d$，在假设H_0成立的条件下，可以通过求 “吸烟且患病”、“吸烟但未患病”、“不吸烟但患病”、“不吸烟且未患病”的概率（观测频率），将各种人群的估计人数用a,b,c,d,n表示出来.

事件A——某人吸烟，　事件B——某人患病，

事件$\overline{A}$——某人不吸烟，事件$\overline{B}$——某人不患病.

假设H_0：患病与吸烟没有关系，即

$$H_0：P(AB)=P(A)P(B).$$

“吸烟且患病”的估计人数为$n\times P(AB)\approx n\times\frac{a+b}{n}\times\frac{a+c}{n}=\frac{(a+b)(a+c)}{n}$；

“吸烟但未患病” 的估计人数为$n\times P(A\overline{B})\approx n\times\frac{a+b}{n}\times\frac{b+d}{n}=\frac{(a+b)(b+d)}{n}$；

“不吸烟但患病”的估计人数为$n\times P(\overline{A}B)\approx n\times\frac{c+d}{n}\times\frac{a+c}{n}=\frac{(c+d)(a+c)}{n}$；

“不吸烟且未患病”的估计人数为$n\times P(\overline{A}\overline{B})\approx n\times\frac{c+d}{n}\times\frac{b+d}{n}=\frac{(c+d)(b+d)}{n}$.

如果实际观测值与假设求得的估计值相差不大，就可以认为所给数据（观测值）不能否定假设H_0．否则，应认为假设H_0不能接受，即可作出与假设H_0相反的结论.

（2）构造统计量

如果实际观测值与由事件A、B相互独立的假设的估计相差不大，那么，我们就可以认为这些差异是由随机误差造成的，假设不能被所给数据否定，否则应认为假设不能接受.

怎样刻画实际观测值与估计值的差异呢？**统计学中采用如下的量（称为χ^2 统计量）来刻画这个差异.**

卡方统计量（$\chi^2=\sum\frac{(\text{观测值}-\text{预期值})^2}{\text{预期值}}$）来进行估计．卡方统计量 χ^2 公式：

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

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

由此若 H_0 成立，即患病与吸烟没有关系，则 χ^2 的值应该很小．把 $a=37,b=183,c=21,d=274$ 代入计算得 $\chi^2=11.8634$，统计学中有明确的结论，在 H_0 成立的情况下，随机事件“$\chi^2\geq6.635$”发生的概率约为 0.01，即 $P(\chi^2\geq6.635)\approx0.01$，也就是说，在 H_0 成立的情况下，对统计量 χ^2 进行多次观测，观测值超过 6.635 的频率约为 0.01．由此，我们有 99% 的把握认为 H_0 不成立，即有 99%的把握认为“患病与吸烟有关系”．以上这种用 χ^2 统计量研究吸烟与患呼吸道疾病是否有关等问题的方法称为独立性检验．

说明：

1) 估计吸烟者与不吸烟者患病的可能性差异是用频率估计概率，利用 χ^2 进行独立性检验，可以对推断的正确性的概率作出估计，观测数据 a,b,c,d 取值越大，效果越好．在实际应用中，当 a,b,c,d 均不小于 5，近似的效果才可接受．

2) 这里所说的“呼吸道疾病与吸烟有关系”是一种统计关系，这种关系是指“抽烟的人患呼吸道疾病的可能性（风险）更大”，而不是说“抽烟的人一定患呼吸道疾病”．

3) 在假设 H_0 下统计量 χ^2 应该很小，如果由观测数据计算得到 χ^2 的观测值很大，则在一定程度上说明假设不合理（即统计量 χ^2 越大，“两个分类变量有关系”的可能性就越大）．

2．独立性检验的一般步骤：

一般地，对于两个研究对象Ⅰ和Ⅱ，Ⅰ有两类取值：类 A 和类 B（如吸烟与不吸烟），Ⅱ也有两类取值：类1和类2（如患呼吸道疾病与不患呼吸道疾病），得到如下表所示：

		II		
		类1	类2	合计
I	类 A	a	b	$a+b$
	类 B	c	d	$c+d$
	合计	$a+c$	$b+d$	$a+b+c+d$

推断“Ⅰ和Ⅱ有关系”的步骤为：

第一步，提出假设 H_0：两个分类变量Ⅰ和Ⅱ没有关系；

第二步，根据 2×2 列联表和公式计算 χ^2 统计量；

第三步，查对课本中临界值表，作出判断.

四.利用 TI-NSPIRE CAS CX 进行独立性检验

（1）创建 2×2 列联表矩阵（即观测矩阵）

操作步骤：（菜单→7 矩阵与向量→创建→1 矩阵）（图 1-4）

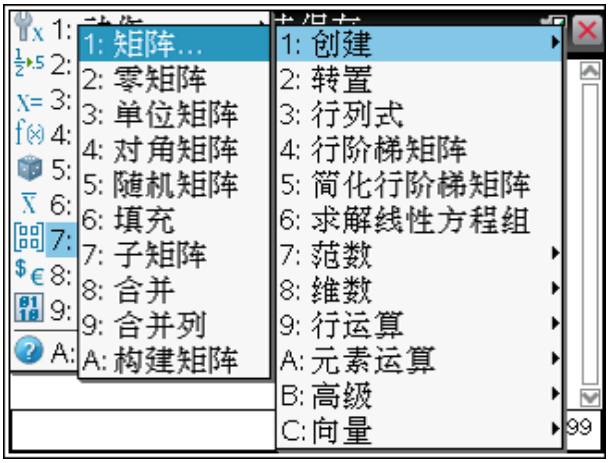

图 1

图 2

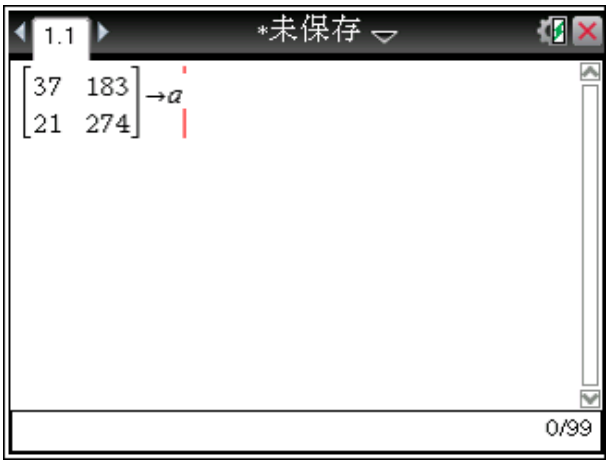

图 3

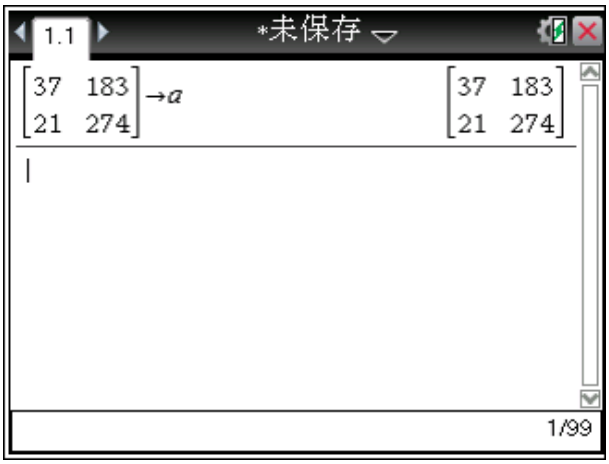

图 4

（2）进行 χ^2 检验

操作步骤：（菜单→6 统计→7 统计检验→8 χ^2 双因素检验 →填写观测矩阵名称

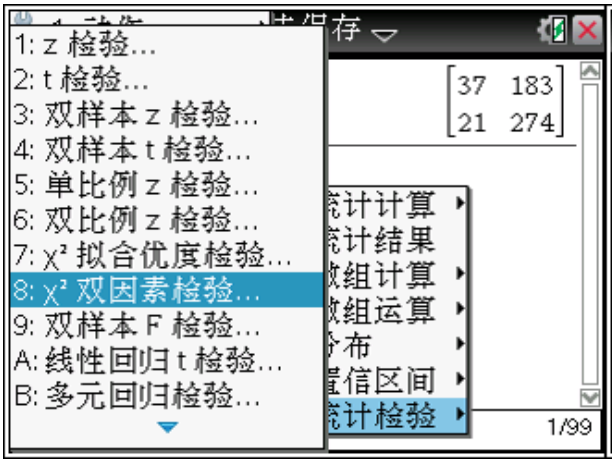

图 5

图 6

（3）统计结果输出及分析（按 var）

输出变量	说明
stat.χ^2	卡方统计：sum（实际值 - 预计值）^{2} 预计值
stat.PVal	可拒绝零假设的最小显著性水平
stat.df	卡方统计的自由度
stat.ExpMat	预期元素计数表的矩阵，假定为零假设
stat.CompMat	元素卡方统计计算值的矩阵

其中：期望矩阵 **stat.expmatrix**

$$\begin{pmatrix} \frac{(a+b)(a+c)}{n} & \frac{(a+b((b+d)}{n} \\ \frac{(c+d)(a+c)}{n} & \frac{(c+d)(b+d)}{n} \end{pmatrix} = \begin{bmatrix} 24.7767 & 195.223 \\ 33.2233 & 261.777 \end{bmatrix}$$

元素卡方矩阵 **stat.compmatrix**

$$\begin{pmatrix} \frac{\left(a-\frac{(a+b)(a+c)}{n}\right)^2}{\frac{(a+b)(a+c)}{n}} & \frac{\left(b-\frac{(a+b)(b+d)}{n}\right)^2}{\frac{(a+b)(b+d)}{n}} \\ \frac{\left(c-\frac{(c+d)(a+c)}{n}\right)^2}{\frac{(c+d)(a+c)}{n}} & \frac{\left(d-\frac{(c+d)(b+d)}{n}\right)^2}{\frac{(c+d)(b+d)}{n}} \end{pmatrix} = \begin{bmatrix} 6.03023 & 0.765324 \\ 4.49712 & 0.57075 \end{bmatrix}$$

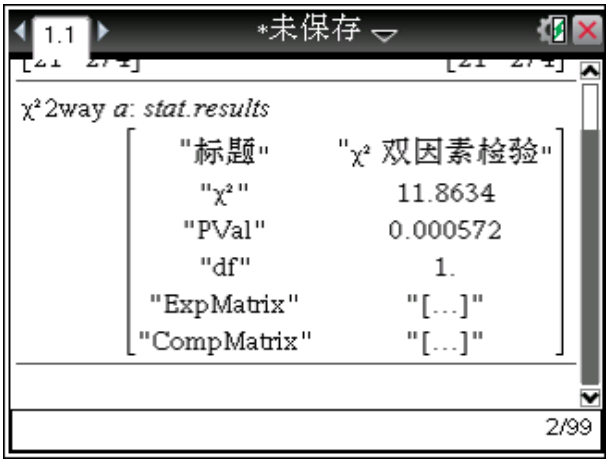

图 7

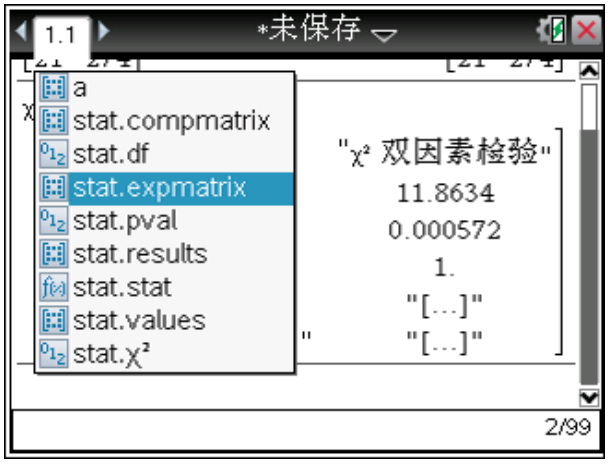

图 8

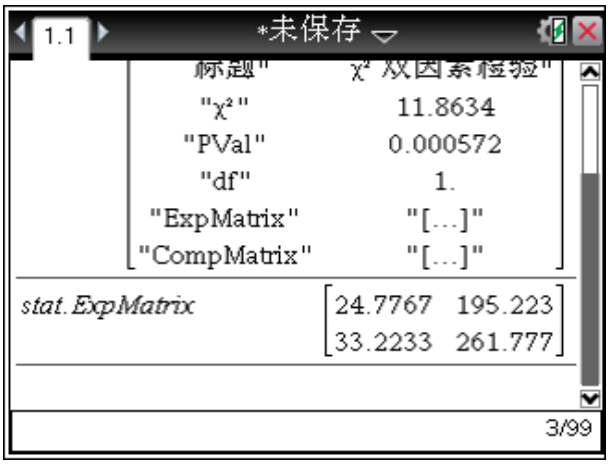

图 9

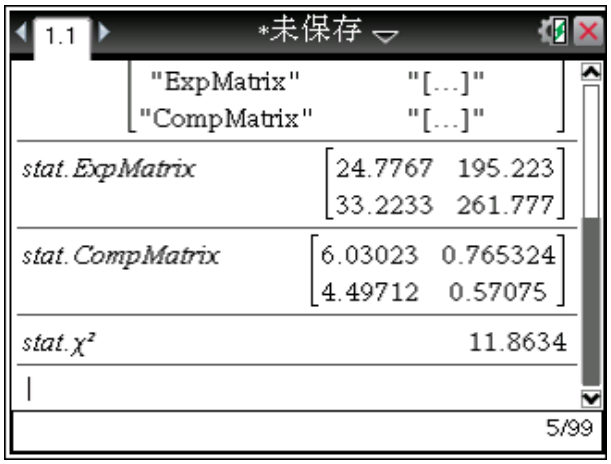

图 10

附：卡方检验表

$P(\chi^2 \geq x_0)$	0.50	0.40	0.25	0.15	0.10	0.05	0.025	0.010	0.005	0.001
x_0	0.455	0.708	1.323	2.072	2.706	3.841	5.024	6.635	7.879	10.828

五. 练习

（1）为了研究色盲与性别的关系，调查了 1000 人，调查结果如下表所示：

	男	女
正常	442	514
色盲	38	6

根据上述数据试问色盲与性别是否是相互独立的？

（2）生物学上对于人类眼睛的颜色是否与头发的颜色有关进行了调研，以下是一次调查结果。

根据上述数据检验眼睛的颜色是否与头发的颜色有关？

眼睛 / 头发	眼睛颜色			
头发颜色	蓝色	棕色	绿色	淡褐色
黑色	20	68	5	15
金色	94	7	16	10
棕色	84	119	29	54
红色	17	26	14	14

注：

1. 统计量 $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$，其中每个单元格频数的期望值为 $E_{ij} = \frac{R_i}{n}\times\frac{C_j}{n}\times n = \frac{R_i \times C_j}{n}$.

2.在零假设成立时，统计量 χ^2 近似服从自由度为(r-1) × (s-1)的 χ^2 分布。当统计量 χ^2 值很大（或 p 值很小）时，就可以拒绝零假设，认为这两个变量不相互独立.