生物统计学（VII）
第五章：χ^2 检验

西安交通大学数学与统计学院

May, 2018

χ^2 检验主要有三种用途

- 样本方差的同质性检验（homogenity test ）
- 适合性检验（compatibility test）与独立性检验（independent test）
  - 适用于离散性资料的假设检验
  - 基本原理是通过 χ^2 值的大小来检验观测值与理论值之间的偏离程度
- 适合性检验：比较观测值与理论值是否符合的假设检验
- 独立性检验：判断两个或两个以上因素之间是否具有关联关系的假设检验，常用列联表进行检验。

- 对计数资料和属性资料，即离散型资料的假设检验，可以采用二项分布概率计算，也可以采用将次数转化为频率再计算其概率来进行
- 还可以采用 χ^2 检验
- 计数资料和属性资料的 χ^2 检验一般有两种类型，即适合性检验与独立性检验

- χ^2 的原意是相互独立的多个正态离差平方值的和，即

$$\chi^2 = \sum_{i=1}^{k}\left(\frac{x_i-\mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{k}(x-\mu)^2$$

- 当用样本平均数 $\bar{x}$ 估计总体平均数 μ 时，则有

$$\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{k}(x-\bar{x})^2$$

样本方差

$$s^2 = \frac{\sum_{i=1}^{k}(x-\bar{x})^2}{k-1}$$

上式可变为

$$\chi^2 = \frac{(k-1)s^2}{\sigma^2}$$

该式中，分子表示样本的离散程度，分母表示总体方差，χ^2 服从自由度为 $k-1$ 的 χ^2 分布。

- 对计数资料或属性资料进行 χ^2 检验，其基本原理是应用观测值 (observed value, O) 与理论值 (expected value, E) 之间的偏离程度来决定其 χ^2 值的大小。
- 观测值与理论值之间偏差越大，越不符合；偏差越小，越趋于符合；若这两个值安全相等，表明他们完全符合
- 在计算观测值 O 与理论值 E 之间的符合程度时，就简单的是比较两者差数的大小，但是由于 $O-E$ 有正有负，因此 $\sum(O-E)$ 趋近于零，不能真实反映两者的差值
- 采用 $\sum(O-E)^2$，可以解决正负抵消的问题
    - 观测值与理论值相差越大，则 $\sum(O-E)^2$ 也越大，反之亦然。

$\sum(O-E)^2$ 似乎可以度量观测值与理论值的相差程度，实际上这个绝对差异数还不足以表示相差程度

例

在某动物育种试验中，两次试验得到 F_2 代数据资料如下表

表: 表 5-1: 某动物育种试验 F_2 代数据资料

	观测值	理论值	$O-E$
试验一	204	200	4
试验二	24	28	-4

- 显然两次试验的 $(O-E)^2$ 都是 16，但二者不能等量齐观。
- 对于 k 组资料，采用

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

使其转化为相对比值，这个值随着 k 的增加渐近于自由度为 $df=k-1$ 的 χ^2 值，即

$$\chi^2=\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}$$

- 由上式可知，χ^2 的最小值为 0，且随着 χ^2 值的增大，观测值与理论值的符合度越来越小，所以 χ^2 的分布是由 0 到无穷大的变数
- 实际上其符合程度由 χ^2 概率决定，由 χ^2 值表可知，在一定自由度下，χ^2 的值与概率 P 成反比
- 由此，可由 χ^2 分布对计数资料或属性资料进行假设检验

χ^2 检验的步骤如下

- 提出假设：包括无效假设H_0: 观测值与理论值的差异由抽样误差引起，即观测值 $=$ 理论值及备择假设H_A：观测值与理论值的差值不等于 0，即观测值 $\neq$ 理论值
- 确定显著水平α。
- 计算样本的 χ^2：求得各个理论值 E_i，并根据各观测值 O_i，代入上式，计算样本的统计数 χ^2
- 进行统计推断：根据 $df = k - 1$，从附表查出 χ^2_α 值，
  - 如果实得 $\chi^2 < \chi^2_\alpha$，则 $P > \alpha$，应接受 H_0，否定 H_A，表明在 α 显著水平下观测值与理论值差异不显著，二者之间的差异是由抽样误差引起；
  - 如果实得 $\chi^2 > \chi^2_\alpha$，则 $P < \alpha$，应否定 H_0，接受 H_A，表明在 α 显著水平下观测值与理论值差异显著，二者之间的差异是真实存在的；

由于 χ^2 分布是连续的，而计数资料或属性资料是离散的，故所得的 χ^2 值是一个近似值。为了使离散型变量的计算结果与连续型变量 χ^2 分布的概率相吻合，在计算 χ^2 时应注意一下几个问题

- 任何一组的理论值 E_i 都必须大于 5，如果 $E_i \leq 5$，统计量会明显偏离 χ^2 分布，则需要并组或增大样本容量，以满足 $E_i > 5$
- 在自由度 $df = 1$ 时，需进行连续性矫正，其矫正的 χ^2 为

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

  对同一资料，进行矫正的 χ_c^2 值要比未矫正的 χ^2 小
- 当自由度 $df \geq 2$ 时，由于 χ_c^2 与 χ^2 相差不大，所以一般不再进行连续性矫正

- 比较观测值与理论值是否符合的假设检验称为适应性检验 (compatibility test)。
- 这种方法是对样本的理论值先通过一定的理论分布推算出来，然后用观测值与理论值比较，从而得出观测值与理论值之间是否吻合的结论，因此适合性检验也称为吻合性检验或拟合优度检验（goodness of fit test）
- 在遗传学上，常用 χ^2 检验来测定所得的结果是否符合孟德尔分离规律、自由组合规律等
- 许多与已有理论比率进行比较的资料，也可用 χ^2 做适合性检验。

做适合性检验的时候，可提出无效假设 $H_0: O-E=0$，即认为观测值与理论值之间没有差异，再计算样本的 χ^2 值，根据规定的显著性水平 α 和自由度 df，从 χ^2 值表中查出 χ^2_α，当 $\chi^2 > \chi^2_\alpha$ 时，拒绝 H_0，接受 H_A；当 $\chi^2 < \chi^2_\alpha$ 时，接受 H_0。

例 5.1

有一个鲤鱼遗传试验，以荷包红鲤（红色，隐性）与湘江野鲤（青灰色，显性）杂交，其 F_2 代获得如下表所列的体色分离尾数，问这一资料的实际观测值是否符合孟德尔一对等位基因的遗传规律，即鲤鱼体色青：红 =3:1?

表: 表 5-2：鲤鱼遗传试验 F_2 代观测结果

体色	青灰色	红色	总数
F_2 代观测尾数/尾	1503	99	1602

本例为判断实际观测值与理论比率是否相符的问题，属于典型的适合性检验问题

(1) 提出假设

- H_0 鲤鱼体色 F_2 代性状分离符合 3:1 比率
- H_A 鲤鱼体色 F_2 代性状分离不符合 3:1 比率

(2) 取显著性水平 $\alpha = 0.01$

(3) 计算统计数 χ^2：由于该资料只有 $k = 2$ 组，故自由度为 1，因而计算 χ^2 时需要进行连续性矫正

- 在无效假设 H_0 正确的前提下，青灰色鲤鱼的理论值为 $E_1 = 1602 \times \frac{3}{4} = 1201.5$ 红色鲤鱼的理论值为 $E_2 = 1602 \times \frac{1}{4} = 400.5$
- 将实际观测值与理论值代入公式计算有

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(|O_i - E_i| - 0.5)^2}{E_i} = 301.63$$

(4) 查 χ^2 表，当 $df = 1$，$\chi^2_{0.01} = 6.63$。现实得 $\chi_c^2 = 301.63$，远大于 $\chi^2_{0.01}$，故否定 H_0，接受 H_A，即鲤鱼体色 F_2 代性状分离不符合 3:1 比率

遗传学中，有许多显性、隐性比率可以划分为两组的资料，如欲测其与某种理论比率的适合性，则 χ^2 值也可用下表来计算

表: 表 5-3：检验两组资料与某种理论比率符合度的 χ^2 值公式

理论比率（显性：隐性）	χ^2 值计算公式
1：1	$\frac{(\lvert A-a\rvert-1)^2}{n}$
2：1	$\frac{(\lvert A-2a\rvert-1.5)^2}{2n}$
3：1	$\frac{(\lvert A-3a\rvert-2)^2}{3n}$
15：1	$\frac{(\lvert A-15a\rvert-8)^2}{15n}$
9：7	$\frac{(\lvert 7A-9a\rvert-8)^2}{63n}$
r：1	$\frac{(\lvert A-ra\rvert-\frac{r+1}{2})^2}{rn}$
r：m	$\frac{(\lvert mA-ra\rvert-\frac{r+m}{2})^2}{rmn}$

注：A 为显性实际观测值，a 为隐性实际观测值，$n=A+a$.

例 5.2-5.3

大豆花色的研究

豌豆黄绿性状的研究

表: 表 5-4：豌豆杂交试验 F_2 代数据资料计算表

项目	黄圆	黄皱	绿圆	绿皱
观测值	315	101	108	32
理论频数	$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$
理论数	312.75	104.25	104.25	34.75
$O-E$	2.25	-3.25	3.75	-2.75
$\frac{(O-E)^2}{E}$	0.016	0.101	0.135	0.218

- 对于资料组多于两组的 χ^2 值，还可以通过下面的简式进行计算
$$\chi^2 = \frac{1}{n}\sum \frac{O_i^2}{p_i} - n$$
其中，O_i 为第 i 组的观测值；p_i 为第 i 组的理论比率；n 为总次数。

第三节：独立性检验

- 独立性检验 (independent test) 是研究两个或以上因子彼此之间是相互独立的还是相互影响的一类统计方法
    - 检验患慢性气管炎病和吸烟有无关系，若无关系则说明两者独立；若有关系，说明吸烟能诱发气管炎病。是否显著？可用 χ^2 检验
- 具体做法
    - 提出无效假设，假设所观测的各属性之间没有关联，根据无效假设计算理论值，在一定的自由度下以给定的显著性水平作出推断，最后证明无效假设是否成立
    - 若拒绝无效假设，则说明两个因子之间的关联是显著的；若接受无效假设，则说明两个因子之间无关联，是相互独立的。
- 独立性检验的形式有多种，常利用列联表 (contingency table) 进行检验。

一：2×2 列联表的独立性检验

- 设 A, B 是一个随机试验中的两个事件，其中 A 可能出现 r_1, r_2 个结果，B 可能出现 c_1, c_2 两个结果
- 两个因子相互作用形成 4 格数，分别以 $O_{11}, O_{12}, O_{21}, O_{22}$ 表示如下表。

表: 表 5-5: 2×2 列联表的一般形式

$A(i)$	$B(j)$ c_1	c_2	总和
r_1	O_{11}.	O_{12}	$R_1 = O_{11} + O_{12}$
r_2	O_{21}.	O_{22}	$R_2 = O_{21} + O_{22}$
总和	$C_1 = O_{11} + O_{21}$	$C_2 = O_{12} + O_{22}$	T

2×2 列联表的 χ^2 检验需要以下步骤：

(1) 提出无效假设 H_0 事件 A 与事件 B 无关，即事件 A 与 B 相互独立；H_A 事件 A 与 B 有关联关系

(2) 给出显著水平 α

(3) 依据 H_0，可以推算出理论值，计算 χ^2 值

(4) 确定自由度, $df = (r-1)(c-1)$，即（行-1）*（列-1）。查附表 4 的 χ^2_α，进行推断：当 $\chi^2 > \chi^2_\alpha$ 时，则 $P < \alpha$，否定 H_0，接受 H_A；当 $\chi^2 < \chi^2_\alpha$ 时，则 $P > \alpha$，接受 H_0，否定 H_A。

例 5.4

随机抽样对吸烟人群和不吸烟人群是否患有气管炎进行了调查，其调查结果如果表 5-6 所示，试检验吸烟与患气管炎有无关联。

表: 表 5-6：不同人群患气管炎病调查资料

不同人群	患病	不患病	总和 (R_i)	患病率 (%)
吸烟人群	50(33).	250(267)	300	16.67%
不吸烟人群	5(22)	195(178)	200	2.50%
总和 (C_i)	55	445	$T = 500$	

括号内为理论值

2×2 列联表的 χ^2 值也可由下面公式计算得出

$$\chi^2 = \frac{\left(|O_{11}O_{22} - O_{12}O_{21}| - \frac{T}{2}\right)^2 T}{R_1 R_2 C_1 C_2}$$

二：$2\times c$ 列联表的独立性检验

- 生物学研究中，经常遇到的是 2×2 列联表，但有时也会遇到 $2\times c$ 列联表，即横行分为两组，纵列分为 $c\geq 3$ 组的资料，$2\times c$ 列联表的一般形式如下：

表: 表 5-7: $2\times c$ 列联表的一般形式

(i)	(j) 1	2	$\cdots$	c	总和
1	$O_{11}.$	O_{12}	$\cdots$	O_{1c}	R_1
2	$O_{21}.$	O_{22}	$\cdots$	O_{2c}	R_2
总和	C_1	C_2	$\cdots$	O_c	T

$2 \times c$ 列联表理论值的计算与 2×2 列联表一样，自由度 $df = (2-1)(c-1) = c-1$，由于 $c \geq 3$，故 $df \geq 2$，计算 χ^2 时不需要做连续性矫正

例 5.5

检测甲乙丙三种农药对烟伢的毒杀效果，结果如下表，试分析这三种农药对烟伢的毒杀效果是否一致。

表: 表 5-8：3 种农药毒杀烟伢的试验资料

毒杀效果	甲	乙	丙	总和
死亡数	37（）	49（）	23（）	109
未死亡数	150（）	100（）	57（）	307
总和	187	149	80	416

括号内应填入理论值

为计算方便，也可不计算理论值，直接代入下面公式

$$\chi^2 = \frac{T^2}{R_1 R_2}\left[\sum\left(\frac{O_{1j}^2}{C_j}\right) - \frac{R_1^2}{T}\right]$$

三：$r \times c$ 列联表的独立性检验

$r \times c$ 列联表是指 $r \geq 3, c \geq 3$ 的计数资料，$r \times c$ 列联表的一般形式如下：

表: 表 5-9: $r \times c$ 列联表的一般形式

(i)	(j) 1	2	$\cdots$	c	总和
1	$O_{11}.$	O_{12}	$\cdots$	O_{1c}	R_1
2	$O_{21}.$	O_{22}	$\cdots$	O_{2c}	R_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	$O_{r1}.$	O_{r2}	$\cdots$	O_{rc}	R_r
总和	C_1	C_2	$\cdots$	C_c	T

- $r \times c$ 列联表各项理论数的计算方法与 2×2 列联表及 $2 \times c$ 列联表一样，即
$$E_{ij} = \frac{R_i C_i}{T}$$
- 其自由度 $df = (r-1)(c-1)$，由于 $r \geq 3, c \geq 3$，所以 $df > 1$，计算 χ^2 时不需要进行连续性矫正
- 其 χ^2 值可由下式进行计算
$$\chi^2 = T\left[\sum\left(\frac{O_{ij}^2}{R_i C_i}\right)\right]$$

式中 $i = 1, 2, \cdots, r; j = 1, 2, \cdots, c$

例 5.6

某医院用碘剂治疗地方性甲状腺肿，不同年龄的治疗效果列于下表，试检验不同年龄的治疗效果有无差异。

表: 表 5-10：不同年龄用碘剂治疗地方性甲状腺肿效果比较

年龄/岁	治愈	显效	好转	无效	总和
11 ～ 30	67（）	9（）	10（）	5（）	91
31 ～ 50	32（）	23（）	20（）	4（）	79
50 以上	10（）	11（）	23（）	5（）	49
总和	109	43	53	14	219

括号内应填入理论值