



独立性检验热点题型例析

江苏省天一中学

孙承辉

独立性检验是统计中的一个重要知识点,主要是根据 2×2 列联表计算 K^2 的值,从而判断两个事件之间是相互独立还是存在某种关联性。考题的呈现形式多种多样,一般与概率统计中的其他知识交汇,体现一定的综合性。本文精选一些典型例题,希望对同学们的复习能有所帮助。

题型一、独立性检验与频率分布直方图交汇

这类题目以频率分布直方图为背景,为检验事件的独立性提供数据。解题的关键是读懂频率分布直方图,明确各小长方形的面积表示相应各组的频率,从而求出相应区间内的样本频数,并完善 2×2 列联表和计算 K^2 的值。

例 1 致敬百年,读书筑梦,某学校组织全校学生参加“学党史颂党恩,党史网络知识竞赛”活动,并从中抽取 100 名学生的竞赛成绩作为样本进行统计,得到如图 1 所

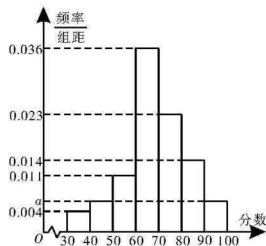


图 1

示的频率分布直方图。规定:成绩在 $[80, 100]$ 内为优秀,成绩低于 60 分为不及格。

(1)求 a 的值,并用样本估算总体,能否认为该校参加本活动的学生成绩符合“不及格的人数低于 20%”的要求?

(2)根据以上数据完成表 1 所示的 2×2 列联表,并判断是否有 99% 的把握认为此次竞赛成绩与性别有关。

表 1

	优秀	非优秀	合计
男	5		
女		35	
合计			

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

其中 $n=a+b+c+d$ 。

表 2

$P(K^2 \geq k_0)$	0.1	0.05	0.01	0.005
k_0	2.706	3.841	6.635	7.879

解析:(1)(0.004 + a + 0.011 + 0.036 + 0.023 + 0.014 + a) $\times 10 = 1$,解得 $a = 0.006$,成绩不及格的频率为 (0.004 + 0.006 + 0.011) $\times 10 = 0.21$,所以“成绩不及格”的概率估计值为 21%。

因为 21% > 20%,所以不能认为该校参加本活动的学生成绩符合“不及格的人数低于 20%”的要求。

(2)由(1)可得,成绩在 $[80, 100]$ 内的人数为 (0.014 + 0.006) $\times 10 \times 100 = 20$,即样本中成绩优秀的有 20 人,由此完成 2×2 列联表,如表 3 所示:

表 3

	优秀	非优秀	合计
男	5	45	50
女	15	35	50
合计	20	80	100

假设 H_0 :此次竞赛成绩与性别无关,则

$$K^2 = \frac{100(35 \times 5 - 15 \times 45)^2}{50 \times 50 \times 20 \times 80} = 6.25 < 6.635.$$

所以没有 99% 的把握认为此次竞赛成绩与性别有关。

点评:本题第(1)问先利用频率分布直方图的性质求出 $a = 0.006$,进而求得 60 分以下的概率估计值,即可判断;第(2)问先根据第(1)问中的结论,求得优秀的人数,再填写列联表,进而求出 K^2 的值,查表后可以判断得出没有 99% 的把握。

题型二、独立性检验与相关系数交汇

样本相关系数 r 可以反映成对样本数据

的变化特征,当 $|r|$ 越接近 1 时,成对样本数据的线性相关程度越强;当 $|r|$ 越接近 0 时,成对样本数据的线性相关程度越弱。解决与相关系数有关的独立性检验问题时,要理解 r 和 K^2 的计算公式中各个数据的含义。

例 2 共享汽车,是指许多人合用一辆车,即开车人对车辆只有使用权,而没有所有权,有点类似于在租车行业里的短时间的租车。它手续简便,打个电话或通过网上就可以预订订车。某市为了了解不同年龄的人对共享汽车的使用体验,随机选取了 100 名使用共享汽车的体验者,让他们根据体验效果进行评分。

(1)设消费者的年龄为 x ,对共享汽车的体验评分为 y 。若根据统计数据,用最小二乘法得到 y 关于 x 的线性回归方程为 $\hat{y} = 1.5x + 15$,且年龄 x 的方差为 $s_x^2 = 9$,评分 y 的方差为 $s_y^2 = 25$ 。求 y 与 x 的相关系数 r ,并据此判断对共享汽车使用体验的评分与年龄的相关性强弱(当 $|r| \geq 0.75$ 时,认为相关性强,否则认为相关性弱)。

(2)现将 100 名消费者的年龄划分为“青年”和“中老年”,评分划分为“好评”和“差评”,整理得到表 4 的一些数据,请将表 4 所示的 2×2 列联表补充完整,并判断是否有 99.9%的把握认为对共享汽车的评价与年龄有关。

表 4

	好评	差评	合计
青年	16		
中老年		12	
合计		44	100

附参考公式:回归直线 $\hat{y} = \hat{b}x + \hat{a}$ 的斜

$$\text{率 } \hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}; \text{ 相关系数 } r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

独立性检验中的 $K^2 =$

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

临界值表(表 5):

表 5

$P(K^2 \geq k_0)$	0.05	0.01	0.001
k_0	3.841	6.635	10.828

解析:(1)因为 $s_x^2 = \frac{\sum_{i=1}^{100} (x_i - \bar{x})^2}{100} = 9$,所以 $\sum_{i=1}^{100} (x_i - \bar{x})^2 = 900$ 。

因为 $s_y^2 = \frac{\sum_{i=1}^{100} (y_i - \bar{y})^2}{100} = 25$,所以 $\sum_{i=1}^{100} (y_i - \bar{y})^2 = 2500$ 。

因为 $\hat{b} = \frac{\sum_{i=1}^{100} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{100} (x_i - \bar{x})^2} = 1.5$,所以 $\sum_{i=1}^{100} (x_i - \bar{x})(y_i - \bar{y}) = 1.5 \times \sum_{i=1}^{100} (x_i - \bar{x})^2 = 1.5 \times 900 = 1350$ 。

所以相关系数 $r = \frac{\sum_{i=1}^{100} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{100} (x_i - \bar{x})^2 \sum_{i=1}^{100} (y_i - \bar{y})^2}} = \frac{1350}{\sqrt{900 \times 2500}} = \frac{1350}{30 \times 50} = 0.9$ 。

因为 $0.9 > 0.75$,所以可判断对共享汽车使用体验的评分与年龄的相关性很强。

(2)根据题意可得完整的 2×2 列联表,如表 6 所示:

表 6

	好评	差评	合计
青年	16	32	48
中老年	40	12	52
合计	56	44	100

因为 $K^2 = \frac{100(16 \times 12 - 32 \times 40)^2}{48 \times 52 \times 56 \times 44} \approx 19.25 > 10.828$,所以有 99.9%的把握认为对共享汽车的评价与年龄有关。

点评:第(1)问根据公式求出相关系数

r , 计算过程中注意整体代入, 再与 0.75 比较即可得出结论; 第(2)问直接根据已有数据即可完成已知列联表, 再根据公式求出 K^2 的值, 然后对照临界值表即可得出结论。

题型三、独立性检验与概率交汇

统计与概率的相关知识密不可分, 独立性检验问题经常与概率交汇在一起, 重点考查古典概型、随机变量的概率分布、数学期望等知识。

例 3 某种疾病可分为 A, B 两种类型, 为了解该疾病的类型与患者性别是否相关, 在某地区随机抽取了若干名该疾病的患者进行调查, 发现女性患者人数是男性患者的 2 倍, 男性患 A 型疾病的人数占男性患者的 $\frac{5}{6}$, 女性患 A 型疾病的人数占女性患者的 $\frac{1}{3}$ 。

(1) 若本次调查得出“在犯错误的概率不超过 0.005 的前提下认为‘所患疾病的类型’与‘性别’有关”的结论, 试问: 被调查的男性患者至少有多少人?

(2) 某团队进行预防 A 型疾病的疫苗的研发试验, 试验期间至多安排 2 个周期接种疫苗, 每人每个周期接种 3 次, 每次接种费用为 $m(m>0)$ 元。该团队研发的疫苗每次接种后产生抗体的概率为 $p(0<p<1)$, 如果一个周期内至少 2 次出现抗体, 则该周期结束后终止试验, 否则进入第二个周期。若 $p=\frac{2}{3}$, 试验人数为 1 000 人, 试估计该试验用于接种疫苗的总费用。

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

其中 $n=a+b+c+d$ 。

表 7

$P(K^2 \geq k_0)$	0.05	0.01	0.005	0.001
k_0	3.841	6.635	7.879	10.828

解析: (1) 设男性患者有 x 人, 则女性患者有 $2x$ 人, 整理可得完整的 2×2 列联表, 如表 8 所示:

表 8

	A 型病	B 型病	合计
男	$\frac{5x}{6}$	$\frac{x}{6}$	x
女	$\frac{2x}{3}$	$\frac{4x}{3}$	$2x$
合计	$\frac{3x}{2}$	$\frac{3x}{2}$	$3x$

假设 H_0 : 患者所患疾病类型与性别之间无关联, 根据列联表中的数据, 经计算得到

$$K^2 = \frac{3x \left(\frac{5x}{6} \cdot \frac{4x}{3} - \frac{x}{6} \cdot \frac{2x}{3} \right)^2}{\frac{3x}{2} \cdot \frac{3x}{2} \cdot 2x \cdot x} = \frac{2x}{3}.$$

要使在犯错误的概率不超过 0.005 的前提下认为“所患疾病类型”与“性别”有关, 则 $\frac{2x}{3} > 7.879$, 解得 $x > 11.8185$ 。

因为 $\frac{x}{6} \in \mathbf{Z}$, $\frac{x}{3} \in \mathbf{Z}$, 所以 x 的最小整数值为 12, 故男性患者至少有 12 人。

(2) 设该试验每人的接种费用为 ξ 元, 则 ξ 的所有可能取值为 $3m, 6m$ 。

则 $P(\xi=3m) = C_3^2 p^2 (1-p) + p^3 = -2p^3 + 3p^2$, $P(\xi=6m) = 1 + 2p^3 - 3p^2$, 所以 $E(\xi) = 3m \cdot (-2p^3 + 3p^2) + 6m \cdot (1 + 2p^3 - 3p^2) = 3m(2p^3 - 3p^2 + 2)$ 。

因为 $p=\frac{2}{3}$, 试验人数为 1 000 人, 所以估计该试验用于接种疫苗的总费用为 $1\,000E(\xi)$, 即 $1\,000 \times 3m \times \left[2 \times \left(\frac{2}{3} \right)^3 - 3 \times \left(\frac{2}{3} \right)^2 + 2 \right] = \frac{34\,000}{9}m$ (元)。

点评: 第(1)问根据数据设男性患者有 x 人, 则女性患者有 $2x$ 人, 即可得到 2×2 列联表, 计算出 K^2 的值, 从而得到不等式, 求出 x 的取值范围是 $x > 11.8185$, 同时注意到 x 必须是 6 的倍数, 所以 x 的最小整数值为 12。第(2)问设该试验每人的接种费用为 ξ 元, 则 ξ 的可能取值为 $3m, 6m$, 求出所对应的概率, 即可求出数学期望, 再由 $p=\frac{2}{3}$ 和试验人数为 1 000 人, 求出总费用的期望值。

(责任编辑 王福华)