独立性检验考向解析与备考建议

秦文波

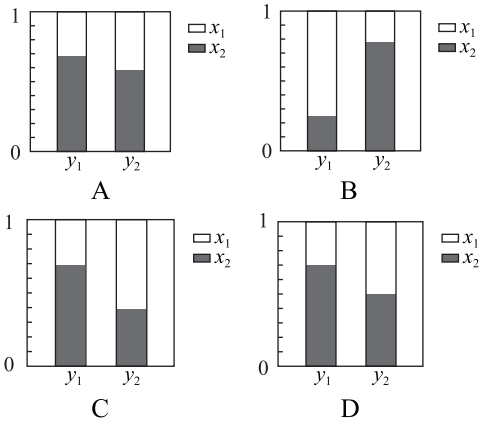
（重庆市璧山区教师进修学校）

独立性检验是高中阶段所学的一种检验方法，是近年来高考的考查热点.为了更好地复习备考，本文就独立性检验的考向进行诠释，并结合新课标要求给出备考建议.

1 考向解析

1.1 等高条形图分析

例1 在下面频率等高条形图中，两个分类变量 x 与 y 关系最强的是（　　）.



解析 由图可知，在四个选项中，选项 B 中 y_1，y_2 的高度差异最大，故 x 与 y 这两个分类变量之间关系最强，选 B.

点评 由于 $\frac{a}{a+b}$ 与 $\frac{c}{c+d}$ 相差越大，这两个分类变量的关系越强，故在频率等高条形图中下方颜色区域的高度相差越大，则这两个分类变量间的关系越强.

1.2 列联表分析

例2 假设两个分类变量 x 与 y 的 2×2 列联表如表 1 所示.

表 1

	y_1	y_2
x_1	a	b
x_2	c	d

对于以下数据，对同一样本能说明 x 与 y 有关系的可能性最大的一组为（　　）.

A. $a=20,b=30,c=40,d=50$

B. $a=50,b=30,c=30,d=40$

C. $a=30,b=60,c=20,d=50$

D. $a=50,b=30,c=40,d=30$

解析 对于 A，$|ad-bc|=200$；对于 B，$|ad-bc|=1\,100$；对于 C，$|ad-bc|=300$；对于 D，$|ad-bc|=300$.显然 B 中 $|ad-bc|$ 最大，该组数据能说明 x 与 y 有关系的可能性最大，故选 B.

点评 对于同一样本，在 2×2 列联表中，$|ad-bc|$ 的值越大，两个分类变量有关系的可能性就越大，反之则越小，当 $|ad-bc|$ 的值为 0 时，两个分类变量无关系，即独立.

1.3 统计量 K^2 的计算

例3 在研究色盲与性别的关系调查中，调查了男性 50 人，其中有 20 人患色盲，调查的 60 个女性中 15 人患色盲，则变量 K^2 的值约为（　　）.

A. 1.60　　B. 2.83　　C. 2.712　　D. 6.004

解析 列出 2×2 列联表如表 2 所示.

表 2

	患色盲	不患色盲	小计
男性	20	30	50
女性	15	45	60
小计	35	75	110

$$K^2=\frac{110\times(20\times45-15\times30)^2}{35\times75\times60\times50}\approx2.83,$$

故选 B.

点评 求 K^2 的观测值关键是完善 2×2 列联表，并能准确理解和解读表中各部分的意义.

1.4 独立性检验的概念及辨析

例4 有关独立性检验的四个命题，其中为假命题的是（　　）.

A. 两个分类变量的 2×2 列联表中，对角线上数据的乘积相差越大，说明这两个变量有关系的可能性就越大

B. 对分类变量 X 与 Y 的随机变量 K^2 的观测值 k 来说，k 越小，“X 与 Y 有关系”的可信程度越小

C. 从独立性检验可知：有 95% 把握认为秃顶与患心脏病有关，我们说某人秃顶，那么他有 95% 可能患有心脏病

D. 从独立性检验可知：有 99% 的把握认为吸烟与患肺癌有关，是指在犯错误的概率不超过 1% 前提下认为吸烟与患肺癌有关

解析 对于 C，从独立性检验可知：有 95% 的把握认为秃顶与患心脏病有关，不是说某人秃顶，那么他有 95% 的可能患有心脏病，C 错误，故选 C.

点评 独立性检验得到的结果可以用多大的把握认为两个分类变量之间有关系来呈现，并不能得到像数值变量那种确定的函数关系.

1.5 独立性检验解决实际问题

例 5 （2022 年全国甲卷文 17，节选）甲、乙两城之间的长途客车均由 A 和 B 两家公司运营，为了解这两家公司长途客车的运行情况，随机调查了甲、乙两城之间的 500 个班次，得到列联表（如表 3）.

表 3

	准点班次数	未准点班次数
A	240	20
B	210	30

能否有 90% 的把握认为甲、乙两城之间的长途客车是否准点与客车所属公司有关？

解析 $K^2=\dfrac{500\times(240\times30-210\times20)^2}{260\times240\times450\times50}\approx3.205>2.706$. 根据临界值表可知，有 90% 的把握认为甲、乙两城之间的长途客车是否准点与客车所属公司有关.

点评 解答独立性检验实际应用问题需要做好三个方面：一是准确解读 2×2 列联表，完善表中相关数据；二是正确利用公式求出 K^2 的观测值，并与临界值比较；三是正确理解“犯错误的概率”和“多大的把握”的含义.

1.6 独立性检验与其他知识交会问题

例 6 为了检测某种抗病毒疫苗的免疫效果，需要进行动物与人体试验，研究人员将疫苗注射到 200 只小白鼠体内，一段时间后测量小白鼠的某项指标值，按 $[0,20)$，$[20,40)$，$[40,60)$，$[60,80)$，$[80,100]$ 分组，绘制频率分布直方图如图 1 所示，试验发现小白鼠体内产生抗体的共有 160 只，其中该项指标值不小于 60 的有 110 只. 假设小白鼠注射疫苗后是否产生抗体相互独立.

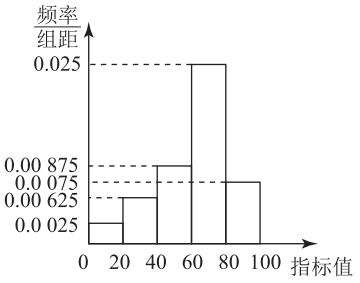


图 1

（1）请根据 $\alpha=0.05$ 的独立性检验，判断能否认为注射疫苗后小白鼠产生抗体与指标值不小于 60 有关.

（2）为检验疫苗二次接种的免疫抗体性，对第一次注射疫苗后没有产生抗体的 40 只小白鼠进行第二次注射疫苗，结果又有 20 只小白鼠产生抗体.

（ⅰ）用频率估计概率，求一只小白鼠注射 2 次疫苗后产生抗体的概率 p；

（ⅱ）以（ⅰ）中确定的概率 p 作为人体注射 2 次疫苗后产生抗体的概率，进行人体接种试验，记 n 个人注射 2 次疫苗后产生抗体的数量为随机变量 X. 试验后统计数据显示，当 $X=90$ 时，$P(X)$ 取最大值，求参加人体接种试验的人数 n 及 $E(X)$.

解析 （1）由频率分布直方图，知 200 只小白鼠按指标值分布如下.

在 $[0,20)$ 内有 $0.002\,5\times20\times200=10$ 只；

在 $[20,40)$ 内有 $0.006\,25\times20\times200=25$ 只；

在 $[40,60)$ 内有 $0.008\,75\times20\times200=35$ 只；

在 $[60,80)$ 内有 $0.025\times20\times200=100$ 只；

在 $[80,100]$ 内有 $0.007\,5\times20\times200=30$ 只.

由题意，有抗体且指标值小于 60 的有 50 只；而指标值小于 60 的小白鼠共有 $10+25+35=70$ 只，所以指标值小于 60 且没有抗体的小白鼠有 20 只. 同理，指标值不小于 60 且没有抗体的小白鼠有 20 只，故列联表（单位：只）如表 4 所示.

表 4

	指标小于 60	指标不小于 60	合计
有抗体	50	110	160
没有抗体	20	20	40
合计	70	130	200

零假设为 H_0：注射疫苗后小白鼠产生抗体与指标值不小于 60 无关联.

根据列联表中数据，得

$$K^2=\frac{200\times(50\times20-20\times110)^2}{160\times40\times70\times130}\approx$$

$$4.945>3.841=x_{0.05}.$$

根据 $\alpha=0.05$ 的独立性检验，推断 H_0 不成立，即认为注射疫苗后小白鼠产生抗体与指标值不小于 60 有关，此推断犯错误的概率不大于 0.05.

(2)(ⅰ)令事件 A 为“小白鼠第一次注射疫苗产生抗体”，事件 B 为“小白鼠第二次注射疫苗产生抗体”，事件 C 为“小白鼠注射 2 次疫苗后产生抗体”.

记事件 A,B,C 发生的概率分别为 $P(A)$，$P(B)$，$P(C)$，则

$$P(A)=\frac{160}{200}=0.8,$$

$$P(B)=\frac{20}{40}=0.5,$$

$P(C)=1-P(\overline{A})P(\overline{B})=1-0.2\times0.5=0.9$，所以一只小白鼠注射 2 次疫苗后产生抗体的概率为 0.9.

(ⅱ)由题意，知随机变量 $X\sim B(n,0.9)$，则

$P(X=k)=C_n^k\times0.9^k\times0.1^{n-k}(k=0,1,2,\cdots,n)$.

因为 $P(X=90)$ 最大，所以

$$\begin{cases}C_n^{90}\times0.9^{90}\times0.1^{n-90}\geqslant C_n^{91}\times0.9^{91}\times0.1^{n-91},\\ C_n^{90}\times0.9^{90}\times0.1^{n-90}\geqslant C_n^{89}\times0.9^{89}\times0.1^{n-89},\end{cases}$$

解得 $99\leqslant n\leqslant\frac{901}{9}$，因为 n 是整数，所以 $n=99$ 或 100，故接受接种试验的人数为 99 或 100.

当接种人数为 99 时，$E(X)=np=99\times0.9=89.1$；当接种人数为 100 时，$E(X)=np=100\times0.9=90$.

点评 独立性检验容易和古典概型、随机变量的分布列及其期望、一元线性回归等综合考查，需要学生牢固掌握各部分的知识、方法和思想.

2 备考建议

2.1 要准确理解基础知识

本部分内容的基础知识包括：2×2 列联表、等高条形图、统计量 K^2、独立性检验的概念、基本思想以及操作步骤等，这些都是高考重点考查的内容，需要准确理解并牢固掌握.

2.2 要熟练掌握基本方法

学生要能快速且准确地列出 2×2 列联表并画出等高条形图，要能根据等高条形图准确列出 2×2 列联表，要熟练独立性检验的操作步骤和结果的表述方式.

2.3 要不断提高运算能力

关于统计量 K^2 的观测值的获得需要较大的运算量，我们平时要提高自己的运算求解能力，计算时最好能提取公因数或借助平方差公式因式分解后再约分化简求解，结果最好先以分式形式呈现，再根据题目要求保留相应小数点位数.

2.4 要重视练习的质和量

平时的练习题要在重视基础的前提下适当顾及广度、深度和难度，练习题尽量选取各省市或名校模拟试题以及高考真题.

2.5 要深入理解独立性检验的基本思想

独立性检验本质上是对两个分类变量独立性的假设检验，基本思路：先假设两个分类变量独立，再在给定显著性水平下计算统计量 K^2 的观测值并与临界值比较，最后根据比较结果作出拒绝假设或接受假设的概率解释.

2×2 列联表、等高条形图和统计量 K^2 的观测值可以判断两个分类变量是否有关系.2×2 列联表对角线乘积相差越大，则两个分类变量有关系的可能性越大；等高条形图下方颜色高度相差越大，则两个分类变量有关系的可能性越大；统计量 K^2 的观测值越大，则两个分类变量有关系的可能性越大.

在给定显著性水平 α 的条件下，利用统计量 K^2 的观测值 k 与 α 对应的临界值 k_0 的大小关系可以从概率的角度判断两个分类变量是否有关系.若 $k\geqslant k_0$，则可表述为能在犯错误的概率不超过 α 的前提下认为这两类分类变量有关系；有 $1-\alpha$(百分比)的把握认为这两个分类变量有关系.若 $k<k_0$，则可表述为不能在犯错误的概率不超过 α 的前提下判定这两个分类变量有关；没有 $1-\alpha$(百分比)的把握认为这两个分类变量有关.

3 备考练习

练习 1 (多选题)为了增强学生的身体素质，某校将冬天长跑作为一项制度固定下来，每天大课间例行跑操.为了调查学生喜欢跑步是否与性别有关，研究人员随机调查了相同人数的男、女学生，发现男生中有 80%喜欢跑步，女生中有 40%不喜欢跑步，且有 95%的把握判断喜欢跑步与性别有关，但没有 99%的把握判断喜欢跑步与性别有关，则被调查的男、女学

生的总人数可能为(　　).

A. 120　　B. 130　　C. 240　　D. 250

答案　AB.

练习 2　某种常见疾病可分为Ⅰ,Ⅱ两种类型.为了了解该疾病类型与地域、初次患该疾病的年龄(以下简称初次患病年龄)的关系,在甲、乙两个地区共随机抽取 100 名患者调查其疾病类型及初次患病年龄,得到的数据如表 5 所示.

表 5

初次患病年龄/岁	甲地Ⅰ型患者/人	甲地Ⅱ型患者/人	乙地Ⅰ型患者/人	乙地Ⅱ型患者/人
[10,20)	8	1	5	1
[20,30)	4	3	3	1
[30,40)	3	5	2	4
[40,50)	3	8	4	4
[50,60)	3	9	2	6
[60,70]	2	11	1	7

记初次患病年龄在[10,40)的患者为低龄患者,初次患病年龄在[40,70]的患者为高龄患者.根据表 5 中数据,解决以下问题:

(1)将以下列联表(表 6 和表 7)补充完整,并判断地域、初次患病年龄这两个变量中哪个变量与该疾病的类型有关联的可能性更大(直接写出结论,不必说明理由).

表 6

	Ⅰ型患者	Ⅱ型患者	总计
甲地			
乙地			
总计			100

表 7

	Ⅰ型患者	Ⅱ型患者	总计
低龄			
高龄			
总计			100

(2)记(1)中与该疾病的类型有关联的可能性更大的变量为 X.问:是否有 99%的把握认为该疾病的类型与 X 有关?

答案　(1)列联表略,初次患病年龄与该疾病的类型有关联的可能性更大.

(2)有 99%的把握认为该疾病类型与初次患病年龄有关.

练习 3　为迎接 2022 年北京冬季奥运会,普及冬奥知识,某校开展了"冰雪答题王"冬奥知识竞赛活动.现从参加冬奥知识竞赛活动的学生中随机抽取 100 名学生,将他们的竞赛成绩(满分为 100 分)分为 6 组:[40,50),[50,60),[60,70),[70,80),[80,90),[90,100],得到如图 2 所示的频率分布直方图.

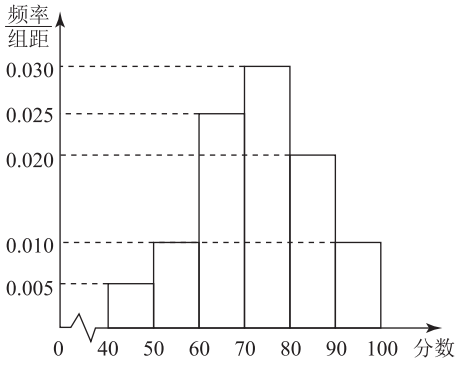


图 2

(1)估计这 100 名学生的平均成绩(同一组中的数据用该组区间的中点值为代表),并估计这 100 名学生成绩的中位数(精确到 0.01);

(2)在抽取的 100 名学生中,规定:竞赛成绩不低于 80 分为"优秀",竞赛成绩低于 80 分为"非优秀".

(ⅰ)请判断是否有 99%的把握认为"竞赛成绩是否优秀与性别有关"?

(ⅱ)求出等高条形图需要的数据,并画出等高条形图(按图 3 中"优秀"和"非优秀"所对应阴影线画),利用条形图判断竞赛成绩优秀与性别是否有关系?

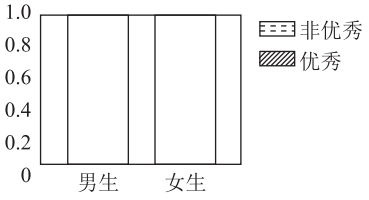


图 3

答案　(1)平均成绩 73,中位数 73.33.

(2)(ⅰ)略,没有;(ⅱ)略,有.

本文可能用到的参考公式及数据:

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

表 8

$P(K^2\geqslant k_0)$	0.10	0.05	0.025	0.010	0.005	0.001
k_0	2.706	3.841	5.024	6.635	7.879	10.828

(完)