条件概率、相互独立事件等在高考中的考查

一、山东省高考大事记：

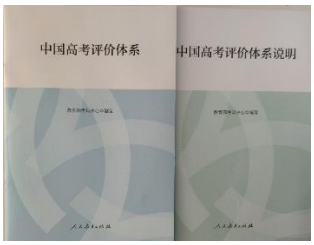
2003--2007 年前，大纲版，山东适用两省一市试验修订本教材，过度

2007 年新课标（2007 版）第一年高考山东、广东自主命题（2007-2017），海南、宁夏新课标全国卷，适用必修 5 本+选修 3 本教材，一纲多本，人教 A 版、B 版、北师大版等

2017 年新高考第一年上海、浙江，（北京、天津、山东、海南启动新课标（2017 版）老高考），老教材

2018、2019 两年山东适用新课标（2017 版）全国 1 卷（河北、河南、山东、山西、安徽、江西、湖南、湖北、福建、广东）过度，老高考，老教材

2020 年新高考全国 1 卷（仅山东适用）、2 卷（仅海南适用），老教材



2021 年新高考全国 1 卷（山东、河北、江苏、湖北、湖南、广东、福建适用）、2 卷（海南、辽宁、重庆适用），老教材

2022 年新高考全国 1 卷（山东、河北、江苏、湖北、湖南、广东、福建适用）、2 卷（海南、辽宁、重庆适用）首次新教材

2023 年新高考全国 1 卷（山东、河北、江苏、浙江、湖北、湖南、广东、福建适用）、2 卷（海南、辽宁、重庆适用）

二、2018、2019 两年新课标Ⅰ卷概率统计综合题：

（2018 年新课标Ⅰ卷山东，22 题是二选一题）新课标、老高考、老教材

20.（12 分）某工厂的某种产品成箱包装，每箱 200 件，每一箱产品在交付用户之前要对产品作检验，如检验出不合格品，则更换为合格品．检验时，先从这箱产品中任取 20 件作检验，再根据检验结果决定是否对余下的所有产品作检验，设每件产品为不合格品的概率都为 $p(0<p<1)$，且各件产品是否为不合格品相互独立．

（1）记 20 件产品中恰有 2 件不合格品的概率为 $f(p)$，求 $f(p)$ 的最大值点 p_0．

（2）现对一箱产品检验了 20 件，结果恰有 2 件不合格品，以（1）中确定的 p_0 作为 p 的值．已知每件产品的检验费用为 2 元，若有不合格品进入用户手中，则工厂要对每

件不合格品支付 25 元的赔偿费用.

（ⅰ）若不对该箱余下的产品作检验，这一箱产品的检验费用与赔偿费用的和记为 X，求 EX；

（ⅱ）以检验费用与赔偿费用和的期望值为决策依据，是否该对这箱余下的所有产品作检验？

20.解：（1）20 件产品中恰有 2 件不合格品的概率为 $f(p)=C_{20}^{2}p^{2}(1-p)^{18}$.因此

$f'(p)=C_{20}^{2}[2p(1-p)^{18}-18p^{2}(1-p)^{17}]=2C_{20}^{2}p(1-p)^{17}(1-10p)$.

令 $f'(p)=0$，得 $p=0.1$.当 $p\in(0,0.1)$ 时，$f'(p)>0$；当 $p\in(0.1,1)$ 时，$f'(p)<0$.

所以 $f(p)$ 的最大值点为 $p_0=0.1$.

（2）由（1）知，$p=0.1$.

（ⅰ）令 Y 表示余下的 180 件产品中的不合格品件数，依题意知 $Y\sim B(180,0.1)$，$X=20\times2+25Y$，即 $X=40+25Y$.

所以 $EX=E(40+25Y)=40+25EY=490$.

（ⅱ）如果对余下的产品作检验，则这一箱产品所需要的检验费为 400 元.

由于 $EX>400$，故应该对余下的产品作检验.

（2019 年新课标Ⅰ卷山东压轴题，22 题是二选一题）新课标、老高考、老教材

21. 为了治疗某种疾病，研制了甲、乙两种新药，希望知道哪种新药更有效，为此进行动物试验. 试验方案如下：每一轮选取两只白鼠对药效进行对比试验. 对于两只白鼠，随机选一只施以甲药，另一只施以乙药. 一轮的治疗结果得出后，再安排下一轮试验. 当其中一种药治愈的白鼠比另一种药治愈的白鼠多 4 只时，就停止试验，并认为治愈只数多的药更有效. 为了方便描述问题，约定：对于每轮试验，若施以甲药的白鼠治愈且施以乙药的白鼠未治愈则甲药得 1 分，乙药得 -1 分；若施以乙药的白鼠治愈且施以甲药的白鼠未治愈则乙药得 1 分，甲药得 -1 分；若都治愈或都未治愈则两种药均得 0 分. 甲、乙两种药的治愈率分别记为 α 和 β，一轮试验中甲药的得分记为 X.

（1）求 X 的分布列；

（2）若甲药、乙药在试验开始时都赋予 4 分，$p_i(i=0,1,\cdots,8)$ 表示“甲药的累计得分为 i 时，最终认为甲药比乙药更有效”的概率，则 $p_0=0$，$p_8=1$，$p_i=ap_{i-1}+bp_i+cp_{i+1}$ $(i=1,2,\cdots,7)$，其中 $a=P(X=-1)$，$b=P(X=0)$，$c=P(X=1)$．假设 $\alpha=0.5$，$\beta=0.8$．

(i)证明：$\{p_{i+1}-p_i\}\ (i=0,1,2,\cdots,7)$ 为等比数列；

(ii)求 p_4，并根据 p_4 的值解释这种试验方案的合理性．

【答案】（1）见解析；（2）（i）见解析；（ii）$p_4=\dfrac{1}{257}$．

【解析】

【分析】

（1）首先确定 X 所有可能的取值，再来计算出每个取值对应的概率，从而可得分布列；（2）（i）求解出 a,b,c 的取值，可得 $p_i=0.4p_{i-1}+0.5p_i+0.1p_{i+1}\left(i=1,2,\cdots,7\right)$，从而整理出符合等比数列定义的形式，问题得证；（ii）列出证得的等比数列的通项公式，采用累加的方式，结合 p_8 和 p_0 的值可求得 p_1；再次利用累加法可求出 p_4．

【详解】（1）由题意可知 X 所有可能的取值为：-1，0，1

$\therefore P\left(X=-1\right)=\left(1-\alpha\right)\beta$；$P\left(X=0\right)=\alpha\beta+\left(1-\alpha\right)\left(1-\beta\right)$；$P\left(X=1\right)=\alpha\left(1-\beta\right)$

则 X 的分布列如下：

X	-1	0	1
P	$\left(1-\alpha\right)\beta$	$\alpha\beta+\left(1-\alpha\right)\left(1-\beta\right)$	$\alpha\left(1-\beta\right)$

（2）$\because\alpha=0.5$，$\beta=0.8$

$\therefore a=0.5\times0.8=0.4$，$b=0.5\times0.8+0.5\times0.2=0.5$，$c=0.5\times0.2=0.1$

（i）$\because\ p_i=ap_{i-1}+bp_i+cp_{i+1}\left(i=1,2,\cdots,7\right)$

即 $p_i=0.4p_{i-1}+0.5p_i+0.1p_{i+1}\left(i=1,2,\cdots,7\right)$

整理可得：$5p_i=4p_{i-1}+p_{i+1}\left(i=1,2,\cdots,7\right)$ $\therefore p_{i+1}-p_i=4\left(p_i-p_{i-1}\right)\left(i=1,2,\cdots,7\right)$

$\therefore \{p_{i+1}-p_i\}\ (i=0,1,2,\cdots,7)$ 是以 p_1-p_0 为首项，4 为公比的等比数列

（ii）由（i）知：$p_{i+1}-p_i=(p_1-p_0)\cdot 4^i=p_1\cdot 4^i$

$\therefore p_8-p_7=p_1\cdot 4^7$，$p_7-p_6=p_1\cdot 4^6$，……，$p_1-p_0=p_1\cdot 4^0$

作和可得：$p_8-p_0=p_1\cdot(4^0+4^1+\cdots+4^7)=\dfrac{1-4^8}{1-4}p_1=\dfrac{4^8-1}{3}p_1=1$

$\therefore p_1=\dfrac{3}{4^8-1}$

$\therefore p_4=p_4-p_0=p_1\cdot(4^0+4^1+4^2+4^3)=\dfrac{1-4^4}{1-4}p_1=\dfrac{4^4-1}{3}\times\dfrac{3}{4^8-1}=\dfrac{1}{4^4+1}=\dfrac{1}{257}$

p_4 表示最终认为甲药更有效的.由计算结果可以看出，在甲药治愈率为 0.5，乙药治愈率为 0.8 时，认为甲药更有效的概率为 $p_4=\dfrac{1}{257}\approx 0.0039$，此时得出错误结论的概率非常小，说明这种实验方案合理.

【点睛】本题考查离散型随机变量分布列的求解、利用递推关系式证明等比数列、累加法求解数列通项公式和数列中的项的问题.本题综合性较强，要求学生能够熟练掌握数列通项求解、概率求解的相关知识，对学生分析和解决问题能力要求较高.

三、2021、2022 年条件概率、相互独立在新高考中的考查：

试题一：（2021 年新高考Ⅰ卷）新课标、新高考、老教材

8. 有 6 个相同的球，分别标有数字 1，2，3，4，5，6，从中有放回的随机取两次，每次取 1 个球，甲表示事件“第一次取出的球的数字是 1”，乙表示事件“第二次取出的球的数字是 2”，丙表示事件“两次取出的球的数字之和是 8”，丁表示事件“两次取出的球的数字之和是 7”，则（　　）

A. 甲与丙相互独立　　B. 甲与丁相互独立

C. 乙与丙相互独立　　D. 丙与丁相互独立

【答案】B

【解析】【分析】根据独立事件概率关系逐一判断

【详解】$P(甲)=\dfrac{1}{6}$，$P(乙)=\dfrac{1}{6}$，$P(丙)=\dfrac{5}{36}$，$P(丁)=\dfrac{6}{36}=\dfrac{1}{6}$，，

$P(甲丙)=0\neq P(甲)P(丙)$，$P(甲丁)=\dfrac{1}{36}=P(甲)P(丁)$，

$P(\text{乙丙})=\dfrac{1}{36}\neq P(\text{乙})P(\text{丙})$，$P(\text{丙丁})=0\neq P(\text{丁})P(\text{丙})$，

故选：B

【点睛】判断事件A,B是否独立，先计算对应概率，再判断$P(A)P(B)=P(AB)$是否成立

试题二：（2022年新高考Ⅰ卷）新课标、新高考、新教材

20. (12分)

一医疗团队为研究某地的一种地方性疾病与当地居民的卫生习惯(卫生习惯分为良好和不够良好两类)的关系，在已患该疾病的病例中随机调查了100例(称为病例组)，同时在未患该疾病的人群中随机调查了100人(称为对照组)，得到如下数据：

	不够良好	良好
病例组	40	60
对照组	10	90

(1) 能否有99%的把握认为患该疾病群体与未患该疾病群体的卫生习惯有差异？

(2) 从该地的人群中任选一人，A表示事件“选到的人卫生习惯不够良好”，B表示事件“选到的人患有该疾病”，$\dfrac{P(B|A)}{P(\overline{B}|A)}$与$\dfrac{P(B|\overline{A})}{P(\overline{B}|\overline{A})}$的比值是卫生习惯不够良好对患该疾病风险程度的一项度量指标，记该指标为R.

(i) 证明：$R=\dfrac{P(A|B)}{P(\overline{A}|B)}\cdot\dfrac{P(\overline{A}|\overline{B})}{P(A|\overline{B})}$；

$P(K^2\geqslant k)$	0.050	0.010	0.001
k	3.841	6.635	10.828

(ii) 利用该调查数据，给出$P(A|B)$，$P(A|\overline{B})$的估计值，并利用(i)的结果给出R的估计值.

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

【解析】(1) 假设患该疾病群体与未患该疾病群体的卫生习惯没有差异，

则$K^2=\dfrac{200(40\times90-60\times10)^2}{50\times150\times100\times100}=24>10.828$，

所以有99%的把握认为患该疾病群体与未患该疾病群体的卫生习惯有差异；

(2) (i)$R=\dfrac{P(B|A)}{P(\overline{B}|A)}\cdot\dfrac{P(\overline{B}|\overline{A})}{P(B|\overline{A})}=\dfrac{\dfrac{P(AB)}{P(A)}}{\dfrac{P(A\overline{B})}{P(A)}}\cdot\dfrac{\dfrac{P(\overline{A}\overline{B})}{P(\overline{A})}}{\dfrac{P(\overline{A}B)}{P(\overline{A})}}=\dfrac{P(AB)}{P(A\overline{B})}\cdot\dfrac{P(\overline{A}\overline{B})}{P(\overline{A}B)}$

$=\dfrac{P(AB)}{P(\overline{A}B)}\cdot\dfrac{P(\overline{A}\overline{B})}{P(A\overline{B})}=\dfrac{\dfrac{P(AB)}{P(B)}}{\dfrac{P(\overline{A}B)}{P(B)}}\cdot\dfrac{\dfrac{P(\overline{A}\overline{B})}{P(\overline{B})}}{\dfrac{P(A\overline{B})}{P(\overline{B})}}=\dfrac{P(A|B)}{P(\overline{A}|B)}\cdot\dfrac{P(\overline{A}|\overline{B})}{P(A|\overline{B})}$，

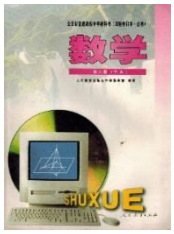
得证：

(ii) 由调查数据可知$P(A|B)=\dfrac{40}{100}=\dfrac{2}{5}$，$P(A|\overline{B})=\dfrac{10}{100}=\dfrac{1}{10}$，

则$P(\overline{A}|B)=1-P(A|B)=\dfrac{3}{5}$，$P(\overline{A}|\overline{B})=\dfrac{9}{10}$，所以$R=\dfrac{\dfrac{2}{5}}{\dfrac{3}{5}}\cdot\dfrac{\dfrac{9}{10}}{\dfrac{1}{10}}=6$.

四、相互独立事件等概念在教材中的变迁：

（一）大纲版《全日制普通高级中学教科书（试验修订本·必修）第二册下A》2001.6



第十章排列、组合和概率

10.7 相互独立事件同时发生的概率

事件 A（或 B）是否发生对事件 B（或 A）发生的概率没有影响，这样的两个事件叫做相互独立事件

独立重复试验

（二）2007 课标版《普通高中课程标准实验教科书·数学·选修 2-3·B 版》2007.4

第二章概率，

2.2 条件概率与事件的独立性

2.2.1 在已知事件 A 发生的条件下，事件 B 发生的概率叫做条件概率，用 P（B | A）表示

2.2.2 事件 A 是否发生对事件 B 发生的概率没有影响，即 P(B | A)=P(B)，这时，我们称两个事件 A，B 相互独立，并把这两个事件叫做相互独立事件

第三章统计案例

3.1 独立性检验

统计量 χ^2 推导，根据概率的统计定义，事件的概率都可用相应的频率来估计

为了把问题讨论清楚，并便于向一般情况推广，我们用字母来代替 2×2 列联表中的事件和数据，得到一张用字母来表示的 2×2 列联表，如下表所示：

	患慢性气管炎 (B)	未患慢性气管炎 (B)	合计
吸　烟(A)	n_{11}	n_{12}	n_{1+}
不吸烟(A)	n_{21}	n_{22}	n_{2+}
合　计	n_{+1}	n_{+2}	n

注

右表中 $n_{1+}=n_{11}+n_{12}$，$n_{2+}=n_{21}+n_{22}$，$n_{+1}=n_{11}+n_{21}$，$n_{+2}=n_{12}+n_{22}$，$n=n_{11}+n_{12}+n_{21}+n_{22}$.

(1) 首先，当吸烟(A)与患慢性气管炎(B)无关时，用概率方法进行推理，看看会出现什么结果.

上面的话的意思是指事件 A 与 B 独立，这时应该有

$$P(AB)=P(A)P(B).$$

成立. 我们用字母 H_0 表示上式，即

$$H_0: P(AB)=P(A)P(B),$$

并称之为统计假设，当 H_0 成立时，下面的三个式子也都成立：

$$P(\overline{A}B)=P(\overline{A})P(B),\ P(A\overline{B})=P(A)P(\overline{B}),\ P(\overline{A}\,\overline{B})=P(\overline{A})P(\overline{B}).$$

根据概率的统计定义，上面提到的众多事件的概率都可用相应的频率来估计. 例如，

$P(AB)$的估计为$\frac{n_{11}}{n}$，$P(A)$的估计为$\frac{n_{1+}}{n}$，$P(B)$的估计为$\frac{n_{+1}}{n}$……

于是$\frac{n_{11}}{n}$与$\frac{n_{1+}}{n}\cdot\frac{n_{+1}}{n}$应该很接近，$\frac{n_{12}}{n}$与$\frac{n_{1+}}{n}\cdot\frac{n_{+2}}{n}$应该很接近……

或者说，

$$\left(\frac{n_{11}}{n}-\frac{n_{1+}}{n}\cdot\frac{n_{+1}}{n}\right)^2,\ \left(\frac{n_{12}}{n}-\frac{n_{1+}}{n}\cdot\frac{n_{+2}}{n}\right)^2,\ \left(\frac{n_{21}}{n}-\frac{n_{2+}}{n}\cdot\frac{n_{+1}}{n}\right)^2,\ \left(\frac{n_{22}}{n}-\frac{n_{2+}}{n}\cdot\frac{n_{+2}}{n}\right)^2$$

应该比较小，从而

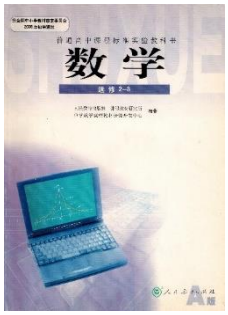
$$\frac{\left(\frac{n_{11}}{n}-\frac{n_{1+}}{n}\cdot\frac{n_{+1}}{n}\right)^2}{\frac{n_{1+}}{n}\cdot\frac{n_{+1}}{n}}+\frac{\left(\frac{n_{12}}{n}-\frac{n_{1+}}{n}\cdot\frac{n_{+2}}{n}\right)^2}{\frac{n_{1+}}{n}\cdot\frac{n_{+2}}{n}}+\frac{\left(\frac{n_{21}}{n}-\frac{n_{2+}}{n}\cdot\frac{n_{+1}}{n}\right)^2}{\frac{n_{2+}}{n}\cdot\frac{n_{+1}}{n}}+\frac{\left(\frac{n_{22}}{n}-\frac{n_{2+}}{n}\cdot\frac{n_{+2}}{n}\right)^2}{\frac{n_{2+}}{n}\cdot\frac{n_{+2}}{n}} \quad ①$$

也应该比较小.

（2）上面的表达式①就是统计中非常有用的χ^2（读作“卡方”）统计量，它可以化简为

$$\chi^2=\frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}. \quad ②$$

《普通高中课程标准实验教科书·数学·选修 2-3·A 版》2006.4



第二章随机变量及其分布

2.2.1 条件概率

一般地，设 A，B 为两个事件，且$P(A)>0$，称$P(B|A)=\frac{P(AB)}{P(A)}$为事件 A 发生的条件下，事件 B 发生的条件概率，一般把$P(B|A)$读作 A 发生的条件下 B 的概率

2.2.2 事件的相互独立性

设 A，B 为两个事件，如果$P(AB)=P(A)P(B)$，则称事件 A 与事件 B 相互独立，

$P(B|A)=P(B)$, $P(AB)=P(A)P(B|A)=P(A)P(B)$

2.2.3 独立重复试验与二项分布

第三章统计案例

3.2 独立性检验的基本思想及其初步应用

统计量K^2推导，根据概率的统计定义，频率近似于概率

H_0：吸烟与患肺癌没有关系.

用 A 表示不吸烟，B 表示不患肺癌，则“吸烟与患肺癌没有关系”等价于“吸烟与患肺癌独立”，即假设 H_0 等价于

$$P(AB)=P(A)P(B).$$

把表 3-7 中的数字用字母代替，得到如下用字母表示的列联表：

表 3-8　吸烟与患肺癌列联表

	不患肺癌	患肺癌	总计
不吸烟	a	b	$a+b$
吸烟	c	d	$c+d$
总计	$a+c$	$b+d$	$a+b+c+d$

在表 3-8 中，a 恰好为事件 AB 发生的频数；$a+b$ 和 $a+c$ 恰好分别为事件 A 和 B 发生的频数. 由于频率近似于概率，所以在 H_0 成立的条件下应该有

$$\frac{a}{n}\approx\frac{a+b}{n}\times\frac{a+c}{n},$$

其中 $n=a+b+c+d$ 为样本容量，即

$$(a+b+c+d)a\approx(a+b)(a+c),$$

即

$$ad\approx bc.$$

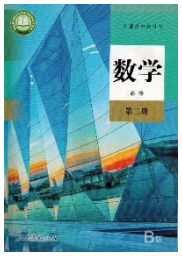
因此，$|ad-bc|$ 越小，说明吸烟与患肺癌之间关系越弱；$|ad-bc|$ 越大，说明吸烟与患肺癌之间关系越强.

为了使不同样本容量的数据有统一的评判标准，基于上面的分析，我们构造一个随机变量

$$K^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)},\tag{1}$$

其中 $n=a+b+c+d$ 为样本容量.

（三）2017 课标版《普通高中教科书・数学・必修 2・B 版》2019.4



第五章统计与概率

5.3.5 随机事件的独立性

一般地，当 $P(AB)=P(A)P(B)$ 时，就称事件 A 与 B 相互独立（简称独立），事件 A 与 B 相互独立的直观理解是，事件 A 是否发生不会影响事件 B 发生的概率，事件 B 是否发生也不会影响事件 A 发生的概率

《普通高中教科书・数学・选择性必修 2・B 版》2019.4

第四章概率与统计

4.1 条件概率与事件的独立性

4.1.1 条件概率

一般地，当事件 B 发生的概率大于 0 时（即 $P(B)>0$），已知事件 B 发生的条件下事件 A 发生的概率，称为条件概率，记作 $P(A|B)$，而且

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

4.1.3 独立性与条件概率的关系

当 $P(B)>0$ 时，A 与 B 独立的充要条件是 $P\left(A\middle|B\right)=P(A)$

4.3.2 独立性检验

统计量 χ^2 推导，根据两个事件相互独立的定义， $P(A)P(B)$ 看做 $P(AB)$ 的近似值

尝试与发现

此时，可以利用 $P(AB)=P(A)P(B)$ 是否成立来判断 A 与 B 是否独立吗？为什么？

因为 $P(A)$，$P(B)$，$P(AB)$都是根据样本数据得到的估计值，而估计是有误差的，因此直接用 $P(AB)=P(A)P(B)$是否成立来判断 A 与 B 是否独立是不合理的.

但是，如果 A 与 B 独立，那么 $P(A)P(B)$应该可以作为 $P(AB)$的近似值. 因此理论上可知，喜欢长跑的女生数可以估计为 $110P(A)P(B)$，注意到实际数为 20（即 $110P(AB)$），因此

$$\frac{[110P(AB)-110P(A)P(B)]^2}{110P(A)P(B)}$$

应该不会太大.

类似地，考虑 $\overline{A}$ 与 B，A 与 $\overline{B}$，$\overline{A}$ 与 $\overline{B}$，可知

$$\frac{[110P(\overline{A}B)-110P(\overline{A})P(B)]^2}{110P(\overline{A})P(B)},$$

$$\frac{[110P(A\overline{B})-110P(A)P(\overline{B})]^2}{110P(A)P(\overline{B})},$$

$$\frac{[110P(\overline{A}\overline{B})-110P(\overline{A})P(\overline{B})]^2}{110P(\overline{A})P(\overline{B})}$$

都应该不会太大.

若记上述四项的和为 χ^2（读作“卡方”），则代入有关数据可以算得 $\chi^2\approx7.8$.

不过，概率学上可以证明，如果 A 与 B 独立，则 $\chi^2\geqslant6.635$ 的概率只有 1%，即 $P(\chi^2\geqslant6.635)=1\%$. 因为算出的 χ^2 值 7.8 大于 6.635，所以若 A 与 B 独立（即“喜欢长跑”与“是女生”独立），那么我们就观察到了一件概率不超过 1%的事件. 这也可以说成，在犯错误的概率不超过 1%的前提下，可以认为“喜欢长跑”与“是女生”不独立（也称为是否喜欢长跑与性别有关）；或说有 99%的把握认为是否喜欢长跑与性别有关.

上述 1%通常称为显著性水平，而 6.635 称为显著性水平 1%所对应的分位数.

一般情况下，可以用完全类似的方法来检验两个随机事件是否独立.

如果随机事件 A 与 B 的样本数据的 2×2 列联表如下.

	A	$\overline{A}$	总计
B	a	b	$a+b$
$\overline{B}$	c	d	$c+d$
总计	$a+c$	$b+d$	$a+b+c+d$

记 $n=a+b+c+d$，则由表可知：

(1) 事件 A 发生的概率可估计为 $P(A)=\frac{a+c}{n}$；

(2) 事件 B 发生的概率可估计为 $P(B)=$ 3 ；

(3) 事件 AB 发生的概率可估计为 $P(AB)=$ 4 .

如果 A 与 B 独立，那么上述 $P(AB)$ 与 $P(A)P(B)$ 的估计值相差不会太大，注意到总数为 n，因此利用后者可以估计出，理论上既是 A 又是 B 的数据有 $nP(A)P(B)$ 个，注意到实际的数据为 a（即 $nP(AB)$）个，因此

$$\frac{[nP(AB)-nP(A)P(B)]^2}{nP(A)P(B)}=\frac{[na-(a+c)(a+b)]^2}{n(a+c)(a+b)}$$

不会太大.

类似地，考虑 $\overline{A}$ 与 B，A 与 $\overline{B}$，$\overline{A}$ 与 $\overline{B}$，可知

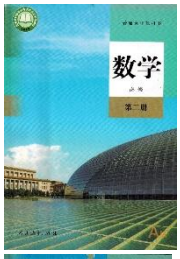
$$\frac{[nb-(b+d)(a+b)]^2}{n(b+d)(a+b)},\ \frac{[nc-(a+c)(c+d)]^2}{n(a+c)(c+d)},\ \frac{[nd-(b+d)(c+d)]^2}{n(b+d)(c+d)}$$

都不会太大，因此这四个数的和

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

也不会太大.

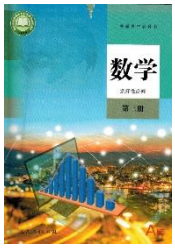
《普通高中教科书・数学・必修 2・A 版》2019.4



第十章概率

10.2 事件的相互独立性

对任意两个事件 A 与 B，如果 $P(AB)=P(A)P(B)$ 成立，则称事件 A 与事件 B 相互独立。简称为独立



《普通高中教科书・数学・选择性必修 3・A 版》2019.4

第七章随机变量及其分布

7.1 条件概率与全概率公式

一般地，设 A，B 为两个事件，且 $P(A)>0$，我们称 $P(B|A)=\frac{P(AB)}{P(A)}$ 为在事件 A 发生的条件下，事件 B 发生的条件概率，简称条件概率

当 $P(A)>0$ 时，当且仅当事件 A 与 B 相互独立时，有 $P(A|B)=P(A)$.

第八章成对数据的统计分析

8.3 列联表与独立性检验

统计量 χ^2 推导，由条件概率的定义提出零假设，进而等价于分类变量相互独立，最后根据频率稳定于概率的原理得出.

考虑以 Ω 为样本空间的古典概型. 设 X 和 Y 为定义在 Ω 上，取值于 $\{0, 1\}$ 的成对分类变量. 我们希望判断事件 $\{X=1\}$ 和 $\{Y=1\}$ 之间是否有关联. 注意到 $\{X=0\}$ 和 $\{X=1\}$，$\{Y=0\}$ 和 $\{Y=1\}$ 都是互为对立事件，与前面的讨论类似，我们需要判断下面的假定关系

$$H_0: P(Y=1 \mid X=0)=P(Y=1 \mid X=1)$$

是否成立，通常称 H_0 为**零假设**或**原假设**（null hypothesis）. 这里，$P(Y=1 \mid X=0)$ 表示从 $\{X=0\}$ 中随机选取一个样本点，该样本点属于 $\{X=0, Y=1\}$ 的概率；而 $P(Y=1 \mid X=1)$ 表示从 $\{X=1\}$ 中随机选取一个样本点，该样本点属于 $\{X=1, Y=1\}$ 的概率.

由条件概率的定义可知，零假设 H_0 等价于

$$\frac{P(X=0, Y=1)}{P(X=0)}=\frac{P(X=1, Y=1)}{P(X=1)},$$

或

$$P(X=0, Y=1)P(X=1)=P(X=1, Y=1)P(X=0). \quad ①$$

注意到 $\{X=0\}$ 和 $\{X=1\}$ 为对立事件，于是 $P(X=0)=1-P(X=1)$. 再由概率的性质，我们有

$$P(X=0, Y=1)=P(Y=1)-P(X=1, Y=1).$$

由此推得①式等价于

$$P(X=1)P(Y=1)=P(X=1, Y=1).$$

因此，零假设 H_0 等价于 $\{X=1\}$ 与 $\{Y=1\}$ 独立.

根据已经学过的概率知识，下面的四条性质彼此等价：

$\{X=0\}$ 与 $\{Y=0\}$ 独立；$\{X=0\}$ 与 $\{Y=1\}$ 独立；

$\{X=1\}$ 与 $\{Y=0\}$ 独立；$\{X=1\}$ 与 $\{Y=1\}$ 独立.

如果这些性质成立，我们就称分类变量 X 和 Y 独立. 这相当于下面四个等式成立：

$$\begin{aligned}
&P(X=0, Y=0)=P(X=0)P(Y=0);\\
&P(X=0, Y=1)=P(X=0)P(Y=1);\\
&P(X=1, Y=0)=P(X=1)P(Y=0);\\
&P(X=1, Y=1)=P(X=1)P(Y=1).
\end{aligned} \quad ②$$

因此，我们可以用概率语言，将零假设改述为

$$H_0: \text{分类变量 } X \text{ 和 } Y \text{ 独立.}$$

假定我们通过简单随机抽样得到了 X 和 Y 的抽样数据列联表，如表 8.3-3 所示.

表 8.3-3

X	Y		合计
	$Y=0$	$Y=1$	
$X=0$	a	b	$a+b$
$X=1$	c	d	$c+d$
合计	$a+c$	$b+d$	$n=a+b+c+d$

表 8.3-3 是关于分类变量 X 和 Y 的抽样数据的 2×2 列联表：最后一行的前两个数分别是事件 $\{Y=0\}$ 和 $\{Y=1\}$ 的频数；最后一列的前两个数分别是事件 $\{X=0\}$ 和 $\{X=1\}$ 的频数；中间的四个数 a，b，c，d 是事件 $\{X=x, Y=y\}$（x，$y=0$，1）的频数；右下角格中的数 n 是样本容量.

对于随机样本，表 8.3-3 中的频数 a，b，c，d 都是随机变量，而表 8.3-2 中的相应数据是这些随机变量的一次观测结果.

思考

如何基于②中的四个等式及列联表 8.3-3 中的数据，构造适当的统计量，对成对分类变量 X 和 Y 是否相互独立作出推断？

在零假设 H_0 成立的条件下，根据频率稳定于概率的原理，由②中的第一个等式，我们可以用概率 $P(X=0)$ 和 $P(Y=0)$ 对应的频率的乘积

$$\frac{(a+b)(a+c)}{n^2}$$

估计概率 $P(X=0, Y=0)$，而把

$$\frac{(a+b)(a+c)}{n}$$

视为事件 $\{X=0, Y=0\}$ 发生的频数的期望值（或预期值）. 这样，该频数的观测值 a 和

期望值$\frac{(a+b)(a+c)}{n}$应该比较接近.

综合②中的四个式子，如果零假设 H_0 成立，下面四个量的取值都不应该太大：

$$\left|a-\frac{(a+b)(a+c)}{n}\right|,\ \left|b-\frac{(a+b)(b+d)}{n}\right|,$$
$$\left|c-\frac{(c+d)(a+c)}{n}\right|,\ \left|d-\frac{(c+d)(b+d)}{n}\right|. \quad ③$$

反之，当这些量的取值较大时，就可以推断 H_0 不成立.

显然，分别考虑③中的四个差的绝对值很困难. 我们需要找到一个既合理又能够计算分布的统计量，来推断 H_0 是否成立. 一般来说，若频数的期望值较大，则③中相应的差的绝对值也会较大；而若频数的期望值较小，则③中相应的差的绝对值也会较小. 为了合理地平衡这种影响，我们将四个差的绝对值取平方后分别除以相应的期望值再求和，得到如下的统计量：

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

该表达式可化简为

$$\chi^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}. \quad (1)$$