第5课时　独立性检验(1)



知识技能

1. 通过实例，理解2×2列联表的统计意义．

2. 通过实例，了解2×2列联表独立性检验的基本思想、方法及初步应用．

思想方法

借助典型案例的探究，了解独立性检验的基本思想，总结独立性检验的基本步骤．

核心素养

经历由实际问题建立数学模型的过程，发展数学建模和数学运算素养．



重点：独立性检验的基本方法．

难点：独立性检验的基本思想．



问题导引

预习教材P158～163，思考下面的问题：

1. 2×2列联表的结构是什么？

|  |  |  |
| --- | --- | --- |
|  | Ⅱ | 合　计 |
| 类1 | 类2 |
| Ⅰ | 类*A* | *a* | *b* | *a*＋*b* |
| 类*B*  | *c* | *d* | *c*＋*d* |
| 合　计 | *a*＋*c* | *b*＋*d* | *a*＋*b*＋*c*＋*d* |

2. *χ*2统计量的公式是什么？

提示　*χ*2＝

(其中*n*＝*a*＋*b*＋*c*＋*d*)．

即时体验

1. 判断(正确的打“√”，错误的打“×”)：

(1) 分类变量中的变量与函数中的变量是同一概念；(×)

(2) 列联表中的数据是两个分类变量的频数；(√)

(3) 独立性检验中，*χ*2的值越大，则认为两个变量有关的把握越大．(√)

2. 若要检验喜欢参加体育活动是否与性别有关，应该假设喜欢参加体育活动与性别无关.

3. 对于第2题，当计算得到的*χ*2≥3.841时才能有95%的把握认为，喜欢参加体育活动与性别有关．

4. 若由一个2×2列联表中的数据计算得*χ*2＝4.013，则认为两个变量有关系的可能性为(A)

A. 95% B. 97.5%

C. 99% D. 99.9%



一、 问题情境

某医疗机构为了了解呼吸道疾病与吸烟是否有关[1]，进行了一次抽样调查，具体数据如下表所示：

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

在吸烟的人中，有≈16.82%的人患病；在不吸烟的人中，有≈7.12%的人患病．因此，从直观上可以得到结论：吸烟者与不吸烟者患病的可能性存在差异．

上述结论给我们的印象是患病与吸烟有关，事实果真如此吗？能有多大的把握认为“患病与吸烟有关”呢？如何说明患病与吸烟是否有关呢？[2]

答：若将“某成年人吸烟”记为事件*A*，“某成年人患病”记为事件*B*，则“某成年人不吸烟”为事件，“某成年人不患病”为事件.这样，说明“患病与吸烟是否有关？”就是说明“事件*A*与*B*是否独立？”．

二、 数学建构

为了便于研究一般情况，我们用字母代替数据，将上表中的“观测值”用字母表示，得到2×2列联表：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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* |

为了回答问题情境中的问题，我们先做出判断“患病与吸烟没有关系”，即提出如下假设

*H*0：患病与吸烟没有关系．[3]

由两个事件相互独立的充要条件，又可将上述假设记为

*H*0: *P*(*AB*)＝*P*(*A*)*P*(*B*)．

问题1　这里的*P*(*A*), *P*(*B*)和*P*(*AB*)的值都不知道，如果用频率来代替概率，你能否估计出*P*(*A*), *P*(*B*)和*P*(*AB*)的值吗？

设*n*＝*a*＋*b*＋*c*＋*d*，则有

*P*(*A*)≈， *P*(*B*)≈，

故*P*(*AB*)≈·.

问题2　如果实际观测值在事件*A, B*相互独立的假设下的估计值相差不“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*H*0不能被所给数据否定．否则，应认为假设*H*0不能接受．怎样描述实际观测值与估计值的差异呢？

在*H*0成立的条件下，吸烟且患病的人数为

*n*·*P*(*AB*)≈*n*··.

同理可得，吸烟但未患病的人数为

*n*·*P*(*A* )≈*n*··，

不吸烟但患病的人数为

*n*·*P*()≈*n*··，

不吸烟且未患病的人数为

*n*·*P*()≈*n*··.

为此，考虑实际观测值与在事件*A, B*独立的假设下的估计值的差(如下表)．

|  |  |  |
| --- | --- | --- |
|  | 患病 | 未患病 |
| 吸烟 | *a*－*n*·· | *b*－*n*·· |
| 不吸烟 | *c*－*n*·· | *d*－*n*·· |

问题3　怎样刻画实际观测值与估计值两者之间差异性的大小？

为了避免正负相消及消除样本容量对差异大小的影响，可以将它们分别平方并除以对应的估计频数(即估计值)，最后相加．

统计量 *χ*2(读作卡方)的计算如下：

*χ*2＝＋

 　＋＋，

化简，得

*χ*2＝

(其中*n*＝*a*＋*b*＋*c*＋*d*)．

把问题情境中的*a*＝37, *b*＝183, *c*＝21, *d*＝274代入计算得*χ*2＝≈11.8634.

问题4　如何根据*χ*2统计量进行推断呢？[4]

统计学家们根据统计数据得到了如下的临界值表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *P*(*χ*2≥*x*0) | 0.50 | 0.40 | 0.25 | 0.15 | 0.10 |
| *x*0 | 0.455 | 0.708 | 1.323 | 2.072 | 2.706 |
| *P*(*χ*2≥*x*0) | 0.05 | 0.025 | 0.010 | 0.005 | 0.001 |
| *x*0 | 3.841 | 5.024 | 6.635 | 7.879 | 10.828 |

统计学已有明确的结论：在*H*0成立的情况下，随机事件“*χ*2≥10.828”发生的概率约为0.001，即

*P*(*χ*2≥10.828)≈0.001.

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10.828的概率约为0.001.由此，我们有99.9%的把握认为*H*0不成立，即有99.9%的把握认为“患病与吸烟有关系”．

以上我们研究了吸烟与患呼吸道疾病是否有关的问题．用这种方法还可以研究类似的问题，如花的颜色与花粉的形状是否有关、用药效果与用药方式是否有关等．

用*χ*2统计量研究这类两个分类变量是否有关的问题的方法称为独立性检验．

一般地，对于两个分类变量Ⅰ和Ⅱ， Ⅰ有两类取值，即类*A*和类*B*(如吸烟与不吸烟)，Ⅱ也有两类取值，即类1和类2(如患呼吸道疾病和未患呼吸道疾病)，得到如下列联表所示的抽样数据：

|  |  |  |
| --- | --- | --- |
|  | Ⅱ | 合　计 |
| 类1 | 类2 |  |
| Ⅰ | 类*A* | *a* | *b* | *a*＋*b* |
| 类*B* | *c* | *d* | *c*＋*d* |
| 合　计 | *a*＋*c* | *b*＋*d* | *a*＋*b*＋*c*＋*d* |

要推断“Ⅰ和Ⅱ有关系”，可按下面的步骤进行：

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的值；

第三步，查对课本中的临界值表，做出判断．

独立性检验的思想：在一个已知假设下，如果一个与该假设矛盾的小概率事件发生，就推断这个假设不成立．

问题5　用*χ*2进行独立性检验时，当抽取的样本量很小时，其结论是否可靠？

当抽取的样本量很小时，用*χ*2进行独立性检验得到的结论不一定可靠，需要采取进一步的检验．

用样本估计总体时，由于抽样的随机性，结果并不唯一．因此，由某个样本得到的推断有可能正确，也有可能错误．利用*χ*2进行独立性检验，可以对推断的正确性的概率做出估计，*n*越大，这个估计越准确．

三、 数学运用

例1　(多选)对于下列实际问题，用独立性检验可以解决的有[5](CD)

A. 一种药物对某种病的治愈率

B. 吸烟者得肺病的概率

C. 吸烟人群是否与性别有关系

D. 上网与青少年的犯罪率是否有关系

(见学生用书课堂本P98)

[处理建议]　利用独立性检验的概念分析推理．

[规范板书]　解析　独立性检验是用来解决两类分类变量的关系问题，A, B不是研究两类分类变量的关系的问题，不符合题意．

[题后反思]　独立性检验是研究两个分类变量是否有关的统计方法，在独立性检验的概念中要准确分析出两个分类变量；在学习中还要注意其与相关关系概念的区别，相关性分析是研究两个变量之间的相关关系的统计方法．

　(多选)关于独立性检验的说法中正确的有(ABD)

A. 样本不同，独立性检验的结论可能有差异

B. 分类变量*X*与*Y*的统计量*χ*2越小，“*X*与*Y*有关系”的可信程度越小

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

D. 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%的前提下认为“吸烟与患肺癌有关”

[规范板书]　解析　有95%的把握认为“秃顶与患心脏病有关”，不表示某人秃顶就有95%的可能患有心脏病，所以C不正确．

[题后反思]　注意D中的两种等价说法．

在犯错误的概率不超过1%的前提下认为“吸烟与患肺癌有关”，是指有1%的可能性使得判断出现错误，但并不是说某人吸烟就有99%的可能患有肺癌，更不是说在100个吸烟的人中必有99人患有肺癌.

例2　下面是一个2×2列联表，则表中*a, b*的值分别为[6](C)

|  |  |  |  |
| --- | --- | --- | --- |
|  | *y*1 | *y*2 | 合计 |
| *x*1 | *a* | 21 | 73 |
| *x*2 | 2 | 25 | 27 |
| 合计 | *b* | 46 | 100 |

A. 94, 96 B. 52, 50

C. 52, 54 D. 54, 52

(见学生用书课堂本P98)

[处理建议]　结合两个分类变量的2×2列联表的结构特征解题．

[规范板书]　解析　由*a*＋21＝73，得*a*＝52；由*b*＋46＝100，得*b*＝54.

[题后反思]　在*n*×*n*列联表中，第*n*列与第*n*行分别是其前(*n*－1)列、前(*n*－1)行数字之和．

　(多选)某村庄对村内50名老年人、年轻人每年是否体检的情况进行了调查，统计数据如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 每年体检 | 每年未体检 | 合计 |
| 老年人 | *a* | 7 | *c* |
| 年轻人 | 6 | *b* | *d* |
| 合计 | *e* | *f* | 50 |

已知抽取的老年人、年轻人各25名，则对列联表数据的分析错误的有(ABC)

A. *a*＝18 B. *b*＝19

C. *c*＋*d*＝50 D. *e*－*f*＝2

[规范板书]　解析　*a*＋7＝*c*＝25, 6＋*b*＝*d*＝25, *a*＋6＝*e,* 7＋*b*＝*f, e*＋*f*＝50，所以*a*＝18, *b*＝19, *c*＋*d*＝50, *e*＝24, *f*＝26，从而*e*－*f*＝－2.

例3　(教材P162例1)在500人身上试验某种血清预防感冒的作用，把他们1年中的感冒记录与另外500名未用血清的人的感冒记录进行比较，结果如下表所示．

|  |  |  |  |
| --- | --- | --- | --- |
|  | 未感冒 | 感冒 | 合计 |
| 使用血清 | 258 | 242 | 500 |
| 未使用血清 | 216 | 284 | 500 |
| 合计 | 474 | 526 | 1000 |

问：该种血清对预防感冒是否有作用？[7]

(见学生用书课堂本P99)

[处理建议]　提出假设，计算*χ*2的值并作出判断．

[规范板书]　解　提出假设

*H*0：感冒与是否使用该种血清没有关系．

根据列联表中的数据，可以求得

*χ*2＝≈7.075.

因为当*H*0成立时，*χ*2≥6.635的概率约为0.01，所以我们有99%的把握认为，该种血清能起到预防感冒的作用．

[题后反思]　在查表时，先根据计算得到的*χ*2的值找到临界值表第二行对应的数值，再找到该数值对应的第一行的值．要注意，表中第一行数据表示两个变量没有关联的可能性*P*，所以其有关联的可能性为1－*P*.

　某医疗机构为了了解肝病与酗酒是否有关，对成年人进行了一次随机抽样抽查，结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 患肝病 | 未患肝病 | 合　计 |
| 酗　酒 | 700 | 200 | 900 |
| 不酗酒 | 60 | 32 | 92 |
| 合　计 | 760 | 232 | 992 |

从直观上你能得到的结论是患肝病与酗酒有关系的可能性很大，得到患肝病与酗酒有关系的判断有99%的把握．

[规范板书]　解析　提出假设

*H*0：患肝病与酗酒没有关系．

根据列联表中的数据，可以求得

*χ*2＝≈7.349.

因为当*H*0成立时，*χ*2≥6.635的概率约为0.01，所以我们有99%的把握认为，患肝病与酗酒有关系．

[题后反思]　解决一般的独立性检验问题，首先由所给2×2列联表确定*a, b, c, d, n*的值，然后根据统计量*χ*2的计算公式确定*χ*2的值，最后根据所求值确定有多大的把握判定两个变量有关联.

例4　甲、乙两个班级进行一门课程的考试，按照学生考试成绩优秀和不优秀统计成绩后，得到如下列联表：

|  |  |  |
| --- | --- | --- |
|  | 优秀 | 不优秀 |
| 甲班 | 10 | 35 |
| 乙班 | 7 | 38 |

问：学生的学习成绩与所在的班级有关吗？

(见学生用书课堂本P99)

[处理建议]　将列联表补充完整，再按步骤解答．

[规范板书]　解　根据表中的数据计算得下表：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 优秀 | 不优秀 | 合计 |
| 甲班 | 10 | 35 | 45 |
| 乙班 | 7 | 38 | 45 |
| 合计 | 17 | 73 | 90 |

提出假设

*H*0：学生的学习成绩与所在班级无关．

根据列联表中的数据，可以求得

*χ*2＝≈0.653.

因为0.653＜2.706，所以没有充分的证据认为，学生的学习成绩与所在的班级有关．

　(教材P163例2)为研究不同的给药方式(口服与注射)和药的效果(有效与无效)是否有关，进行了相应的抽样调查，调查结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 有效 | 无效 | 合计 |
| 口服 | 58 | 40 | 98 |
| 注射 | 64 | 31 | 95 |
| 合计 | 122 | 71 | 193 |

根据所选择的193个病人的数据，能否作出药的效果与给药方式有关的结论？

[处理建议]　计算在口服的病人中，有≈59%的人有效；在注射的病人中，有≈67%的人有效．因此，从直观上来看，口服与注射的病人的用药效果有一定的差异．为了研究能否认为用药效果与用药方式一定有关，下面用独立性检验的方法加以说明．

[规范板书]　解　提出假设

*H*0：药的效果与给药方式没有关系．

根据列联表中的数据，可以求得

*χ*2＝≈1.3896<2.072.

因为当*H*0成立时，*χ*2≥1.3896的概率大于15%，这个概率比较大，所以根据目前的调查数据，不能否定假设*H*0，即不能作出药的效果与给药方式有关的结论．

[题后反思]　独立性检验的步骤：

① 构造2×2列联表(若已知，则省略此步)；

② 提出假设，并计算*χ*2；

③ 查表，确定有多大的把握判定两个变量有关系．

四、 课堂练习

1. 下列关于回归分析与独立性检验的说法中正确的是(C)

A. 回归分析和独立性检验没有什么区别

B. 回归分析是对两个变量准确关系的分析，而独立性检验分析的是两个变量之间的不确定性关系

C. 回归分析研究两个变量之间的相关关系，独立性检验是对两个变量是否具有某种关系的一种检验

D. 独立性检验可以100%确定两个变量之间是否具有某种关系

2. 通过随机询问110名性别不同的大学生是否爱好某项运动，得到如下列联表：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 男 | 女 | 合计 |
| 爱好 | 40 | 20 | 60 |
| 不爱好 | 20 | 30 | 50 |
| 合计 | 60 | 50 | 110 |

由*χ*2＝算得*χ*2＝≈7.822，则下列结论中正确的是(C)

A. 在犯错误的概率不超过0.1%的前提下，认为“爱好该项运动与性别有关”

B. 在犯错误的概率不超过0.1%的前提下，认为“爱好该项运动与性别无关”

C. 有99%以上的把握认为“爱好该项运动与性别有关”

D. 有99%以上的把握认为“爱好该项运动与性别无关”

3. 某次国际会议为了搞好对外宣传工作，会务组选聘了50名记者担任对外翻译工作，在如下2×2列联表中，*a*＋*b*＋*d*＝44.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 会外语 | 不会外语 | 合计 |
| 男 | *a* | *b* | 20 |
| 女 | 6 | *d* |  |
| 合计 | 18 |  | 50 |

4. 某医疗机构为了了解呼吸道疾病与吸烟是否有关，进行了一次随机抽样调查，调查结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 患病 | 未患病 | 合计 |
| 吸　烟 | 30 | 170 | 200 |
| 不吸烟 | 20 | 280 | 300 |
| 合　计 | 50 | 450 | 500 |

问：患呼吸道疾病是否与吸烟有关？

解　提出假设

*H*0：患呼吸道疾病与吸烟无关．

根据列联表中的数据，可以求得

*χ*2＝≈9.259.

因为当*H*0成立时，*χ*2≥7.879的概率约为0.005，所以有99.5%的把握认为，患呼吸道疾病与吸烟有关．

五、 课堂小结

1. 独立性检验的含义：用*χ*2统计量研究两个分类变量是否有关等问题的方法称为独立性检验．

2. 独立性检验的一般步骤：提出假设*H*0；根据2×2列联表和公式计算*χ*2的值；查对临界值表，作出判断．



[1] 这里所说的“呼吸道疾病与吸烟有关”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”，应特别注意这两者的区别．

[2] 从实际问题入手，抽象出数学问题，培养数学抽象素养，引出独立性检验的课题．

[3] *H*0称为原假设．*H*0不成立，即“患病与吸烟有关系”，称为备择假设．

[4] 卡方越小，独立性越强，相关性越弱；卡方越大，独立性越弱，相关性越强．

[5] 独立性检验相关概念的理解．

[6] 对2×2列联表结构的理解．

[7] 利用独立性检验的思想解决实际问题．