

第二章 统计学

2.6 假设检验

2.6.1 引言

刑事诉讼遵循无罪推定原则：在人民法院依法作出生效有罪判决之前，应当推定被告人无罪。此时，证明责任由检察机关承担，只有当其提出的证据达到法定证明标准，足以排除合理怀疑时，法庭才会作出有罪判决；否则，依据疑罪从无原则，应当作出无罪判决。

这与我们要讲的假设检验流程非常像。

看一个贴近生产实际的例子，某罐装食品厂规定，每罐产品的标准质量为 500 克，工作人员随机抽取了 10 罐产品，测得质量（单位：克）如下：

495, 510, 505, 498, 503, 492, 502, 512, 497, 506

问该厂生产线是否出现了异常？

第一眼看过去，这些数据没有一个等于 500，能据此判断生产线出现了异常吗？

在实际生产过程中，受设备精度、原材料以及环境的影响，每件产品的质量都不可避免地存在随机波动，显然不能通过上述观察立即断定生产出现了异常。但问题在于，我们实际观察到的偏差究竟是正常的随机波动，还是生产线异常所致？一种自然的检验思路是，先假定生产线正常工作，即产品质量围绕 500 克随机波动，只有当抽样数据的偏离足够明显，以至于这种结果几乎不可能由正常波动产生时，我们才认为生产线可能出现了异常。

我们自然又要问，什么叫“偏离得足够明显”？仅仅因为出现了几个 510 克或 492 克的数据，就足以判定生产异常吗？如果不能，又应当以什么标准来衡量“证据是否充分”？我们应当如何完成这一检验流程呢？这些问题，正是假设检验所要解决的核心内容。

2.6.2 假设检验的基本概念

以上面的问题为例，用 X 表示该厂生产的罐装食品的质量，可以假定其服从正态分布¹ $N(\mu, \sigma^2)$ ，此处我们假定²标准差 $\sigma = 10$ ，“该厂生产是否正常”这一问题等价于“总体的均值 μ 是否等于 500”。

将“ $\mu = 500$ ”作为待检验的假设，记为 $H_0: \mu = 500$ ，称为**原假设（Null Hypothesis）**或**零假设**；与之对立的假设记为 $H_1: \mu \neq 500$ ，称为**备择假设（Alternative Hypothesis）**或**对立假设**。

原假设与备择假设并非随意决定，而应结合实际问题的背景。看下面若干例子。

¹在实际生产过程中，一件产品的质量通常受到设备误差、环境波动、原材料等众多独立因素共同影响。大量独立的小随机因素叠加后，产品质量往往可以用正态分布作为良好的近似模型。

²标准差未知的情况将在后续进行说明。

例 1 药厂研发了一种新药, 对新药的有效性进行假设检验. 应取 H_0 : 新药无效, H_1 : 新药有效, 因为新药一旦被认为有效, 就将进入市场并作用于真实病人. 若实际上新药无效, 却因偶然实验结果而被判定为有效, 就会造成医疗风险、资源浪费, 甚至延误患者接受真正有效治疗. 因此在没有充足证据前, 宁可暂时认为该药“无效”, 只有当实验数据强到足以推翻这一判断时, 才承认其有效性.

例 2 电商网站想把“购买按钮”的颜色从蓝色改成红色, 对销量是否提升进行假设检验. 应取 H_0 : 颜色改变对销量没有影响, H_1 : 颜色改变对销量有影响, 因为更换按钮颜色会带来设计成本、开发成本、用户习惯改变以及误判商业策略的风险. 因此在没有充足证据前, 应默认“颜色改变没有影响”, 只有当数据表现出足够显著的差异时, 才认为颜色改变确实产生了影响.

对最开始的问题而言, 应取 H_0 : 生产正常, H_1 : 生产不正常, 这是因为生产线处于持续运行状态, 若没有明显异常证据, 通常不应轻易判定其失控; 否则可能导致停工检查、产品召回、经济损失以及不必要的管理成本. 因此在常规审计中应默认“生产正常”, 只有当抽检数据明显偏离正常标准时, 才认为流水线存在问题.

(1) 一般地, 原假设表示当前默认成立的状态, 如“无变化”、“无影响”、“无差异”、“生产正常”等;

(2) 原假设表示的状态在没有充足证据的情况下不能被轻易推翻;

(3) 原假设的选择应当避免人们做出错误的判断, 避免不必要的损失.

假设检验的目的, 就是根据样本数据判断是否有充分的证据推翻原假设, 假设检验的结论有两种: “**拒绝原假设** (认为生产线工作不正常)”或“**不能拒绝原假设** (认为生产线工作正常)”

确定了假设后, 怎么进行接下来的检验呢?

还是看最开始的问题, 根据统计学知识, 样本均值 $\bar{X}$ 是总体均值 μ 的一个合理估计³. 因此要判断生产线的平均质量是否仍为 500 克, 一个自然的想法是考察样本均值 $\bar{X}$ 与 500 的差距 (注意, 不能因为 $\bar{X} \neq 500$ 而直接断定生产线出现了问题.) 我们关心的是: 样本均值偏离 500 到什么程度, 才能认为这种偏离已经不是随机误差所能解释的. 于是, 我们希望确定一个界限 k , 并按照如下规则进行判断:

$|\bar{X} - \mu_0| < k$, 认为样本均值与 500 的差异属于正常波动, 不能拒绝原假设;

$|\bar{X} - \mu_0| \geq k$, 认为样本均值与 500 的差异过大, 拒绝原假设.

如何确定这个界限 k 呢?

若原假设 H_0 成立, 则总体均值确为 $\mu_0 = 500$, 因此样本均值满足

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right),$$

³对于正态分布 $N(\mu, \sigma^2)$ 中 μ 的参数估计, $\bar{X}$ 既是它的矩估计, 也是它的极大似然估计.

标准化样本均值, 构造统计量

$$U = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1).$$

于是 $|\bar{X} - \mu_0| \geq k$ 等价于 $|U| \geq \frac{k}{\sigma/\sqrt{n}}$

k 的选取关系着我们如何划分“正常波动”和“异常波动”, 那么什么样的事件可以称为“极不正常”呢?

在前面我们讲了小概率原理: 如果某事件发生的概率极小, 那么在一次随机试验中, 它几乎不会发生. 一旦在一次试验中观测到了这样一个“几乎不会发生”的结果, 就有理由怀疑这并非抽样的偶然性所致, 而是最初作出的假设有误.

“概率极小”是一种人为规定的决策界限, 通常事先给定一个较小的概率 α (常取 (0.05, 0.01, 0.001) 等) 作为判断“小概率事件”的标准, 我们要求: 若原假设成立, 样本落入“异常区域”的概率恰好等于 α , 即

$$P\left(|U| \geq \frac{k}{\sigma/\sqrt{n}}\right) = \alpha$$

由此确定了界限 $k = \frac{\sigma}{\sqrt{n}} u_{\alpha/2}$, 其中 $u_{\alpha/2}$ 满足 $P(U > u_{\alpha/2}) = \frac{\alpha}{2}$. 假设原问题取显著性水平 $\alpha = 0.05$, 查表知 $u_{0.025} = 1.96$, 代入 $\sigma = 10$, $n = 10$, 得到 $k \approx 6.2$. 如果原假设确实成立, 那么有

$$P(|\bar{X} - 500| < 6.2) = 0.95,$$

也就是说, 在生产正常的前提下, 有 95% 的样本均值都会落在这一范围内; 只有 5% 的样本均值会落在该范围外. 若某次抽样满足 $|\bar{X} - 500| < 6.2$, 那么这属于正常波动, 并没有足够的证据表明生产线已经发生异常; 反之, 若出现 $|\bar{X} - 500| \geq 6.2$, 由于在原假设成立的前提下, 这样的结果发生概率仅有 5%. 根据“小概率事件在一次试验中几乎不可能发生”的原理, 我们有充分理由怀疑原假设不成立, 从而拒绝原假设. 所有导致我们拒绝原假设的样本均值的取值范围称为**拒绝域**, 在本例中为 $\{X \mid |\bar{X} - 500| \geq 6.2\}$.

最开始的例子中, 平均数 $\bar{X} = 502$, 可知 $|\bar{X} - 500| = 2 < 6.2$, 因此不能拒绝原假设, 认为生产线工作正常.

回顾整个假设检验流程, 我们可以将其总结为以下几个步骤:

确定假设 $\Rightarrow$ 构造统计量 $\Rightarrow$ 确定拒绝域 $\Rightarrow$ 计算并做出决策

2.6.3 独立性检验

假设我们要研究“抽烟”与“患病”是否独立, 数据如下

	患病	健康	合计
抽烟	a	b	$a + b$
不抽烟	c	d	$c + d$
合计	$a + c$	$b + d$	N

在独立性检验中, 原假设 H_0 : “抽烟”与“患病”独立, 备择假设 H_1 : “抽烟”与“患病”不独立. 如果原假设成立, 那么一个人“既抽烟又患病”的概率是

$$P(\text{抽烟且患病}) = P(\text{抽烟}) \times P(\text{患病}) = \frac{a+b}{N} \cdot \frac{a+c}{N}$$

将这个概率乘上总人数 N 可以得到预期值 $\frac{(a+b)(a+c)}{N}$, 同理, 可以计算出其余三个格子的预期人数 $E(a', b', c', d')$. 我们用观测值 $O(a, b, c, d)$ 与对应预期值 $E(a', b', c', d')$ 差值的平方作为误差, 同时对每个误差除以 E 作标准化, 由此构造出统计量

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

化简得到

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

根据统计学知识, 对于 $r \times s$ 独立性检验, 在样本量足够大时, 通过上述方法得到的统计量 $\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ 服从自由度为 $(r-1)(s-1)$ 的 χ^2 分布. 对于 2×2 独立性检验, 查阅 χ_1^2 , 得到常用的小概率值与界限:

$P(\chi^2 \geq k)$	0.1	0.05	0.025	0.01	0.005	0.001
k	2.706	3.841	5.024	6.635	7.879	10.828

回到最初的问题, 对于某个小概率值 α , 其对应的界限为 k , 若计算样本得到的统计量 $\chi^2 \geq k$, 则拒绝 H_0 , 认为“吸烟”与“患病”不独立(相关), 否则不能拒绝 H_0 , 认为“吸烟”与“患病”独立(不相关)

第二章 统计学

2.6 假设检验

习 题

A 组

- 【2015 湖北文 4】已知变量 x 和 y 满足关系 $y = -0.1x + 1$, 变量 y 与 z 正相关. 下列结论中正确的是

(A) x 与 y 正相关, x 与 z 负相关 (B) x 与 y 正相关, x 与 z 正相关

(C) x 与 y 负相关, x 与 z 负相关 (D) x 与 y 负相关, x 与 z 正相关
- 【2011 江西 6】变量 X 与 Y 相对应的一组数据为 $(10, 1), (11.3, 2), (11.8, 3), (12.5, 4), (13, 5)$; 变量 U 与 V 相对应的一组数据为 $(10, 5), (11.3, 4), (11.8, 3), (12.5, 2), (13, 1)$. r_1 表示变量 Y 与 X 之间的线性相关系数, r_2 表示变量 V 与 U 之间的线性相关系数, 则

(A) $r_2 < r_1 < 0$ (B) $0 < r_2 < r_1$ (C) $r_2 < 0 < r_1$ (D) $r_2 = r_1$
- 【2019 年新高高考适应性测试】某大学在校学生中, 理科生多于文科生, 女生多于男生, 则下述关于该大学在校学生的结论中, 一定成立的是

(A) 理科男生多于文科女生 (B) 文科女生多于文科男生

(C) 理科女生多于文科男生 (D) 理科女生多于理科男生
- 【2011 湖南文 5 理 4】通过随机询问 110 名性别不同的大学生是否爱好某项运动, 得到列联表, 由 $K^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ 算得, $K^2 = \frac{110 \times (40 \times 30 - 20 \times 20)^2}{60 \times 50 \times 60 \times 50} \approx 7.8$. 可以得到的正确结论是

(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”
- 【2014 江西文 7 理 6】某人研究中学生的性别与成绩、视力、智商、阅读量这 4 个变量之间的关系, 随机抽查 52 名中学生, 得到统计数据如下. 则与性别有关联的可能性最大的变量是

性别	成绩不及格	及格	视力好	差	智商偏高	正常	阅读量丰富	不丰富
男	6	14	4	16	8	12	14	6
女	10	22	12	20	8	24	2	30

- (A) 成绩 (B) 视力 (C) 智商 (D) 阅读量

6. 【2025 “fiddie” 模拟考 8】设随机变量 X 服从正态分布 $N(0, 1)$. 已知部分小概率值和相应的临界值如下表:

α	2.706	3.841	5.024	6.635	7.879	10.828
$P(X^2 \geq \alpha)$	0.1	0.05	0.025	0.01	0.005	0.001

若实数 m 满足 $P(m \leq X < 3) = 0.025$, 则

- (A) $2.706 < m^2 < 3.841$ (B) $3.841 < m^2 < 5.024$ (C) $5.024 < m^2 < 6.635$ (D) $6.635 < m^2 < 7.879$
7. 某公司在 x 年的销售额 y (万元) 如下表, 根据表中数据用最小二乘法得到的回归方程为 $\hat{y} = 2x + 35$.

x	2016	2017	2018	2019	2020	2021
y	y_1	y_2	y_3	y_4	y_5	y_6

当关于 a, b 的表达式 $\sum_{k=1}^6 (y_k - bk - a)^2$ 取最小值时, 求 $a + b$ 的值.

8. 【2012 课标全国文 3】在一组样本数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ($n \geq 2, x_1, x_2, \dots, x_n$ 不全相等) 的散点图中, 若所有样本点 (x_i, y_i) ($i = 1, 2, \dots, n$) 都在直线 $y = \frac{1}{2}x + 1$ 上, 求这组样本数据的样本相关系数.
9. 【人教 A 版选修三复习参考题 8 第 3 题】根据分类变量 x 与 y 的成对样本数据, 计算得到 $\chi^2 = 2.974$. 依据 $\alpha = 0.05$ 的独立性检验, 结论为

- (A) 变量 x 与 y 不独立
 (B) 变量 x 与 y 不独立, 这个结论犯错误的概率不超过 0.05
 (C) 变量 x 与 y 独立
 (D) 变量 x 与 y 独立, 这个结论犯错误的概率不超过 0.05

10. 回答下述问题:

- (1) 为什么必须基于成对样本数据推断两个分类变量之间是否有关联?
- (2) 借助等高堆积条形图直观判断两分类变量关系的方法有何不足?
- (3) 独立性检验能否得到两分类变量间的因果关系?
- (4) χ^2 独立性检验是否适用于普查数据? 如果不适用, 应当采用什么方法研究可以获得普查数据的两个分类变量之间的关联性?
- (5) 在进行独立性检验时, 一般要求列联表中的数据 a, b, c, d 大于 5, 且样本容量 n 大于 40, 试解释其原因.
- (6) 原假设 H_0 与对立假设 H_1 的位置可以互相交换吗? 为什么?
- (7) 存在“接受原假设”这一假设检验结论吗, 为什么?

- (8) 甲、乙两袋中均混有黑、白两色的 200 粒棋子. 从甲袋中随机取出 20 粒, 得到白色 18 粒、黑色 2 粒; 从乙袋中随机取出 20 粒, 得到白色 11 粒, 黑色 9 粒. 是否可以认为甲、乙两袋中黑白棋子的数量不同? 上述判断的把握分别有多大? (参考数据: $\frac{C_{100}^{18}C_{100}^2}{C_{200}^{20}} \approx 10^{-4}$, $\frac{C_{100}^{11}C_{100}^9}{C_{200}^{20}} \approx 0.167$)
- (9) 独立性检验的基本思想是什么? 它和反证法有何异同?
- (10) 为什么选用统计量 $\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ (其中 $a+b+c+d=n$) 取值的大小作为判断 A, B 是否独立的依据?
- (11) 对于已经获取的成对样本数据, 检验结论“两个变量之间有关联”“两个变量之间没有关联”的实际含义分别是什么?
- (12) 根据同一抽样数据推断两个分类变量之间是否有关联, 先采用小概率值 $\alpha = 0.01$ 检验, 发现没有充分的证据推断 H_0 不成立. 此时能否将小概率值调整至 0.05 或 0.1, 使得有充分的证据推断 H_0 不成立, 从而推翻原假设? 为什么?
- (13) 对于给定的列联表, 先进行一次独立性检验, 若将表中的所有数据扩大为原来的 10 倍, 在相同的检验标准下, 再进行一次独立性检验, 所得检验结果是否相同? 试解释原因;
- (14) 若通过独立性检验发现两变量有关联, 能否借助一元回归模型建立它们之间的具体关系? 为什么?
- (15) 若通过独立性检验发现, 有一定的把握认为 A 与 B 有关, 此时是否有更好的调查方法来估计某范围内事件 A (或事件 B) 发生的比例?
- (16) 当我们推断零假设 H_0 不成立的时候, 有可能犯错误 (称之为第 I 类错误), 记犯这类错误的概率为 α ; 当我们接受零假设时, 也可能犯错误 (一般称之为第 II 类错误), 记犯这类错误的概率为 p . 随着 α 的变化, p 如何变化?

11. 两分类变量 X, Y 的取值分别为 $\{A, \bar{A}\}$ 和 $\{B, \bar{B}\}$, 其样本频数列联表如下:

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

- (1) 若 $\frac{77}{20} < \chi^2 < 4$, 是否有 95% 的把握认为 X 与 Y 有关?
- (2) 设随机变量 $W = \left| \frac{a}{a+b} - \frac{c}{c+d} \right|$, 根据独立性检验原理, 现要用 W 构造一个判断 X 和 Y 是否有关系的规则 Q , 使得在规则 Q 下把“ X 和 Y 没有关系”错误判成“ X 和 Y 有关系”的概率不超过 0.01. 设临界值 $w_{0.01}$ 为正实数, 试从以下 4 个规则中选择一个符合上述要求的, 并说明理由:

$$Q_1: P(W \geq w_{0.01}) = 0.01 \quad Q_2: P(W \leq w_{0.01}) = 0.01$$

$$Q_3: P(W \geq w_{0.01}) \leq 0.01 \quad Q_4: P(W \leq w_{0.01}) \leq 0.01$$

- (3) 在 (2) 选取的规则下, 求 $w_{0.01}$ 并分析该检验规则的优势.

12. 【2023 “fiddie” 模拟测试 21】某校 20 名学生的数学成绩 $x_i (i = 1, 2, \dots, 20)$ 和知识竞赛成绩 $y_i (i = 1, 2, \dots, 20)$ 如下表:

学生编号 i	1	2	3	4	5	6	7	8	9	10
数学成绩 x_i	100	99	96	93	90	88	85	83	80	77
知识竞赛成绩 y_i	290	160	220	200	65	70	90	100	60	270
学生编号 i	11	12	13	14	15	16	17	18	19	20
数学成绩 x_i	75	74	72	70	68	66	60	50	39	35
知识竞赛成绩 y_i	45	35	40	50	25	30	20	15	10	5

计算可得数学成绩的平均值是 $\bar{x} = 75$, 知识竞赛成绩的平均值是 $\bar{y} = 90$, 并且 $\sum_{i=1}^{20} (x_i - \bar{x})^2 =$

$$6464, \sum_{i=1}^{20} (y_i - \bar{y})^2 = 149450, \sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y}) = 21650.$$

- (1) 求这组学生的数学成绩和知识竞赛成绩的样本相关系数 (精确到 0.01);
- (2) 设 N 是正整数, 变量 x 和变量 y 的一组样本数据为 $\{(x_i, y_i) \mid i = 1, 2, \dots, N\}$, 其中 x_i 两两不同, y_i 两两不同. 记 x_i 在 $\{x_n \mid n = 1, 2, \dots, N\}$ 中的排名是第 R_i 位, y_i 在 $\{y_n \mid n = 1, 2, \dots, N\}$ 中的排名是第 S_i 位, $i = 1, 2, \dots, N$. 定义变量 x 和变量 y 的“斯皮尔曼相关系数” (记为 ρ) 为变量 x 的排名和变量 y 的排名的样本相关系数.

i. 记 $d_i = R_i - S_i, i = 1, 2, \dots, N$. 证明: $\rho = 1 - \frac{6}{N(N^2 - 1)} \sum_{i=1}^N d_i^2$;

- ii. 用 (i) 的公式求这组学生的数学成绩和知识竞赛成绩的“斯皮尔曼相关系数” (精确到 0.01);

- (3) 比较 (1) 和 (2.ii) 的计算结果, 简述“斯皮尔曼相关系数”在分析线性相关性时的优势.

参考数据: $\sqrt{6464 \times 149450} \approx 31000$.

1. C.

新答案（来源：2024 上海，13（统计）.md）：C

【解题思路】 x, y 负相关， y, z 正相关，则 x, z 负相关.

2. C.

3.

4. C. 评：本题理科难度为 0.650，文科难度为 0.572.5. D. 变量之间的独立性检验的卡方随机变量

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

现在四个表中的 $a+b, a+c, b+d, c+d$ 都相同，若 $ad-bc$ 越大，则两个变量关联的可能性更大.

6.

7. A. 因为 $P(m \leq X < 3) = 0.025$ ，所以 $P(m^2 \leq X^2 < 9) = 0.05$. 而 $P(X^2 \leq 7.879) = 0.995 < P(X^2 \leq 9) < P(X^2 < 10.828) = 0.999$ ，故 $0.945 < P(X^2 < m^2) < 0.949 \Rightarrow 0.051 < P(X^2 \leq m^2) < 0.055$ ，于是 $2.706 < m^2 < 3.841$.

Fiddie 评：命制本题时，首先取材书上“卡方独立性检验”一节，考虑到很多学生可能不熟悉卡方随机变量，而书上给出的卡方随机变量的自由度为 1，相当于标准正态分布的平方. 所以本题最终改成了标准正态分布. 希望能用这题让同学和老师对卡方独立性检验的内容进行更全面的了解.

如果误认为 $P(m^2 \leq X^2 < 9) = 0.025$ ，就会得到 $P(X^2 \geq m^2) \in (0.026, 0.03)$ ，从而错选 B. 也有可能因为得到 $P(X^2 \geq m^2) \in (0.02, 0.024)$ 从而错选 C；因为得到 $P(X^2 \geq m^2) \in (0.045, 0.049)$ 从而错选 B. 本题错选 B 的占 40%，错选 C 的占 20%.

8.

9.

10. 1.11. C. 零假设 H_0 ：分类变量 x 和 y 独立.

卡方分布满足 $P(\chi^2 \geq 3.841) = 0.05$.

当计算得到的 $\chi^2 \geq 3.841$ 时，零假设 H_0 被否定（推断 H_0 不成立），就认为 X 和 Y 不独立，该推断犯错误的概率不超过 0.05.

当计算得到的 $\chi^2 < 3.841$ 时，没有充分证据推断 H_0 不成立，可以认为 X 和 Y 独立.

因为本题算出来 $\chi^2 = 2.974 < 3.841$ ，所以认为 X 和 Y 独立. 故选 C.

注：若把本题的 2.974 改为更大的数，例如 4.974，答案选 B.

12. 1. 因为要研究两个分类变量 X 与 Y 是否有关联, 核心是看它们的联合概率分布是否等于各自边缘分布的乘积, 即判断 $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$ 是否成立。要估计联合概率, 必须对同一个个体同时观测两个特征。如果数据不是成对的 (例如在 A 学校调查学生的性别, 在 B 学校调查学生对某学科的喜好), 将无法建立两变量间的联合对应关系, 因而无法研究它们的关联性。

2. 两大不足, 第一是缺乏量化的可信度: 图形展现的只是样本内的比例差异, 但无法告诉我们这种差异是由于变量关联引起的, 还是由于随机抽样误差引起的。第二是忽略了样本容量 n 的影响: 如果样本量很小 (10 人), 即使条形图比例差异巨大也可能是随机波动; 如果样本量极大 (10 万人), 即使条形图上极其微小的比例偏差, 在统计学上也有可能是高度显著的。

3. 不能。独立性检验只能确证两个变量之间存在统计学上的相关性, 相关性无法推导得到因果性。两个变量相关可能是因为 A 导致 B , 可能是 B 导致 A , 也可能是因为它们共同受到第三个混杂变量 C 的影响, 即共因关系, 例如冰淇淋销量与溺水人数正相关, 共因是夏季气温升高。

4. 不适用。假设检验的思想是“以样本推断总体”。普查已经获取了总体的全部数据, 无需从样本向总体进行统计推断。对于普查数据, 可直接使用描述性相关性指标, 例如计算 ϕ 相关系数 (Phi Coefficient) 或 Cramer's V 关联度:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

若该值不为 0, 则两变量在总体上就是直接存在关联的, 其大小反映了关联的紧密程度。

5. 独立性检验的数学本质是利用了渐近积分分布。卡方统计量 χ^2 只有在样本量趋近于无穷大时, 其极限分布才严格服从自由度为 1 的 χ^2 分布。

* 若格子的期望频数 (Theoretical/Expected Frequency) $T < 5$ (通常用实际频数 $a, b, c, d > 5$ 作为简易判定标准) 或者样本量 $n \leq 40$, 这种连续型的卡方分布去近似离散型的频数分布的误差会变得非常大, 导致第一类错误率 (假阳性率) 失控。* 数理拓展: 当不满足这一条件时, 根据数理统计理论, 应当执行 Yates 连续性修正 (将分子 $|ad - bc|$ 修改为 $|ad - bc| - \frac{n}{2}$) 或直接使用 Fisher 确切概率法 (Fisher's Exact Test) 进行超几何分布的精确条件概率计算。

6. 不可以 (可参考讲义的相关内容叙述。) 我们总是把“没有把握、不轻易认同、或者维持原有常规”的断言放在原假设 H_0 (如 H_0 : 变量独立/新药无效), 把“需要强有力证据支持”的断言放在对立假设 H_1 。如果调换位, 整个推断的统计学保护方向将完全相反。

7. 不存在, 检验的结论只能是“拒绝原假设” (强有力的结论) 或者“产生妥协, 没有充分证据拒绝原假设” (弱结论, 不等于原假设一定成立)。

8. 可以认为甲袋的黑白棋子数量不同; 无法认为乙袋的黑白棋子数量不同。若假设配比相同的袋子抽样为 H_0 。甲袋抽样得到 18 白 2 黑的概率为 10^{-4} 。这是一个极小概率事件。因为小概率事件在一次试验中几乎不可能发生, 现在它发生了, 所以有极大的把握指出原假设不成立。其把握为: $1 - 10^{-4} = 99.99\%$

乙袋抽样得到 11 白 9 黑的概率为 0.167。这一概率远大于常规的显著性水平 (如 0.05), 说明在两色棋子平分的袋中摸出该结果是相当常见的随机波动。样本没有提供足够坚实的证据证明其不同。

9. 独立性检验是基于小概率原理的带有容错空间的统计学推断方法。先假设 H_0 (两变量独立) 成立, 若在该前提下, 代表独立偏差程度的随机变量 χ^2 算得的观测值超过了临界值 (即发生了发

15. 有。可以采用分层抽样与估计条件概率来替代简单的无差别随机抽样。既然已经确认 A 的发生概率受到 B 的规律性干扰。可以将总体按照变量 B 划分为不同的层（子群体），在各层内部进行抽样计算 $P(A|B)$ ，再利用全概率公式估计 $P(A)$ 。由分层抽样理论可证，当层内个体的波动偏差小于层与层之间的偏差时，分层估计估计量的方差 $\text{Var}(\bar{X}_{\text{stratified}})$ 控制得最小，显著优于简单随机抽样的估计精度。

16. 在样本量 n 固定的前提下， β 会随之增大。两类错误就像天平的两端。第一类错误 α 是指“为了防止冤枉好人（错判独立）把标准订得特别严苛”。如果强行拉低 α （例如从 0.05 到 0.001），即要求卡方值极大才肯推翻独立原假设。这会使得我们的推断变得过度谨慎保守，导致很多真实存在轻微关联的情况被我们“略过”当成独立处理，从而增加了放过事实（即接受了不成立的原假设，犯第 II 类错误 β ）的几率。

13.

14.

15.

16.

第二章 统计学

2.6 假设检验

习题组 II

1. 【2023 全国甲卷文 19 , 理 19 改编】一项试验旨在研究臭氧效应, 试验方案如下: 选 40 只小白鼠, 随机地将其中 20 只分配到试验组, 另外 20 只分配到对照组, 试验组的小白鼠饲养在高浓度臭氧环境中, 对照组的小白鼠饲养在正常环境中, 一段时间后统计每只小白鼠体重的增加量 (单位: g). 试验结果如下: 对照组的小白鼠体重的增加量从小到大排序为

15.2 18.8 20.2 21.3 22.5 23.2 25.8 26.5 27.5 30.1 32.6 34.3 34.8 35.6 35.6 35.8 36.2 37.3 40.5 43.2

试验组的小白鼠体重的增加量从小到大排序为

7.80 9.20 11.4 12.4 13.2 15.5 16.5 18.0 18.8 19.2 19.8 20.2 21.6 22.8 23.6 23.9 25.1 28.2 32.3 36.5

- (1) 设 X 表示指定的两只小白鼠中分配到对照组的只数, 求 X 的分布列和数学期望;
- (2) 求 40 只小白鼠体重的增加量集中位数 m ;
- (3) 结合统计学的知识, 设计一个列联表, 并判断能否有 95% 的把握认为小白鼠在高浓度臭氧环境中与在正常环境中体重的增加量有差异?

1.

第二章 统计学

2.6 假设检验

第二章 统计学

2.6 假设检验