

第八章 成对数据的统计分析

8.1 成对数据的统计相关性

8.1.1 变量的相关关系

8.1.2 样本相关系数

刷基础

1. C 【解析】A, D 是函数关系; B 中学生的成绩和身高不是相关关系, 也不是函数关系; C 中小麦产量与施肥量是相关关系, 一般来说, 在一定范围内, 农田的施肥量越大, 小麦产量会越高. 故选 C.

名师点拨 相关关系是一种非确定性关系, 可以是因果关系, 也可以是伴随关系.

2. A 【解析】由题中散点图可知, A 选项中的散点均匀地分布在一条直线附近, 则变量 x, y 具有线性相关关系, B, C, D 选项中的散点没有分布在一条直线附近, 故 B, C, D 选项中变量 x, y 不具有线性相关关系. 故选 A.

3. BD 【解析】题中散点图①, ②中, y 与 x 呈负相关, 则 $r_1 < 0, r_2 < 0$, 散点图②中 y 与 x 的线性相关程度更强, 则 $|r_1| < |r_2|$, 因此 $r_2 < r_1 < 0$; 题中散点图

③, ④中, y 与 x 呈正相关, 则 $r_3 > 0, r_4 > 0$, 散点图④中 y 与 x 的线性相关程度更强, 则 $|r_3| < |r_4|$, 因此 $r_4 > r_3 > 0$, 所以 $r_2 < r_1 < 0 < r_3 < r_4$. 故选 BD.

$$4. 【解】\bar{x} = \frac{1+2+3+4+5}{5} = 3, \bar{y} = \frac{120+100+80+70+55}{5} = 85,$$

$$\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y}) = (-2) \times 35 + (-1) \times 15 + 0 \times (-5) + 1 \times (-15) + 2 \times (-30) = -160,$$

$$\sum_{i=1}^5 (x_i - \bar{x})^2 = (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 = 10,$$

$$\sum_{i=1}^5 (y_i - \bar{y})^2 = 35^2 + 15^2 + (-5)^2 + (-15)^2 + (-30)^2 = 2\,600,$$

所以样本相关系数

$$r = \frac{\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^5 (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^5 (y_i - \bar{y})^2}} = \frac{-160}{\sqrt{10} \times \sqrt{2\,600}} = \frac{-160}{10 \sqrt{260}} \approx -0.997.$$

≈ -0.99 ,

因为 $|r|$ 非常接近 1, 所以变量 x 与 y 之间具有很强的线性相关关系.

刷易错

★易错点 混淆相关关系与函数关系致误

5. A 【解析】对于 A, 一般情况下, 汽车越重, 则每千米耗油量越多, 呈正相关, 故 A 正确;

对于 B, 圆的半径与面积是函数关系, 故 B 错误;

对于 C, 一般情况下, 若花费在体育活动上的时间越长, 则期末考试的数学成绩可能会降低, 故不为正相关, 故 C 错误;

对于 D, 人的身高与受教育的程度没有相关关系, 故 D 错误.

故选 A.

易错警示 函数关系是一种确定性

关系, 相关关系是一种非确定性关系, 判断两个变量间的关系是否为相关关系的关键是看这个关系是否具有不确定性, 但又不能毫无关联, 如 D 选项.

8.2 一元线性回归模型及其应用

8.2.1 一元线性回归模型

8.2.2 一元线性回归模型参数的最小二乘估计

刷基础

1. B 【解析】对于 A, 样本相关系数的取值范围是 $[-1, 1]$, 故 A 错误;

对于 B, 回归系数 $\hat{b}$ 的含义: 当自变量每增加 1 个单位时, 因变量平均增加的量, 这里自变量是每日自主阅读时间, 因变量是语文成绩, 所以阅读时间每增加 1 小时, 语文成绩平均增加 5.2 分, 故 B 正确;

点悟: 这里的 5.2 是斜率参数, 是单位增加量

对于 C, 经验回归直线是对样本数据拟合的直线, 不一定经过样本点, 故 C 错误;

对于 D, 当 $x=2$ 时, $\hat{y}=5.2 \times 2 + 72.4 = 82.8$, 82.8 为预测值, 不是确定值, 故

敲黑板: 预测值是当解释变量取某个具体值时, 响应变量的均值, 不是观测值

D 错误. 故选 B.

2. C 【解析】由题意可知, $\bar{x} = \frac{1}{5} \times (2+4+5+6+8) = 5$,

因为经验回归直线 $\hat{y} = 7x + 15$ 过样本点的中心 $(\bar{x}, \bar{y})$, 所以 $\bar{y} = 7 \times 5 + 15 = 50$.

故 $\frac{1}{5} \times (30+40+50+60+a) = 50$, 解得 $a = 70$. 故选 C.

3. C 【解析】因为 x 与 y 正相关, 所以 A, D 不正确;

对于 B, 因为经验回归直线过点 $(\bar{x}, \bar{y})$, 当 $x=3$ 时, $\hat{y} = 2 \times 3 - 2.4 = 3.6$, 故 B 不正确;

对于 C, 当 $x=3$ 时, $\hat{y} = 0.4 \times 3 + 2.3 = 3.5$, 故 C 正确. 故选 C.

4. 【解】(1) 根据散点图直观判断 y 与 x 之间线性相关.

由题易知 $\bar{x} = 3, \bar{y} = 1$, $\sum_{i=1}^5 x_i y_i - 5\bar{x}\bar{y} =$

$$2.6, \sum_{i=1}^5 x_i^2 - 5\bar{x}^2 = 10, \sum_{i=1}^5 y_i^2 - 5\bar{y}^2 = 0.68,$$

$$\text{则 } r = \frac{\sum_{i=1}^5 x_i y_i - 5\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^5 x_i^2 - 5\bar{x}^2} \sqrt{\sum_{i=1}^5 y_i^2 - 5\bar{y}^2}} = \frac{2.6}{\sqrt{10} \times \sqrt{0.68}} \approx 0.997.$$

$$= \frac{2.6}{\sqrt{10 \times 0.68}} \approx 0.997,$$

所以 y 与 x 的线性相关程度强.

$$(2) \text{ 设 } \hat{y} = \hat{b}x + \hat{a}, \text{ 则 } \hat{b} = \frac{\sum_{i=1}^5 x_i y_i - 5\bar{x}\bar{y}}{\sum_{i=1}^5 x_i^2 - 5\bar{x}^2} =$$

$$\frac{2.6}{10} = 0.26, \hat{a} = \bar{y} - \hat{b}\bar{x} = 1 - 0.26 \times 3 =$$

$$0.22,$$

$$\text{所以 } \hat{y} = 0.26x + 0.22,$$

$$\text{故 } x=6 \text{ 时, } \hat{y} = 0.26 \times 6 + 0.22 = 1.78.$$

所以第 6 个月日平均参与劳动实践的学生人数预计为 1.78 万.

5. B 【解析】由题设条件, 知 $\bar{x} = \frac{2+3+4+5+6}{5} = 4$.

$\therefore$ 点 $(\bar{x}, \bar{y})$ 在经验回归直线 $\hat{y} = 12x - 18$ 上,

$$\therefore \bar{y} = 12 \times 4 - 18 = 30,$$

$$\therefore 30 = \frac{10+15+29+40+c}{5}, \text{ 解得 } c = 56,$$

$\therefore$ 当 $x=6$ 时, 残差为 $56 - (12 \times 6 - 18) = 2$. 故选 B.

6. D 【解析】模型 I 所得经验回归方程

高中必刷题 数学

为 $\hat{y} = 9.15 - 1.14x$, 因为 $-1.14 < 0$, 所以两个变量负相关, 故 **A 不正确**;

若选用模型 I, 当自变量 x 每增加 1 个单位时, 因变量 y 估计减少 1.14 个单位, 故 **B 不正确**;

若选用模型 II, 则此品牌轮胎行驶里程越多, 其轮胎凹槽深度可能越小, 故 **C 不正确**;

由于模型 I 的决定系数为 0.95, 模型 II 的决定系数为 0.99, $0.95 < 0.99$, 则模型 II 的拟合效果比模型 I 的拟合效果好, 故 **D 正确**. 故选 D.

链接教材 本题是教材第 115 页问题的延伸, 决定系数 R^2 的计算公式

$$\text{为 } R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

在 R^2 表达式中, $\sum_{i=1}^n (y_i - \bar{y})^2$ 与经验回归方程无关, 残差平方和 $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ 与经验回归方程有关. 因此 R^2

越大, 表示残差平方和越小, 即模型的拟合效果越好; R^2 越小, 表示残差平方和越大, 即模型的拟合效果越差. 拟合效果越好, 说明选择的模型越接近实际情况.

7. A 【解析】已知 $u = x, v = \ln y$,

点悟: 将非线性转换为线性, 可以使用最小二乘法计算系数 $\hat{b}, \hat{a}$

由 $\ln y$ 与 x 呈线性相关关系, 其经验回归方程为 $\hat{v} = \hat{b}u + \hat{a}$,

$$\text{得 } \hat{b} = \frac{\sum_{i=1}^5 u_i v_i - 5\bar{u}\bar{v}}{\sum_{i=1}^5 u_i^2 - 5\bar{u}^2} \approx \frac{42.5 - 5 \times 3 \times 2.5}{55 - 5 \times 3^2} = 0.5, \hat{a} = \bar{v} - \hat{b}\bar{u} \approx$$

$$2.5 - 0.5 \times 3 = 1,$$

因此 $\hat{v} = 0.5u + 1$, 即 $\ln \hat{y} = 0.5x + 1$, 所以 y 关于 x 的非线性经验回归方程为 $\hat{y} = e^{0.5x+1}$. 故选 A.

8. 【解】(1) 由题中散点图可知, $y = c \ln x + d$ 更适合作为每年 1 月份来哈尔滨的

游客数量 y 关于年份代码 x 的经验回归方程类型.

因为 $u = \ln x$, 所以 $y = cu + d$.

$$\text{因为 } \bar{u} = 1.3, \bar{y} = 165, \sum_{i=1}^8 u_i^2 = 17.02, \sum_{i=1}^8 u_i y_i = 1901.5,$$

$$\text{所以 } \hat{c} = \frac{\sum_{i=1}^8 u_i y_i - 8\bar{u}\bar{y}}{\sum_{i=1}^8 u_i^2 - 8\bar{u}^2} = \frac{1901.5 - 8 \times 1.3 \times 165}{17.02 - 8 \times 1.3^2} = \frac{185.5}{3.5} = 53,$$

所以 $\hat{d} = \bar{y} - \hat{c}\bar{u} = 165 - 53 \times 1.3 = 96.1$, 所以 $\hat{y} = 53u + 96.1$, 即 $\hat{y} = 53 \ln x + 96.1$.

所以每年 1 月份来哈尔滨的游客数量 y 关于年份代码 x 的经验回归方程为 $\hat{y} = 53 \ln x + 96.1$.

(2) 由 (1) 知 y 关于 x 的经验回归方程为 $\hat{y} = 53 \ln x + 96.1$,

当 $x = 11$ 时, $\hat{y} \approx 53 \times 2.4 + 96.1 = 223.3$,

所以预测 2027 年 1 月份来哈尔滨的游客数量为 223.3 万.

第 8.1, 8.2 节综合训练

刷能力

1. C 【解析】∵ 第一组散点图的点整体呈下降趋势, 两变量负相关, $\therefore r_1 < 0$, 故 **A 正确**, 不符合题意;

∵ 第二组散点图的点整体呈上升趋势, 两变量正相关, $\therefore r_2 > 0$, 故 **B 正确**, 不符合题意;

观察散点图可知第一组数据的散点更集中在一条直线附近, 线性相关程度更强, $\therefore |r_1| > |r_2|$, 故 **C 错误**, 符合题意;

∵ $r_1 < 0, r_2 > 0$ 且 $|r_1| > |r_2|$, $\therefore r_1 + r_2 < 0$, 故 **D 正确**, 不符合题意. 故选 C.

2. D 【解析】由题中折线图可知, 新能源汽车销量逐月增加, 所以新能源汽车销量与月份正相关, 故 **A 正确**; 由题中折线图可知, 燃油车销量逐月减少, 可预测燃油车销量呈现下降趋势, 故 **B 正确**; 由题中折线图可知, 新能源汽车销量逐月增长率大致相同, 故 **C 正确**; 由题中折线图可知, 燃油车销量随月份的递增呈现下降趋势, 且近似分布在一条直线上, 所以燃油车销量与月份的样本相关系数 r 接近 -1, 故 **D 错误**. 故选 D.

3. C 【解析】设第 1 代至第 11 代的平均质量分别为 $a_1, a_2, \dots, a_{11}$, 易知 $a_1 = 3.5$.

又 $a_2 + a_3 + \dots + a_{10} = 20$,

$$\text{前 10 代平均质量 } \bar{x} = \frac{a_1 + a_2 + \dots + a_{10}}{10} = \frac{3.5 + 20}{10} = 2.35 \text{ 千克, 后 10 代平均质量}$$

点悟: 样本数据共有 10 组: $(a_1, a_2), (a_2, a_3), \dots, (a_{10}, a_{11})$, 所以 $\bar{x}$ 为前 10 代平均质量, $\bar{y}$ 为后 10 代平均质量

$$\bar{y} = \frac{a_2 + a_3 + \dots + a_{11}}{10} = \frac{20 + a_{11}}{10},$$

将 $(\bar{x}, \bar{y})$ 代入经验回归方程中可得 $\frac{20 + a_{11}}{10} = 0.8 \times 2.35 + 0.3$,

解得 $a_{11} = 1.8$. 故选 C.

4. C 【解析】将 $\hat{y} = e^{bx-0.5}$ 两边同时取对数, 得到 $\ln \hat{y} = bx - 0.5$, 令 $\hat{z} = \ln \hat{y}$, 得到 $\hat{z} = bx - 0.5$.

根据已知表格数据, 得到 x, z 的取值对照表如下:

x	1	2	3	4
z	1	3	4	6

由上述表格可知,

$$\bar{x} = \frac{1+2+3+4}{4} = 2.5, \bar{z} = \frac{1+3+4+6}{4} = 3.5,$$

则 $3.5 = 2.5\hat{b} - 0.5$, 解得 $\hat{b} = 1.6$, 则 $\hat{z} = 1.6x - 0.5$, 进而得到 $\hat{y} = e^{1.6x-0.5}$, 将 $x = 5$ 代入, 解得 $\hat{y} = e^{7.5} = e^{\frac{15}{2}}$. 故选 C.

5. C 【解析】由统计学知识可知, R^2 越大, 拟合效果越好,

又 A 小组的决定系数 $R_1^2 = 0.8732$, B 小组的决定系数 $R_2^2 = 0.9375$,

$\therefore$ B 小组的拟合效果好, 则经验回归方程为 $\hat{v} = -0.5006u + 0.4922$,

又 $u = x^2, v = y^2, \therefore y^2 = -0.5006x^2 + 0.4922$, 即 $\frac{0.5006x^2}{0.4922} + \frac{y^2}{0.4922} = 1$.

故选 C.

6. ABD 【解析】由题知 $\bar{x} = \frac{3+4+5+6}{4} =$

$$4.5, \bar{y} = \frac{5+8+m+17}{4} = \frac{30+m}{4}.$$

将 $(\bar{x}, \bar{y})$ 的坐标代入 $\hat{y} = 4.2x - 7.9$ 中, 得 $\frac{30+m}{4} = 4.2 \times 4.5 - 7.9$, 所以 $m = 14$,

A 选项正确;

由 $\hat{y} = 4.2x - 7.9$ 中 $4.2 > 0$, 得变量 x 和 y 呈正相关关系, **B 选项正确**;

由题知 $\bar{x} = 4.5, \bar{y} = \frac{30+14}{4} = 11$, 该经验回归直线必过点 $(4.5, 11)$, **C 选项错误**;

当 $x = 8$ 时, $\hat{y} = 4.2 \times 8 - 7.9 = 25.7$,

故当某户家庭人口数为 8 时, 预测该户茶产业的年收入为 25.7 万元, **D 选项正确**. 故选 ABD.

7. AC 【解析】10 人的身高的平均值为 $\frac{1}{10} \times (165 + 168 + 170 + 172 + 173 + 174 + 175 + 177 + 179 + 182) = 173.5$ cm, 从散点图分析可知, x 与 y 正相关, 且异常数据 (168, 89) 的横坐标 168 小于平均值, 纵坐标 89 相对过大, 所以去掉异常数据后经验回归直线的斜率变大, 即 $\hat{b}_1 < \hat{b}_2$, 因为去掉异常数据后成对样本数据的线性相关程度更强, 拟合效果会更好, 所以 $r_1 < r_2$. 又因为重新测量后的数据恰好为余下 9 组数据的平均值, 所以 $\hat{b}_3 = \hat{b}_2, r_3 = r_2$. 综上, $\hat{b}_1 < \hat{b}_2 = \hat{b}_3, r_1 < r_2 = r_3$. 故选 AC.

8. -0.05 【解析】日期代码的平均值为 $\bar{x} = \frac{1+2+3+4+5}{5} = 3$, 电池级碳酸锂价格的平均值为 $\bar{y} = \frac{4.1+3.9+3.8+3.8+3.9}{5} = 3.9$, 因为经验回归直线过样本点的中心 $(\bar{x}, \bar{y})$, 所以将 $\bar{x} = 3, \bar{y} = 3.9$ 代入 $\hat{y} = -0.05x + \hat{a}$ 中, 可得 $3.9 = -0.05 \times 3 + \hat{a}$, 解得 $\hat{a} = 4.05$, 则经验回归方程为 $\hat{y} = -0.05x + 4.05$. 当 $x = 4$ 时, 预测值 $\hat{y} = -0.05 \times 4 +$

$4.05 = 3.85$, 根据残差的定义, 样本点 (4, 3.8) 处的残差为 $y - \hat{y} = 3.8 - 3.85 = -0.05$.

9. 【解】 (1) 由题表知, $\bar{x} = \frac{1}{8} \times (1+2+3+4+5+6+7+8) = 4.5$, $\bar{y} = \frac{1}{8} \times (0.07+0.12+0.18+0.25+0.28+0.30+0.35+0.45) = 0.25$, 所以 $\hat{b} = \frac{\sum_{i=1}^8 x_i y_i - 8\bar{x}\bar{y}}{\sum_{i=1}^8 x_i^2 - 8\bar{x}^2} = \frac{11.1 - 8 \times 4.5 \times 0.25}{204 - 8 \times 4.5^2} = 0.05$, 则 $\hat{a} = 0.25 - 0.05 \times 4.5 = 0.025$, 所以 y 关于 x 的经验回归方程为 $\hat{y} = 0.05x + 0.025$.

(2) 由已知得 $|y_1 - \hat{y}_1| = |0.07 - (0.05 \times 1 + 0.025)| = 0.005 < 0.01, |y_2 - \hat{y}_2| = |0.12 - (0.05 \times 2 + 0.025)| = 0.005 < 0.01,$ $|y_3 - \hat{y}_3| = |0.18 - (0.05 \times 3 + 0.025)| = 0.005 < 0.01, |y_4 - \hat{y}_4| = |0.25 - (0.05 \times 4 + 0.025)| = 0.025 > 0.01,$ $|y_5 - \hat{y}_5| = |0.28 - (0.05 \times 5 + 0.025)| =$

$0.005 < 0.01, |y_6 - \hat{y}_6| = |0.30 - (0.05 \times 6 + 0.025)| = 0.025 > 0.01,$ $|y_7 - \hat{y}_7| = |0.35 - (0.05 \times 7 + 0.025)| = 0.025 > 0.01, |y_8 - \hat{y}_8| = |0.45 - (0.05 \times 8 + 0.025)| = 0.025 > 0.01,$ 即有 4 组数据为“正常数据”, 所以 X 的所有可能取值为 0, 1, 2, 3, 4, 则 $P(X=0) = \frac{C_4^0 C_4^4}{C_8^4} = \frac{1}{70}$, $P(X=1) = \frac{C_4^1 C_3^3}{C_8^4} = \frac{8}{35}$, $P(X=2) = \frac{C_4^2 C_2^2}{C_8^4} = \frac{18}{35}$, $P(X=3) = \frac{C_4^3 C_1^1}{C_8^4} = \frac{8}{35}$, $P(X=4) = \frac{C_4^4 C_0^0}{C_8^4} = \frac{1}{70}$, 所以 X 的分布列为

X	0	1	2	3	4
P	$\frac{1}{70}$	$\frac{8}{35}$	$\frac{18}{35}$	$\frac{8}{35}$	$\frac{1}{70}$

故期望 $E(X) = 0 \times \frac{1}{70} + 1 \times \frac{8}{35} + 2 \times \frac{18}{35} + 3 \times \frac{8}{35} + 4 \times \frac{1}{70} = 2$.

8.3 列联表与独立性检验

8.3.1 分类变量与列联表

8.3.2 独立性检验

刷基础

1. BCD 【解析】散点图适用于推断两个变量之间是否有相关关系, 等高堆积条形图, 列联表, χ^2 独立性检验适用于推断两个分类变量之间是否有关联. 故选 BCD.

2. D 【解析】在四个等高堆积条形图中, D 图中 x_1, x_2 所占比例相差最大, 分类变量 x, y 之间关系最强. 故选 D.

3. 28 【解析】由题得 $\begin{cases} a+b=20, \\ a+6=18, \\ b+d=32, \end{cases}$ 解得

$$\begin{cases} a=12, \\ b=8, \\ d=24, \end{cases} \text{ 所以 } a-b+d=28.$$

4. D 【解析】要使是否喜欢羽毛球运动与性别无关的可能性最大, 则 χ^2 最小, 即 $\chi^2 = 0$, 所以 $100m = 50n$, 所以 $\frac{m}{n} = \frac{1}{2}$.

5. AC 【解析】对于 A, 根据表格数据可知 $b = 200 - 150 = 50, c = 270 - 150 = 120$, A 正确; 对于 B, 为了研究荧光抗体法与常规

培养法在沙门氏菌检验结果中是否存在差异, 零假设为 H_0 : 在沙门氏菌检验中荧光抗体法与常规培养法无差异, B 错误;

对于 C, 由题意得 $\chi^2 = \frac{400 \times (150 \times 80 - 50 \times 120)^2}{200 \times 200 \times 270 \times 130} \approx 10.256 >$

$6.635 = x_{0.010}$, 所以零假设 H_0 不成立, 依据小概率值 $\alpha = 0.01$ 的 χ^2 独立性检验, 认为荧光抗体法与常规培养法在沙门氏菌检验中有差异, C 正确;

对于 D, 由表格数据知, 常规培养法检测沙门氏菌阳性的频率为 $\frac{120}{200} = \frac{3}{5}$, D

错误. 故选 AC.

6. 300 【解析】设男、女生人数均为 x ($x \in \mathbb{N}^*$), 可得如下 2×2 列联表:

性别	对科普的关注度		合计
	关注	不关注	
男生	$\frac{5x}{6}$	$\frac{x}{6}$	x
女生	$\frac{2x}{3}$	$\frac{x}{3}$	x
合计	$\frac{3x}{2}$	$\frac{x}{2}$	$2x$

$$\text{依题意, } \chi^2 = \frac{2x \left(\frac{5x}{6} \cdot \frac{x}{3} - \frac{2x}{3} \cdot \frac{x}{6} \right)^2}{x^2 \cdot \frac{3x}{2} \cdot \frac{x}{2}} =$$

$$\frac{2x}{27} \geq 10.828,$$

【敲黑板】依据小概率值 $\alpha = 0.001$ 的独立性检验认为有关联, 则 χ^2 的值必须大于或等于 0.001 对应的临界值

因此 $2x \geq 292.356$, 显然 x 必须为 6 的整数倍, $2x$ 必须为 12 的整数倍, 则 $(2x)_{\min} = 300$, 所以调查的总人数最少为 300.

7. 【解】 (1) 根据题中数据完成 2×2 列联表如下所示.

	优良	优良以下	合计
2025 年下学期	20	30	50
2026 年上学期	40	10	50
合计	60	40	100

(2) 零假设为 H_0 : 该班体测等级优良率与参加体测的学期无关.

$$\text{由 (1) 得 } \chi^2 = \frac{100 \times (20 \times 10 - 40 \times 30)^2}{50 \times 50 \times 60 \times 40} =$$

$$\frac{50}{3} \approx 16.667 > 7.879 = x_{0.005},$$

依据小概率值 $\alpha = 0.005$ 的独立性检验,我们推断 H_0 不成立,即认为该班体测等级优良率与参加体测的学期有关,该推断犯错误的概率不超过 0.005.

规律方法 独立性检验的具体做法

- (1) 提出零假设 H_0 : X 和 Y 相互独立.
- (2) 根据实际问题的需要确定允许推断“事件 X 与 Y 有关系”犯错误的概率的上界 α ,然后查表确定临界值 x_α .
- (3) 利用公式 $\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ 计算随机变量 χ^2 .
- (4) 如果 $\chi^2 \geq x_\alpha$,那么“ X 与 Y 有关系”这种推断犯错误的概率不超过 α ;否则,就认为在犯错误的概率不超过 α 的前提下不能推断“ X 与 Y 有关系”,或者在样本数据中没有发现足够的证据支持结论“ X 与 Y 有关系”.

刷易错

★易错点 1 不理解独立性检验的基本思想

8. B 【解析】因为利用独立性检验时与样本的选取有关,所以得到的结论可能有误,因此 B 中说法错误. 故选 B.

易错警示 回归分析是对具有相关关系的两个变量进行统计分析的方法,独立性检验是利用随机变量 χ^2 来判断两个分类变量是否有关联的方法. 独立性检验得到的结论不一定可靠,只能做理论上的参考.

★易错点 2 对独立性检验的结果判断错误

9. D 【解析】∵ 概率 $P(\chi^2 \geq 10.828) \approx 0.001$, ∴ 两个变量有关联的可信度是 $1 - 0.001 = 99.9\%$. 故选 D.

易错警示 对独立性检验的结果进行描述时,要注意区分“在犯错误概率不超过 α 的前提下”与“有 $1 - \alpha$ 以上的把握”,同时要注意判断两个变量是有关还是无关.

第 8.3 节综合训练

刷能力

1. D 【解析】独立性检验假设有反证法的意味,应假设两分类变量(而非变量的属性)无关联,这时的 χ^2 应该很小,若 χ^2 很大,则可以否定假设,若 χ^2 很小,

则不能够肯定或者否定假设. 故选 D.

2. A 【解析】根据题中表格数据完成 2×2 列联表,如下:

性别	意愿程度		合计
	愿意参加	不愿意参加	
男生	30	20	50
女生	20	30	50
合计	50	50	100

则 $\chi^2 = \frac{100 \times (30 \times 30 - 20 \times 20)^2}{50 \times 50 \times 50 \times 50} = 4$. 故

选 A.

3. BC 【解析】独立性检验是通过卡方计算来判断两个分类变量存在关联的可能性的一种方法,并非检验二者是否线性相关, A 错误;

由独立性检验可知, χ^2 的值越大,则两个变量有关联的可能性就越大,故 B 正确;

残差平方和反映的是观测值与预测值间的差异,值越小,说明拟合效果越好,故 C 正确;

决定系数 R^2 越接近 1,模型的拟合效果越好,决定系数 R^2 越小,拟合效果越差,故 D 错误. 故选 BC.

4. AC 【解析】对于 A,在调查的高一学生中,科技类占比为 0.6,若用按比例分层随机抽样法抽取 20 人,则参加科技类的学生应为 $20 \times 0.6 = 12$ (人),故 A 正确;

对于 B,在调查的高二学生中,选择文艺类的人数为 $200 \times 0.6 = 120$,选择科技类的人数为 $200 \times 0.4 = 80$,选择文艺类比选择科技类的学生多 $120 - 80 = 40$ (人),故 B 错误;

对于 C,因为 $16 > 10.828$,所以依据 $\alpha = 0.001$ 的独立性检验,

可认为年级与校园活动偏好类型的选择有关联,此推断犯错误的概率不大于 0.001,故 C 正确;

对于 D,当 $\alpha = 0.01$ 时, $16 > 6.635$,则依据 $\alpha = 0.01$ 的独立性检验,

可认为年级与校园活动偏好类型的选择有关联,此推断犯错误的概率不大于 0.01,即有 99% 的把握认为年级与校园活动偏好类型的选择有关联,故 D 错误. 故选 AC.

5. 5 【解析】根据独立性检验思想可得,

$$\chi^2 = \frac{20x \cdot (16x^2 - 36x^2)^2}{10x \cdot 10x \cdot 10x \cdot 10x} = \frac{4x}{5} \geq$$

3.841,解得 $x \geq 4.80125$. 因为 $x \in \mathbf{N}^*$ 且 $x < 6$,所以 $x = 5$.

6. 【解】(1) 零假设为 H_0 : 模块工作状态与测试环境无关.

根据列联表中数据,得 $\chi^2 = \frac{100 \times (50 \times 10 - 35 \times 5)^2}{55 \times 45 \times 85 \times 15} \approx 3.347 <$

6.635 = $x_{0.01}$,

所以依据小概率值 $\alpha = 0.01$ 的独立性检验,没有充分证据推断 H_0 不成立,因此可以认为 H_0 成立,即认为模块工作状态与测试环境无关.

(2) (i) 由题意可知 $X \sim B\left(4, \frac{3}{4}\right)$,

$$P(X=0) = C_4^0 \times \left(\frac{3}{4}\right)^0 \times \left(\frac{1}{4}\right)^4 = \frac{1}{256},$$

$$P(X=1) = C_4^1 \times \left(\frac{3}{4}\right)^1 \times \left(\frac{1}{4}\right)^3 = \frac{12}{256} = \frac{3}{64},$$

$$P(X=2) = C_4^2 \times \left(\frac{3}{4}\right)^2 \times \left(\frac{1}{4}\right)^2 = \frac{54}{256} = \frac{27}{128},$$

$$P(X=3) = C_4^3 \times \left(\frac{3}{4}\right)^3 \times \left(\frac{1}{4}\right)^1 = \frac{108}{256} = \frac{27}{64},$$

$$P(X=4) = C_4^4 \times \left(\frac{3}{4}\right)^4 \times \left(\frac{1}{4}\right)^0 = \frac{81}{256},$$

则 X 的分布列如下:

X	0	1	2	3	4
P	$\frac{1}{256}$	$\frac{3}{64}$	$\frac{27}{128}$	$\frac{27}{64}$	$\frac{81}{256}$

$$E(X) = 4 \times \frac{3}{4} = 3.$$

(ii) 当 $n = 5$ 时记系统 M 中正常工作的模块个数为随机变量 Y ,则 $Y \sim B\left(5, \frac{3}{4}\right)$.

记 $n = 4$ 时系统 M 的可靠性为 P_1 , $n = 5$ 时系统 M 的可靠性为 P_2 .

$$\text{故 } P_1 = P(X=3) + P(X=4) = \frac{27}{64} +$$

悟: 有超过一半的模块正常工作,系统才正常工作

$$\frac{81}{256} = \frac{189}{256},$$

$$P_2 = P(Y=3) + P(Y=4) + P(Y=5) = C_5^3 \times \left(\frac{3}{4}\right)^3 \times \left(\frac{1}{4}\right)^2 + C_5^4 \times \left(\frac{3}{4}\right)^4 \times \left(\frac{1}{4}\right)^1 + C_5^5 \times \left(\frac{3}{4}\right)^5 \times \left(\frac{1}{4}\right)^0 = \frac{270}{1024} + \frac{405}{1024} + \frac{243}{1024} = \frac{918}{1024} = \frac{459}{512},$$

$$\text{故 } P_1 - P_2 = \frac{189}{256} - \frac{459}{512} = \frac{378}{512} - \frac{459}{512} < 0, \text{ 即}$$

$$P_1 < P_2,$$

故增加一个模块(即 $n = 5$)能提高系统 M 的可靠性.

第八章素养检测

刷速度

1. C 【解析】对于 A, χ^2 小于临界值 $x_{0.05}$, 并不意味着一定相互独立, 只是无足够证据反对独立, 故 A 错误; 对于 B, χ^2 小于临界值 $x_{0.05}$, 并不意味着一定不独立, 故 B 错误; 对于 C, 这是独立性检验的基本逻辑: 当 $\chi^2 < x_{0.05}$ 时, 无充分证据支持变量有关, 故 C 正确; 对于 D, $\chi^2 > x_{0.05}$ 对应 95% 的把握认为两个变量有关联, 而实际上 $\chi^2 < x_{0.05}$, 故 D 错误. 故选 C.

2. D 【解析】对于 A: 散点杂乱无章, 无规律可言, 看不出两个变量有什么相关性, 故 A 错误; 对于 B: 两个变量不具有线性相关性, 故 B 错误; 对于 C: 两个变量之间的关系为负相关关系, 故 C 错误; 对于 D: 两个变量之间的关系为正相关关系, 且散点图中的点分布在一条直线附近, 线性相关程度较高, 故 D 正确. 故选 D.

3. D 【解析】因为 $\sum_{i=1}^9 x_i = 33$, 所以 $\bar{x} = \frac{1}{9} \sum_{i=1}^9 x_i = \frac{11}{3}$, 所以 $\bar{y} = 2\bar{x} - 1 = \frac{19}{3}$. 所以新增加数据 $(-3, 3)$ 后, $\bar{x}' = \frac{33-3}{10} = 3$, $\bar{y}' = \frac{9 \times \frac{19}{3} + 3}{10} = 6$, 因为新的经验回归直线过点 $(3, 6)$, 且斜率为 2.1, 所以 $6 = 2.1 \times 3 + \hat{a}$, 得 $\hat{a} = -0.3$, 所以修正后的经验回归方程为 $\hat{y} = 2.1x - 0.3$. 则当 $x = 4$ 时, $\hat{y} = 2.1 \times 4 - 0.3 = 8.1$, 则修正后数据 $(4, 8)$ 的残差为 $8 - 8.1 = -0.1$. 故选 D.

4. B 【解析】依题意, 设男、女学生的人数都为 $5x, x \in \mathbf{N}^*$, 则男、女学生的总人数为 $10x$, 可得 2×2 列联表如下,

性别	网络课程的热衷程度		合计
	喜欢网络课程	不喜欢网络课程	
男生	4x	x	5x
女生	3x	2x	5x
合计	7x	3x	10x

故 $\chi^2 = \frac{10x \cdot (8x^2 - 3x^2)^2}{5x \cdot 5x \cdot 7x \cdot 3x} = \frac{10x}{21}$, 由题意可得 $6.635 \leq \frac{10x}{21} < 10.828$,

所以 $139.335 \leq 10x < 227.388$. 结合选项可知, 只有 B 符合题意. 故选 B.

5. C 【解析】根据题图①可知样本中选择物理学科的人数较多, 故 C 正确; 根据题图②可知样本中男生人数多于女生人数, 故 D 错误; 样本中选择物理的人数多于选择历史的人数, 而选择物理的男生比例高, 选择历史的女生比例低, 所以样本中选择物理的男生人数多于选择历史的女生人数, 故 A 错误; 根据题图②可知样本中女生选择历史的人数少于男生选择历史的人数, 故 B 错误. 故选 C.

6. C 【解析】对于 A, B, $a + c = 1\,000 \times 40\% = 400$, $b + d = 1\,000 - 400 = 600$, $a + b = 1\,000 \times 20\% = 200$, $c + d = 1\,000 - 200 = 800$, 因为每天玩手机超过 1 小时的学生近视率为 50%, 所以 $a = b$, 所以 $a = 100$, $b = 100$, 则 $c = 300$, $d = 500$, 故 A, B 错误; 对于 C, D, 零假设 H_0 : 认为玩手机时长与是否近视之间无关, 则 $\chi^2 = \frac{1\,000 \times (50\,000 - 30\,000)^2}{200 \times 800 \times 400 \times 600} \approx 10.417 > 7.879 = x_{0.005}$, 所以依据小概率值 $\alpha = 0.005$ 的独立性检验, 我们推断 H_0 不成立, 即认为玩手机时长与是否近视之间有关, 即在犯错误概率不超过 0.005 的前提下, 可以认为玩手机时长与是否近视之间有关, 故 C 正确, D 错误. 故选 C.

7. B 【解析】 $\bar{x} = \frac{3+4+5+6+6+7+8+9}{8} = 6$, $\sum_{i=1}^8 (x_i - \bar{x})^2 = (3-6)^2 + (4-6)^2 + (5-6)^2 + 2 \times (6-6)^2 + (7-6)^2 + (8-6)^2 + (9-6)^2 = 28$, 由条件可知 $r = \frac{\sum_{i=1}^8 (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^8 (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^8 (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^8 (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{28 \times 252}} = \frac{83}{84}$, 得 $\sum_{i=1}^8 (x_i - \bar{x})(y_i - \bar{y}) = 83$, 所以 $\hat{b} = \frac{\sum_{i=1}^8 (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^8 (x_i - \bar{x})^2} = \frac{83}{28}$, 故选 B.

8. C 【解析】 $\bar{x} = \frac{1+2+3+4+5}{5} = 3$, 经验回归直线必过样本点的中心 $(\bar{x}, \bar{y})$, 把

$\bar{x} = 3$ 代入经验回归方程得 $\bar{y} = 20 \times 3 + 11 = 71$.

根据题中表格数据, 得 $\bar{y} = \frac{31+9a+80+91+108}{5} = \frac{310+9a}{5}$,

因此 $\frac{310+9a}{5} = 71$, 解得 $a = 5$.

选项 A: 2026 年对应 $x = 6$, 代入 $\hat{y} = 20x + 11$ 得预测值 $\hat{y} = 131$, 该值为估计值, 不是确定值, 故 A 错误;

选项 B: $a = 5 \neq 6$, 故 B 错误;

选项 C: 当 $x = 2$ 时, 观测值 $y = 9a = 45$, 预测值 $\hat{y} = 20 \times 2 + 11 = 51$, 残差为 $45 - 51 = -6$, 故 C 正确;

❶ 敲黑板: 残差 = 观测值 - 预测值

选项 D: $a = 5$, 点 $(4, 20a)$ 即 $(4, 100)$, 将 $x = 4$ 代入经验回归方程得 $\hat{y} = 20 \times 4 + 11 = 91 \neq 100$, 故该点不在经验回归直线上, D 错误. 故选 C.

9. AC 【解析】对于 A, 两个变量的线性相关性越强, $|r|$ 越大, A 错误; 对于 B, 根据经验回归方程的意义, 样本点有可能都不在经验回归直线上, B 正确; 对于 C, χ^2 的值越大, 判断“两变量有关联”犯错误的概率越小, C 错误; 对于 D, 若 $\chi_1^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$, 每个数据都变为原来的 2 倍, 则 $\chi_2^2 = \frac{2n(2a \cdot 2d - 2b \cdot 2c)^2}{(2a+2b)(2c+2d)(2a+2c)(2b+2d)} = 2\chi_1^2$, D 正确. 故选 AC.

10. ACD 【解析】先补全列联表, 车流量 $\geq 1\,500$ 辆/h 且 $\text{PM}_{2.5} \geq 75$ 的频数为 $11 - 4 = 7$, 车流量 $< 1\,500$ 辆/h 且 $\text{PM}_{2.5} < 75$ 的频数为 $12 - 4 = 8$, 车流量 $\geq 1\,500$ 辆/h 的合计频数为 $7 + 1 = 8$, $\text{PM}_{2.5} < 75$ 的合计频数为 $1 + 8 = 9$.

	车流量 $\geq 1\,500$	车流量 $< 1\,500$	合计
$\text{PM}_{2.5} \geq 75$	7	4	11
$\text{PM}_{2.5} < 75$	1	8	9
合计	8	12	20

对于选项 A, 车流量 $\geq 1\,500$ 辆/h 时, $\text{PM}_{2.5} \geq 75$ 的概率估计值为 $\frac{7}{8}$, A 正确.

对于选项 B, 统计量 χ^2 的值越大, 说明车流量大小与空气污染的关联性越强, B 错误.

高中必刷题 数学

对于选项 C, 零假设为 H_0 : 车流量大小与空气污染无关联, 则 $\chi^2 = \frac{20 \times (7 \times 8 - 4 \times 1)^2}{11 \times 9 \times 8 \times 12} \approx 5.69 > 5.024 =$

$\chi_{0.025}$, 根据小概率值 $\alpha = 0.025$ 的独立性检验, 我们推断 H_0 不成立, 即认为车流量大小与空气污染有关联, 此推断犯错误的概率不大于 0.025, C 正确.

对于选项 D, $\chi^2 \approx 5.69 < 6.635 = \chi_{0.01}$, 故根据小概率值 $\alpha = 0.01$ 的独立性检验, 没有充分证据推断 H_0 不成立, 因此可以认为 H_0 成立, 即认为车流量大小与空气污染没有关联, D 正确. 故选 ACD.

11. ACD 【解析】设 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} =$

$\frac{1}{n} \sum_{i=1}^n y_i$, 则 $\bar{x} = x_{n+1}, \bar{y} = y_{n+1}$, 所以 $\frac{1}{n+1} \sum_{i=1}^{n+1} x_i = \bar{x}, \frac{1}{n+1} \sum_{i=1}^{n+1} y_i = \bar{y}$.

对于 A, B, $\sum_{i=1}^{n+1} (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + (x_{n+1} - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})^2$,

同理, $\sum_{i=1}^{n+1} (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \bar{y})^2$, $\sum_{i=1}^{n+1} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$,

则样本相关系数 $r' = \frac{\sum_{i=1}^{n+1} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n+1} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n+1} (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = r$,

可得样本相关系数不变, 所以变量 x 与 y 的线性相关性不变, 故 A 正确, B 错误;

对于 C, D, 因为 $\hat{b}' = \frac{\sum_{i=1}^{n+1} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n+1} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \hat{b}$, 且经验回归

直线过点 $(\bar{x}, \bar{y})$, 所以 $\hat{b}, \hat{a}$ 均不变, 所以经验回归方程不变, 故 C, D 正确.

【点悟】: 增加一个在经验回归直线上的点, 对变量之间的线性关系无影响

故选 ACD.

12. 有 【解析】由题意可得 2×2 列联表如下:

语文成绩	英语成绩		合计
	英语成绩及格	英语成绩不及格	
语文成绩及格	20	4	24
语文成绩不及格	5	11	16
合计	25	15	40

零假设为 H_0 : 学生的英语成绩与语文成绩无关.

根据列联表中的数据得 $\chi^2 = \frac{40 \times (20 \times 11 - 4 \times 5)^2}{24 \times 16 \times 25 \times 15} \approx 11.111 >$

$6.635 = \chi_{0.01}$, 根据小概率值 $\alpha = 0.01$ 的独立性检验, 我们推断 H_0 不成立, 即认为学生的英语成绩与语文成绩有关联, 此推断犯错误的概率不大于 0.01, 即有 99% 的把握认为学生的英语成绩与语文成绩有关.

13. 42 0.99 【解析】计算得 $\bar{x} = 11, \bar{y} =$

$24, \sum_{i=1}^5 x_i y_i = 917, \sum_{i=1}^5 x_i^2 = 811,$

所以 $\hat{b} = \frac{\sum_{i=1}^5 x_i y_i - 5 \bar{x} \bar{y}}{\sum_{i=1}^5 x_i^2 - 5 \bar{x}^2} \approx -1.96$, 所以

以 $\hat{a} = \bar{y} - \hat{b} \bar{x} \approx 46$, 故经验回归方程为 $\hat{y} = -2x + 46$, 当 $x = 2$ 时, $\hat{y} = 42$, 所以可估计其保鲜时长约为 42 小时. 结合已知与经验回归方程可得 $\sum_{i=1}^5 (y_i - \hat{y}_i)^2 = 10, \sum_{i=1}^5 (y_i - \bar{y})^2 = 798$, 所以 $R^2 \approx 0.99$.

14. -0.5 【解析】因为 $\frac{1}{7} \sum_{i=1}^7 \ln(x_i -$

$1.895) = \frac{23.10}{7} = 3.3, \frac{1}{7} \sum_{i=1}^7 y_i =$

$\frac{73.50}{7} = 10.5$,

所以 $10.5 = 3.3a + 12.15$, 解得 $a = -0.5$.

15. 【解】(1) 依题意, 得到 2×2 列联表如下:

	喜欢足球	不喜欢足球	合计
中学生	60	40	100
非中学生	30	70	100
合计	90	110	200

零假设为 H_0 : 该校学生喜欢足球与中学生身份无关.

根据列联表数据, 计算得到 $\chi^2 = \frac{200 \times (60 \times 70 - 40 \times 30)^2}{100 \times 100 \times 90 \times 110} \approx 18.182 >$

10.828,

根据小概率值 $\alpha = 0.001$ 的独立性检

验, 推断 H_0 不成立, 即认为该校学生喜欢足球与中学生身份有关.

(2) 依题意, 3 人进球总次数 ξ 的所有可能取值为 0, 1, 2, 3,

则 $P(\xi = 0) = \left(\frac{1}{5}\right)^2 \times \frac{1}{4} = \frac{1}{100}$,

$P(\xi = 1) = C_2^1 \times \frac{4}{5} \times \frac{1}{5} \times \frac{1}{4} + \left(\frac{1}{5}\right)^2 \times \frac{3}{4} = \frac{11}{100}$,

$P(\xi = 2) = C_2^1 \times \frac{4}{5} \times \frac{1}{5} \times \frac{3}{4} + \left(\frac{4}{5}\right)^2 \times \frac{1}{4} = \frac{2}{5}, P(\xi = 3) = \left(\frac{4}{5}\right)^2 \times \frac{3}{4} = \frac{12}{25}$,

所以随机变量 ξ 的分布列为:

ξ	0	1	2	3
P	$\frac{1}{100}$	$\frac{11}{100}$	$\frac{2}{5}$	$\frac{12}{25}$

所以 ξ 的数学期望为 $E(\xi) = 0 \times \frac{1}{100} +$

$1 \times \frac{11}{100} + 2 \times \frac{2}{5} + 3 \times \frac{12}{25} = \frac{47}{20}$.

16. 【解】(1) 由数据得 $\bar{x} = \frac{1+2+3+4+5}{5} =$

$3, \bar{y} = \frac{2+6+6+7+9}{5} = 6$.

于是 $\sum_{i=1}^5 (x_i - \bar{x})^2 = (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 = 10$.

又 $\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y}) = -2 \times (-4) + (-1) \times 0 + 0 \times 0 + 1 \times 1 + 2 \times 3 = 15$.

所以 $\hat{b} = \frac{\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^5 (x_i - \bar{x})^2} =$

$\frac{15}{10} = \frac{3}{2}$.

从而 $\hat{a} = \bar{y} - \hat{b} \bar{x} = 6 - \frac{3}{2} \times 3 = \frac{3}{2}$.

故经验回归方程为 $\hat{y} = \frac{3}{2}x + \frac{3}{2}$.

(2) 当 $x = 6$ 时, $\hat{y} = \frac{3}{2} \times 6 + \frac{3}{2} = \frac{21}{2} = 10.5$.

所以估计每周锻炼 6 小时时体测成绩的提高分数为 10.5.

若训练效果明显, 则 $\hat{y} \geq 10$,

由 $\frac{3}{2}x + \frac{3}{2} \geq 10$, 解得 $x \geq \frac{17}{3}$.

由于每周锻炼时间取整数, 所以估计每周锻炼时间至少应为 6 小时.

17. 【解】(1) 零假设为 H_0 : 观看这两部影片的观众的男女比例没有差异.

根据列联表中的数据, 经计算得到

$\chi^2 = \frac{200 \times (50 \times 60 - 50 \times 40)^2}{100 \times 100 \times 90 \times 110} = \frac{200}{99} \approx$

2.020,

因为 $2.020 < 2.706 = x_{0.1}$, 所以根据小概率率 $\alpha = 0.1$ 的独立性检验, 没有充分证据推断 H_0 不成立, 因此可以认为观看这两部影片的观众的男女比例没有差异.

(2) 由题意知 $\chi^2 =$

$$\frac{kn(ka \cdot kd - kb \cdot kc)^2}{k(a+b) \cdot k(c+d) \cdot k(a+c) \cdot k(b+d)} = \frac{kn(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{200k}{99},$$

$$\text{令 } \frac{200k}{99} \geq 6.635, \text{ 即 } k \geq \frac{6.635 \times 99}{200},$$

$$\text{又 } \frac{6.635 \times 99}{200} \approx 3.284, k \in \mathbf{N}^*, \text{ 所以 } k$$

的最小值为 4.

18. 【解】(1) 由题中表格可知散点增长趋势较为平缓, 故选择对数函数模型 $y = a + b \ln x$.

(2) 令 $u = \ln x$, 则模型化为线性形式 $y = a + bu$,

$$\text{则 } \hat{b} = \frac{\sum_{i=1}^6 (y_i - \bar{y})(u_i - \bar{u})}{\sum_{i=1}^6 (u_i - \bar{u})^2} \approx \frac{3.35}{0.67} = 5,$$

$$\text{又 因 为 } \bar{y} = \frac{5.2 + 6.7 + 7.8 + 8.7 + 9.4 + 10.2}{6} = 8,$$

$$\bar{u} \approx 1.65,$$

$$\text{所以 } \hat{a} = \bar{y} - \hat{b}\bar{u} \approx 8 - 5 \times 1.65 = -0.25,$$

故所求非线性经验回归方程为 $\hat{y} = -0.25 + 5 \ln x$.

$$(3) \text{ 当 } x = 14 \text{ 时, } \hat{y} = -0.25 + 5 \ln 14 \approx -0.25 + 5 \times 2.64 = 12.95,$$

所以当光照时长 x 为 14 h 时, 预估花卉株高增长量为 12.95 cm.

19. 【解】(1) 已知 $\{x_i\}$ 是首项为 2, 公差为 4 的等差数列, $u_i = x_i^2$,

则由等差数列前 n 项和公式得,

$$\sum_{i=1}^{10} x_i^2 = \sum_{i=1}^{10} (4i-2)^2 = 16 \sum_{i=1}^{10} i^2 -$$

$$16 \sum_{i=1}^{10} i + 40 = 16 \times \frac{10 \times 11 \times 21}{6} - 16 \times$$

$$\frac{10 \times 11}{2} + 40 = 5320, \text{ 故 } \bar{u} = \frac{5320}{10} = 532.$$

对于模型 $y = \alpha x^2 + \beta + e$,

令 $u = x^2$, 则 $y = \alpha u + \beta + e$.

悟: 通过换元将非线性模型转换成线性模型, 易求解

$$\text{又 } \hat{\alpha} = \frac{\sum_{i=1}^{10} (u_i - \bar{u})(y_i - \bar{y})}{\sum_{i=1}^{10} (u_i - \bar{u})^2} =$$

$$\frac{128448}{2247168} = \frac{669}{11704} \approx 0.057, \hat{\beta} = \bar{y} -$$

$$\hat{\alpha}\bar{u} = 32 - \frac{669}{11704} \times 532 \approx 1.591,$$

因此乙团队建立的一元非线性经验回归方程为 $\hat{y} = 0.057x^2 + 1.591$.

(2) 甲团队(线性模型): 当 $x = 40$ 时, $\hat{y}_{\text{甲}} = 2.2 \times 40 - 12 = 76$, 残差的绝对值为 $|91 - 76| = 15$;

乙团队(非线性模型): 当 $x = 40$ 时, $\hat{y}_{\text{乙}} = 0.057 \times 40^2 + 1.591 = 92.791$, 残差的绝对值为 $|91 - 92.791| = 1.791$.

因为乙团队模型的残差的绝对值更小,

所以乙团队的模型更优, 能更好地拟合数据, 反映年研发资金投入量与年销售额之间的关系.

(3) 已知 $y \sim N(\mu, 7.795^2)$, 要保证有 97.725% 的把握获得年销售额 100 亿元, 需 $\mu - 2\sigma \geq 100$,

$$\text{悟: } P(X \geq \mu - 2\sigma) \approx 0.5 + \frac{0.9545}{2} = 0.97725$$

代入乙模型 $\mu = 0.057x^2 + 1.591, \sigma = 7.795$,

$$\text{得 } 0.057x^2 + 1.591 - 2 \times 7.795 \geq 100,$$

$$\text{解得 } x \geq \sqrt{\frac{113999}{57}} \approx 44.721 (\text{亿元}).$$

故预计至少需要投入研发资金 44.73 亿元.

第八章高考强化

刷真题

1. A 【解析】对于 A, 由经验回归方程

$$\hat{y} = -1.17x + 1370.7, \text{ 知 } \hat{b} = -1.17 < 0,$$

所以 y 与 x 呈负相关, 故 A 正确; 对于 B, C, 当 $x = 10$ 时, $\hat{y} = -1.17 \times 10 + 1370.7 = 1359$, 但该值仅为预测值, 并非准确值, 故 B 错误, C 错误; 对于 D, 因为 $r = -0.91$, $|r|$ 接近 1, 说明两变量线性相关性强, 故 D 错误. 故选 A.

2. B 【解析】A 选项, 由 $X \sim N(\mu, \sigma^2)$,

可知正态曲线关于直线 $x = \mu$ 对称, 故有 $P(X \leq \mu - \sigma) = P(X \geq \mu + \sigma)$, A 正确;

B 选项, 由 $X \sim N(1, 2^2)$, 可知 $P(X < 1) = 0.5$,

由 $Y \sim N(2, 2^2)$, 可知 $P(Y < 2) = 0.5$, 故 $P(X < 1) = P(Y < 2)$, B 错误;

C 选项, 样本相关系数 r 的范围是 $[-1, 1]$, $|r|$ 越接近于 1, 则两个变量之间线性相关程度越强, C 正确;

D 选项, 样本相关系数 r 接近于 0 时, 两个变量之间几乎没有线性相关性, D 正确. 故选 B.

3. 【解】(1) 9 年间共有 7 年此颗粒物密度大于二氧化硫密度, 故概率为 $\frac{C_7^1}{C_9^1} = \frac{7}{9}$.

(2) 统计图表需要呈现出当二氧化硫密度变化时, 此颗粒物密度的变化趋

势, 故需要散点图进行呈现.

随着二氧化硫密度的增加, 此颗粒物密度呈现增加趋势, 故二者正相关, 相关系数为正, 又因为相关系数 r 满足 $|r| \leq 1$, 故相关系数在区间 $(0, 1)$ 内.

(3) 采用方程 $y = 105.644e^{-0.461(x-2.014)}$ 时, 2023 年的预测值 $y_1 = 105.644e^{-0.461(2.023-2.014)} \approx 1.667$, 预测值与实际值的差值的绝对值约为 $|1.667 - 3.881| = 2.213$.

$$\text{由题知, } \bar{x} = 2.018, \bar{y} = \frac{1}{9} \times (101.02 +$$

$$87.02 + 57.46 + 21.85 + 11.76 + 8.86 +$$

$$5.03 + 4.63 + 3.86) \approx 33.499,$$

$$\text{代入回归方程得 } 33.499 = a(2.018 - 2.014) + 83.743, \text{ 解得 } a = -12.561,$$

悟: 回归直线过样本点的中心
采用方程 $y = a(x - 2.014) + 83.743$ 时, 2023 年的预测值 $y_2 = -12.561 \times (2.023 - 2.014) + 83.743 = -29.306$, 预测值与实际值的差值的绝对值约为 $|-29.306 - 3.881| = 33.186$.

因为 $2.213 < 33.186$,

所以采用方程 $y = 105.644e^{-0.461(x-2.014)}$

对于 2023 年的预测值与实际值的差值的绝对值更小.

4. 【解】(1) 由题意可知 $\sum_{i=1}^{10} x_i = 0.6, \sum_{i=1}^{10} y_i = 3.9$, 所以这 10 棵树木的根部横截面积和

$$\text{材积量的平均值分别为 } \bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i =$$

0.06, $\bar{y} = \frac{1}{10} \sum_{i=1}^{10} y_i = 0.39$, 所以估计该林区这种树木平均一棵的根部横截面积与平均一棵的材积量分别为 0.06 m^2 和 0.39 m^3 .

$$\begin{aligned} (2) \text{ 因为 } r &= \frac{\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{10} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{10} (y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^{10} x_i y_i - 10 \bar{x} \bar{y}}{\sqrt{\sum_{i=1}^{10} x_i^2 - 10 \bar{x}^2} \cdot \sqrt{\sum_{i=1}^{10} y_i^2 - 10 \bar{y}^2}} \\ &= \frac{0.2474 - 10 \times 0.06 \times 0.39}{\sqrt{0.038 - 10 \times 0.06^2} \times \sqrt{1.6158 - 10 \times 0.39^2}} \\ &= \frac{0.0134}{\sqrt{0.002} \times \sqrt{0.0948}} = \frac{0.0134}{\sqrt{1.896 \times 10^{-4}}} \\ &\approx \frac{134}{137.7} \approx 0.97, \end{aligned}$$

所以该林区这种树木的根部横截面积与材积量的样本相关系数是 0.97.

(3) 因为树木的材积量与其根部横截面积近似成正比, 所以设 $y = kx$ (k 为常数).

由 (1) 知该林区这种树木平均一棵的根部横截面积与平均一棵的材积量的估计值分别为 0.06 和 0.39,

$$\text{所以 } 0.39 = 0.06k, \text{ 即 } k = \frac{13}{2}, \text{ 所以 } y = \frac{13}{2}x.$$

又因为 $\sum_{i=1}^n x_i = 186$, 所以 $\sum_{i=1}^n y_i = \frac{13}{2} \sum_{i=1}^n x_i =$

$$\frac{13}{2} \times 186 = 1209,$$

故该林区这种树木的总材积量的估计值为 $1\,209\text{ m}^3$.

5. 【解】(1) 根据题表数据可知, 超声波检查结果不正常的有 200 人, 其中患该疾病的有 180 人, 因此估计超声波检查结果不正常者患该疾病的概率 $p = \frac{180}{200} = \frac{9}{10}$.

(2) 零假设为 H_0 : 超声波检查结果与患该疾病无关.

$$\chi^2 = \frac{1\,000 \times (20 \times 20 - 180 \times 780)}{200 \times 800 \times 800 \times 200} =$$

$$765.625 > 10.828.$$

根据小概率值 $\alpha = 0.001$ 的独立性检验, 超声波检查结果与患该疾病有关, 此推断犯错误的概率不大于 0.001.

链接教材 本题是教材第 133 页例 4 的同类试题, 考查利用独立性检验分析两类变量是否有关联, 难度及运算量都不大, 步骤中注意给出零假设及判断结果.

基于小概率值 α 的独立性检验是: 当 $\chi^2 \geq x_\alpha$ 时, 我们就推断 H_0 不成立, 即认为 X 和 Y 不独立, 该推断犯错误的概率不超过 α ; 当 $\chi^2 < x_\alpha$ 时, 我们没有充分证据推断 H_0 不成立, 可以认为 X 和 Y 独立.

6. 【解】(1) 补全列联表如下:

	优品	非优品
甲车间	26	24
乙车间	70	30

$$K^2 = \frac{150 \times (26 \times 30 - 24 \times 70)^2}{50 \times 100 \times 96 \times 54} = \frac{75}{16} =$$

4.6875 > 3.841, 4.6875 < 6.635, 所以有 95% 的把握认为甲、乙两车间产品的优品率存在差异, 没有 99% 的把握认为甲、乙两车间产品的优品率存在差异.

$$(2) \text{由题中数据知, } \bar{p} = \frac{96}{150} = 0.64,$$

$$p + 1.65 \sqrt{\frac{p(1-p)}{n}} = 0.5 + 1.65 \cdot$$

$$\sqrt{\frac{0.5 \times 0.5}{150}} = 0.5 + 1.65 \times \frac{0.5}{\sqrt{150}} \approx 0.5 +$$

$$1.65 \times \frac{0.5}{12.247} \approx 0.57.$$

$$\text{又 } 0.64 > 0.57,$$

所以可以认为生产线智能化升级改造后, 该工厂产品的优品率提高了.

刷原创

1. C 【解析】由题可得 $2.706 = x_{0.1} < \chi^2 = 2.974 < x_{0.05} = 3.841$, 则根据小概率值 α 的独立性检验知, 变量 x, y 不相互独立, 这个结论犯错误的概率不超过 0.1. 故选 C.

2. 【解】(1) 作出散点图, 由图知 $y = bx + a$ 更加适合, 由题可知 $\bar{x} = \frac{1+2+3+4+5}{5} = 3$, $\bar{y} = \frac{22+39+59+82+98}{5} = 60$, $\sum_{i=1}^5 x_i y_i = 1 \times 22 + 2 \times 39 + 3 \times$

$$59 + 4 \times 82 + 5 \times 98 = 1\,095,$$

$$\sum_{i=1}^5 x_i^2 - 5\bar{x}^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 - 5 \times 3^2 = 10,$$

$$\hat{b} = \frac{\sum_{i=1}^5 x_i y_i - 5\bar{x}\bar{y}}{\sum_{i=1}^5 x_i^2 - 5\bar{x}^2} = \frac{1\,095 - 5 \times 3 \times 60}{10} =$$

19.5, 所以 $\hat{a} = 60 - 19.5 \times 3 = 1.5$, 所以 y 关于 x 的经验回归方程为 $\hat{y} = 19.5x + 1.5$.

(2) 将 $x = 6$ 代入上述经验回归方程, 得 $\hat{y} = 19.5 \times 6 + 1.5 = 118.5$, 所以预测第 6 个月的销售额约为 119 万元.

(3) 补全缺失数据后的 2×2 列联表如下:

	喜欢	不喜欢	总计
男	40	50	90
女	80	30	110
总计	120	80	200

零假设为 H_0 : 消费者选择意向与性别不存在关联.

$$\text{则 } \chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{200 \times (40 \times 30 - 80 \times 50)^2}{120 \times 80 \times 110 \times 90} \approx 16.5 >$$

10.828, 所以依据小概率值 $\alpha = 0.001$ 的独立性检验, 推断 H_0 不成立, 可以认为消费者选择意向与性别存在关联.

专练 1 新情境、新定义专练

刷素养

1. D 【解析】因为 $1\,260 = 2 \times 2 \times 3 \times 3 \times 5 \times 7$, 所以从 2, 2, 3, 3, 5, 7 中任选 3 个数组成三位数, 可以分成两类:

①若三个数都不相同, 则可以组成三位数 $A_4^3 = 24$ 个;

②若含有 2 个 2 或 2 个 3, 则可以组成三位数 $2 \times C_3^1 \times C_3^1 = 18$ 个.

故由分类加法计数原理, 共可以组成不同三位数的个数为 $24 + 18 = 42$. 故选 D.

2. ABD 【解析】对 A: 观察方阵可知, $a_{i,j} = C_{i+j-2}^{i-1}$, 所以 $a_{j,i} = C_{j+i-2}^{j-1}$, 根据组合数性质得 $a_{j,i} = C_{j+i-2}^{j-1} = C_{j+i-2}^{(j+i-2)-(j-1)} = C_{i+j-2}^{i-1} = a_{i,j}$, 故 A 正确; 对 B: 由 A 分析知 $a_{i,j} = C_{i+j-2}^{i-1}$, 所以 $a_{i-1,j} + a_{i,j-1} = C_{i+j-3}^{i-2} + C_{i+j-3}^{i-1} = C_{i+j-2}^{i-1} = a_{i,j}$, 即 $a_{i,j} = a_{i-1,j} + a_{i,j-1}$ ($i > 1, j > 1$), 故 B 正确;

对 C: 由 A 分析知 $a_{i,j} = C_{i+j-2}^{i-1}$, 可得 $a_{i,1} + a_{i,2} + a_{i,3} = C_{i-1}^{i-1} + C_i^{i-1} + C_{i+1}^{i-1} = C_i^i + C_{i+1}^{i-1} = C_{i+1}^i + C_{i+1}^{i-1} = C_{i+2}^i = a_{i+1,3} \neq a_{i+1,4}$, 故 C 错误;

对 D: 由组合数公式得 $a_{i,j} = C_{i+j-2}^{i-1} =$

$\frac{(i+j-2)!}{(i-1)!(j-1)!}$, 故 D 正确. 故选 ABD.

3. 4 【解析】依题意, 1 的质数分布区间 (1, 4) 内的质数有 2, 3 两个; 2 的质数分布区间 (4, 9) 内的质数有 5, 7 两个; 3 的质数分布区间 (9, 16) 内的质数有 11, 13 两个; 4 的质数分布区间 (16, 25) 内的质数有 17, 19, 23 三个; 5 的质数分布区间 (25, 36) 内的质数有 29, 31 两个; 6 的质数分布区间 (36, 49) 内的质数有 37, 41, 43, 47 四个.

因此从集合 $\{1, 2, 3, 4, 5, 6\}$ 中任取两个元素, 则它们在其质数分布区间上的质数个数之和为 5 的选取方案为取 1 和 4, 2 和 3, 4 和 5, 共 4 种方案.

4. $C_n^k C_n^2 C_{n-3}^2$ 【解析】由题意得, $(a+b+c)^n = [(a+b)+c]^n = C_n^0 (a+b)^n + C_n^1 (a+b)^{n-1} c + C_n^2 (a+b)^{n-2} c^2 + \cdots + C_n^k (a+b)^{n-k} c^k + \cdots + C_n^n c^n$ ($0 \leq k \leq n$ 且 $k \in \mathbf{N}$), 则 $(a+b)^{n-k} c^k$ 的系数为 C_n^k . 在 n 个因式 $(a+b+c)$ 中取 3 个 a , 2 个

$b, n-5$ 个 c 相乘就是 $a^3 b^2 c^{n-5}$, 其方法数为 $C_n^3 C_{n-3}^2 C_{n-5}^{n-5} = C_n^3 C_{n-3}^2$, 故 $a^3 b^2 c^{n-5}$ 的系数为 $C_n^3 C_{n-3}^2$.

5. D 【解析】不超过 30 的自然数有 31 个, 其中素数有 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 共 10 个, 孪生素数有 3 和 5, 5 和 7, 11 和 13, 17 和 19, 共 4 组.

$$\text{所以 } P(A) = \frac{C_{10}^2}{C_{31}^2} = \frac{45}{C_{31}^2}, P(AB) =$$

$$\frac{C_{10}^2 - 4}{C_{31}^2} = \frac{41}{C_{31}^2},$$

点悟: 正难则反思想的运用

$$\text{所以 } P(B|A) = \frac{P(AB)}{P(A)} = \frac{\frac{41}{C_{31}^2}}{\frac{45}{C_{31}^2}} = \frac{41}{45}. \text{ 故}$$

选 D.

6. D 【解析】依题意, 小球从最上层 3 个缝隙中的一个落下的概率都相等, 往后每一层左右两边落下的概率相同, 由对称性可知, $P_1 = P_8 = \frac{1}{3}$.

$$\left(\frac{1}{2}\right)^5 = \frac{1}{96}, P_2 = P_7 = \frac{1}{3} \cdot (C_5^1 + C_5^0) \cdot$$