第八章　成对数据的统计分析

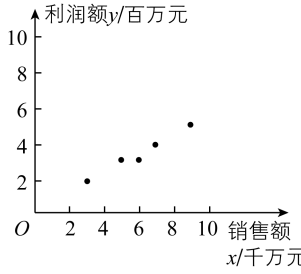
8.1　成对数据的统计相关性

8.1.1　变量的相关关系+8.1.2　样本相关系数

1. D　【解析】选项 A,B 中两个变量间是函数关系;选项 C 中两个变量之间没有什么关系;
选项 D 中,学习成绩与平均学习时间有关,但不仅与时间有关,还与其他变量有关,如学习时的专注性,个人的学习习惯等,因此 D 中两个变量是相关关系. 故选 D.

2. C　【解析】A 中的散点无规律可言,看不出两个变量有什么相关性;
B 中两个变量具有正相关关系;
C 中两个变量具有负相关关系;
D 中两个变量具有相关性,但既不是正相关,也不是负相关. 故选 C.

3. 【解】根据该连锁经营公司的 5 个零售店某月的销售额和利润额资料画出散点图如图所示.



从图中可以看出,这两个变量具有线性相关关系,且是正相关.

4. A　【解析】由题设 $1>|r_1|>|r_4|>|r_2|>|r_3|>0$,则线性相关程度最强的是 A 组成对样本数据(提示:$|r|$ 越接近 1,相关性越强,易直接错选成 r 较大组). 故选 A.

5. 【解】(1)样本中 10 个这种零件的横截面面积的平均值 $\bar{x}=\frac{0.52}{10}=0.052$,
样本中 10 个这种零件的耗材量的平均值 $\bar{y}=\frac{3.90}{10}=0.39$,

据此可估计刘铭同学制作的这种零件平均每个的横截面面积为 0.052 mm^2，平均一个零件的耗材量为 0.39 mm^3.

(2) 样本相关系数 $r=\dfrac{\sum\limits_{i=1}^{10}x_iy_i-10\bar{x}\,\bar{y}}{\sqrt{\left(\sum\limits_{i=1}^{10}x_i^2-10\bar{x}^2\right)\left(\sum\limits_{i=1}^{10}y_i^2-10\bar{y}^2\right)}}$

$=\dfrac{0.0115}{\sqrt{0.000\,149\,136}}$

$\approx\dfrac{0.0115}{0.012\,21}$

≈ 0.94.

(3) 设这种零件的总耗材量的估计值为 y mm^3.

又已知这种零件的耗材量和其横截面面积近似成正比，

可得 $\dfrac{0.052}{0.39}=\dfrac{182}{y}$，解得 $y=1\,365$ mm^3，

故刘铭制作的零件的总耗材量的估计值为 1 365 mm^3.

6. 【解】(1) 由题可知 $\bar{x}=\dfrac{1}{5}(12+12.5+13+13.5+14)=13$，$\bar{y}=\dfrac{1}{5}(14+13+11+9+8)=11$.

(2) 因为 $\sum\limits_{i=1}^{5}(x_i-\bar{x})^2=(12-13)^2+(12.5-13)^2+(13-13)^2+(13.5-13)^2+(14-13)^2=2.5$，

$\sum\limits_{i=1}^{5}(y_i-\bar{y})^2=(14-11)^2+(13-11)^2+(11-11)^2+(9-11)^2+(8-11)^2=26$，

所以 $r=\dfrac{\sum\limits_{i=1}^{5}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum\limits_{i=1}^{5}(x_i-\bar{x})^2}\sqrt{\sum\limits_{i=1}^{5}(y_i-\bar{y})^2}}=-\dfrac{8}{\sqrt{65}}\approx-0.992$.

因为 $|r|\approx 0.992>0.75$，所以可以推断 y 与 x 的线性相关性很强.

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

1. A 【解析】根据题意，适合用线性回

归模型拟合其中两个变量的散点图中,点的分布必须比较集中,且大体接近某一条直线,分析选项可得 A 选项的散点图杂乱无章,最不符合条件. 故选 A.

2. D 【解析】由题意可得 $\overline{x}=\frac{1}{5}\times(6+7+10+12+15)=10$.

∵ 经验回归方程为 $\hat{y}=0.7x-6$,

∴ $\overline{y}=0.7\times10-6=1$.

∵ y_1,y_2,y_3,y_4,y_5 成等差数列,

∴ $\overline{y}=\frac{1}{5}(y_1+y_2+y_3+y_4+y_5)=y_3=1$.

故选 D.

3. ABC 【解析】对于 A,因为 $0.85>0$,所以 y 与 x 是正相关的,所以 A 正确. 对于 B,经验回归直线恒过样本点的中心,所以经验回归直线 $\hat{y}=0.85x-85.71$ 必过点 $(\overline{x},\overline{y})$,所以 B 正确. 对于 C,由于经验回归方程为 $\hat{y}=0.85x-85.71$,所以可知该中学某高中女生身高增加 1 cm,则其体重约增加 0.85 kg,所以 C 正确. 对于 D,当 $x=160$ 时,$\hat{y}=0.85\times160-85.71=50.29$,所以该中学某高中女生身高为 160 cm 时,其体重约为 50.29 kg,所以 D 错误. 故选 ABC.

4. 【解】(1)当 $x=3$ 时,$\hat{y}_1=10.7\times3+3.4=35.5$,所以 $M=43-35.5=7.5$;

当 $x=4$ 时,$\hat{y}_2=35.5\times\sqrt{4}-22.8=48.2$,所以 $n=45-48.2=-3.2$.

则模型①残差值的绝对值之和为 $1.1+2.8+7.5+1.2+1.9+0.4=14.9$,

模型②残差值的绝对值之和为 $0.3+5.4+4.3+3.2+1.6+3.8=18.6$,

因为 $14.9<18.6$,所以模型①的拟合效果较好,应该选模型①.

(2)由题意剔除异常数据即第 3 天的数据后,

得 $\overline{x}=\frac{1}{5}\times(3.5\times6-3)=3.6$,

$\overline{y}=\frac{1}{5}\times(41\times6-43)=40.6$,

$\sum_{i=1}^{5}x_iy_i=1\,049-3\times43=920$,

$\sum_{i=1}^{5} x_i^2 = 91 - 3^2 = 82$,

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

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

$= \dfrac{920 - 5 \times 3.6 \times 40.6}{82 - 5 \times 3.6 \times 3.6} = 11$,

$\hat{a} = \bar{y} - \hat{b}\bar{x} = 40.6 - 11 \times 3.6 = 1$,

故 y 关于 x 的经验回归方程为 $\hat{y} = 11x + 1$.

5. 【解】(1) $\bar{x} = \dfrac{2+5+8+9+11}{5} = 7$, $\bar{y} = \dfrac{12+10+8+8+7}{5} = 9$.

(2) $\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = (-5) \times 3 + (-2) \times 1 + 1 \times (-1) + 2 \times (-1) + 4 \times (-2) = -28$,

$\sqrt{\sum_{i=1}^{5}(x_i - \bar{x})^2} = \sqrt{25+4+1+4+16} = 5\sqrt{2}$,

$\sqrt{\sum_{i=1}^{5}(y_i - \bar{y})^2} = \sqrt{9+1+1+1+4} = 4$,

代入公式得, $r = \dfrac{-28}{5\sqrt{2} \times 4} \approx \dfrac{-7}{7.07} \approx -0.99$.

(3) 由于 $R^2 = 1 - \dfrac{\sum_{i=1}^{5}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{5}(y_i - \bar{y})^2} = 1 - \dfrac{0.32}{16} = 0.98$, 故响应变量 y 的差异有 98% 由解释变量 x 引起.

6. 【解】(1) 根据散点图, 开始的点在某条直线旁, 但后面的点越来越偏离这条直线, 因此 $y = c + d \cdot \ln x$ 更适合作为经验回归方程类型.

(2) 由 $w = \ln x$, 则 $y = c + d \cdot \ln x$ 可化为 $y = c + d \cdot w$,

$\bar{y} = \dfrac{\sum_{i=1}^{20} y_i}{20} = \dfrac{102.4}{20} = 5.12$, $\bar{w} = \dfrac{\sum_{i=1}^{20} w_i}{20} = \dfrac{52}{20} = 2.6$,

$\hat{d} = \dfrac{\sum_{i=1}^{20} w_i y_i - 20\bar{w}\,\bar{y}}{\sum_{i=1}^{20} w_i^2 - 20\bar{w}^2}$

$$= \frac{272.1 - 20 \times 2.6 \times 5.12}{137 - 20 \times 2.6^2}$$

≈ 3.26,

$\hat{c} = \bar{y} - \hat{d}\bar{w} \approx 5.12 - 3.26 \times 2.6 \approx -3.36$,

所以 $\hat{y} = 3.26w - 3.36$,即 $\hat{y} = 3.26\ln x - 3.36$.

(3)当 $x = e^2$ 时,$\hat{y} = 3.26\ln e^2 - 3.36 = 3.16$. 故当光照时长为 e^2 小时时,大棚蔬菜每公顷产量约为 3.16 吨.

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

1. B 【解析】当 ad 与 bc 差距越大时,两个变量有关联的可能性就越大. 检验四个选项中所给的 ad 与 bc 的差距.

A 中,$ad-bc = 10-12 = -2$;

B 中,$ad-bc = 20-9 = 11$;

C 中,$ad-bc = 15-12 = 3$;

D 中,$ad-bc = 15-12 = 3$.

显然 B 中 $|ad-bc|$ 最大. 故选 B.

2. B 【解析】根据题意,在等高堆积条形图中,当 $x_1(x_2)$ 在 y_1, y_2 中所占比例相差越大时,越有把握认为两个分类变量 x, y 之间有关联,

由选项可得,B 选项中,$x_1(x_2)$ 在 y_1, y_2 中所占比例相差无几,所以最有把握认为两个分类变量 x, y 之间没有关联. 故选 B.

3. D 【解析】因为 $\chi^2 = 7.233 > 6.635$,

所以在犯错误的概率不超过 0.01 的前提下,认为“该校高中生爱好数学与性别有关”. 故选 D.

4. 12 【解析】由题意得到如下列联表:

	喜欢看篮球赛	不喜欢看篮球赛	合计
男生	$\frac{5n}{6}$	$\frac{n}{6}$	n
女生	$\frac{n}{6}$	$\frac{n}{3}$	$\frac{n}{2}$
合计	n	$\frac{n}{2}$	$\frac{3n}{2}$

所以$\chi^2=\dfrac{\dfrac{3n}{2}\left(\dfrac{5n}{6}\cdot\dfrac{n}{3}-\dfrac{n}{6}\cdot\dfrac{n}{6}\right)^2}{n\cdot\dfrac{n}{2}\cdot\dfrac{n}{2}\cdot n}=\dfrac{3n}{8}$.

因为有95%的把握认为喜欢看篮球赛与性别有关，

所以$\chi^2\geqslant 3.841$，即$\dfrac{3n}{8}\geqslant 3.841$，$n\geqslant\dfrac{3.841\times 8}{3}\approx 10.24$.

又$\dfrac{n}{2}$，$\dfrac{n}{3}$，$\dfrac{n}{6}$为整数，所以n的最小值为12.

5. 【解】(1)甲学校竞赛成绩优秀的频率为$\dfrac{60}{100}=\dfrac{3}{5}$，

乙学校竞赛成绩优秀的频率为$\dfrac{70}{100}=\dfrac{7}{10}$.

(2)由列联表可得$\chi^2=\dfrac{200\times(60\times 30-40\times 70)^2}{100\times 100\times 130\times 70}=\dfrac{200}{91}\approx 2.198<3.841$，

故没有95%的把握认为甲校成绩优秀与乙校成绩优秀有差异.

6. 【解】(1)列联表补充完整如下：

性别	参与意愿		合计
	愿意参与	不愿意参与	
男性	48	12	60
女性	22	18	40
合计	70	30	100

零假设为H_0：参与意愿与性别无关联，根据列联表的数据可得，$\chi^2=\dfrac{100\times(48\times 18-22\times 12)^2}{60\times 40\times 70\times 30}=\dfrac{50}{7}\approx 7.143>6.635=x_{0.01}$，

对照附表，依据小概率值$\alpha=0.01$的独立性检验，我们推断H_0不成立，所以认为参与意愿与性别有关联，此推断犯错误的概率不大于0.01.

根据表中数据计算男性和女性愿意参与活动的频率分别为$\dfrac{48}{60}=\dfrac{4}{5}$，$\dfrac{22}{40}=\dfrac{11}{20}$，可得$\dfrac{\frac{4}{5}}{\frac{11}{20}}=\dfrac{16}{11}\approx 1.45$，可见，在被调查者

中,男性愿意参与活动的频率是女性愿意参与活动频率的 1.4 倍以上,于是,根据频率稳定于概率的原理,我们可以认为男性比女性更愿意参与活动.

(2)X 的可能取值为 0,1,2,3,

$$P(X=0)=\frac{C_4^0C_3^3}{C_7^3}=\frac{1}{35},$$

$$P(X=1)=\frac{C_4^1C_3^2}{C_7^3}=\frac{12}{35},$$

$$P(X=2)=\frac{C_4^2C_3^1}{C_7^3}=\frac{18}{35},$$

$$P(X=3)=\frac{C_4^3C_3^0}{C_7^3}=\frac{4}{35}.$$

所以 X 的分布列为

X	0	1	2	3
P	$\frac{1}{35}$	$\frac{12}{35}$	$\frac{18}{35}$	$\frac{4}{35}$

根据超几何分布的均值得 $E(X)=\frac{4}{7}\times 3=\frac{12}{7}$.