

成对数据的统计分析

【考试要求】 1. 了解样本相关系数的统计含义. 2. 理解一元线性回归模型和 2×2 列联表, 会运用这些方法解决简洁的实际问题. 3. 会利用统计软件进行数据分析.

落实 主干知识

【学问梳理】

1. 变量的相关关系

(1) 相关关系 两个变量有关系, 但又没有准确到可由其中的一个去精确地确定另一个的程度, 这种关系称为相关关系.

(2) 相关关系的分类: 正相关和负相关.

(3) 线性相关 一般地, 假如两个变量的取值呈现正相关或负相关, 而且散点落在一条直线旁边, 我们称这两个变量线性相关.

2. 样本相关系数

$$(1) r = \frac{\sum_{i=1}^n (X^i - \bar{X})(Y^i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X^i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y^i - \bar{Y})^2}}.$$

(2) 当 $r > 0$ 时, 称成对样本数据正相关; 当 $r < 0$ 时, 称成对样本数据负相关.

(3) $|r| \leq 1$; 当 $|r|$ 越接近 1 时, 成对样本数据的线性相关程度越强; 当 $|r|$ 越接近 0 时, 成对样本数据的线性相关程度越弱.

3. 一元线性回归模型

(1) 我们将 $y = bx + a$ 称为 Y 关于 x 的阅历回来方程,

$$\text{其中} \begin{cases} b = \frac{\sum_{i=1}^n (X^i - \bar{X})(Y^i - \bar{Y})}{\sum_{i=1}^n (X^i - \bar{X})^2}, \\ a = \bar{Y} - b\bar{X}. \end{cases}$$

(2) 残差: 观测值减去预料值, 称为残差.

4. 列联表与独立性检验

(1) 关于分类变量 X 和 Y 的抽样数据的 2×2 列联表:

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

(2) 计算随机变量 $\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$, 利用 χ^2 的取值推断分类变量 X 和 Y 是否独立的方法称为 χ^2 独立性检验.

α	0.10	0.05	0.010	0.005	0.001
χ^2_{α}	2.706	3.841	6.635	7.879	10.828

【常用结论】

1. 阅历回来直线过点 $(\bar{x}, \bar{y})$.

2. 求 b 时, 常用公式 $b = \frac{\sum_{i=1}^n x^i y^i - n\bar{x}\bar{y}}{\sum_{i=1}^n x^i - n\bar{x}}$.

3. 回来分析和独立性检验都是基于成对样本观测数据进行估计或推断, 得出的结论都可能犯错误.

【思索辨析】

推断下列结论是否正确(请在括号中打“√”或“×”)

- (1) 散点图是推断两个变量相关关系的一种重要方法和手段. (√)
- (2) 阅历回来直线 $y = bx + a$ 至少经过点 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 中的一个点. (×)
- (3) 样本相关系数的肯定值越接近 1, 成对样本数据的线性相关程度越强. (√)
- (4) 若分类变量 X, Y 关系越亲密, 则由观测数据计算得到的 χ^2 的观测值越小. (×)

【教材改编题】

1. 试验测得四组 (x, y) 的值为 $(1, 2), (2, 3), (3, 4), (4, 5)$, 则 y 与 x 之间的阅历回来方程为()

- A. $y = x + 1$
- B. $y = x + 2$
- C. $y = 2x + 1$
- D. $y = x - 1$

答案 A

解析 由已知可得 $\bar{y} = \frac{1}{4} \times (1+2+3+4) = 2.5$,

$$\bar{y} = \frac{1}{4} \times (2+3+4+5) = 3.5,$$

所以阅历回来直线过点 $(2.5, 3.5)$,

所以把点 $(2.5, 3.5)$ 代入四个选项中验证, 可得只有 $y = x + 1$ 成立.

2. (多选) 下列关于成对样本数据的统计分析的推断中正确的有 ()

- A. 若样本相关系数 $r=0$, 则说明成对样本数据没有相关性
- B. 样本相关系数 r 越大, 成对样本数据的线性相关性越强
- C. 用最小二乘法求得的一元线性回来模型的残差和肯定是 0
- D. 确定系数 R^2 越大, 残差平方和越小, 模型的拟合效果越好

答案 CD

解析 对于 A, 当 $r=0$ 时, 只表明成对样本数据间没有线性相关关系, 但是不解除它们之间有其他相关关系, 故 A 错误;

对于 B, 样本相关系数 $|r|$ 越大, 成对样本数据的线性相关性越强, 故 B 错误;

对于 C, 残差和为

$$\sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n [y_i - (bx_i + a)]$$

$$= \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i - \sum_{i=1}^n a$$

$$= n\bar{y} - nb\bar{x} - na$$

$$= n(\bar{y} - b\bar{x} - a) = 0, \text{ 故 C 正确;}$$

对于 D, 确定系数 R^2 越大, 表示残差平方和越小, 即模型的拟合效果越好, 故 D 正确.

3. (多选) 为了解阅读量多少与华蜜感强弱之间的关系, 一个调查机构依据所得到的数据, 绘制了如下所示的 2×2 列联表(个别数据暂用字母表示):

	华蜜感强	华蜜感弱	合计
阅读量多	m	18	72
阅读量少	36	n	78
合计	90	60	150

计算得: $\chi^2 \approx 12.981$, 参照下表:

α	0.10	0.05	0.025	0.010	0.005	0.001
----------	------	------	-------	-------	-------	-------

x_{α}	2.706	3.841	5.024	6.635	7.879	10.828
--------------	-------	-------	-------	-------	-------	--------

对于下面的选项，正确的为()

- A. 依据小概率值 $\alpha=0.010$ 的独立性检验，可以认为“阅读量多少与华蜜感强弱无关”
- B. $m=54$
- C. 依据小概率值 $\alpha=0.005$ 的独立性检验，可以在犯错误的概率不超过 0.5%的前提下认为“阅读量多少与华蜜感强弱有关”
- D. $n=52$

答案 BC

解析 $\because \chi^2 \approx 12.981 > 7.879 > 6.635$,

$\therefore$ 依据小概率值 $\alpha=0.010$ 的独立性检验，可以在犯错误的概率不超过 1%的前提下认为“阅读量多少与华蜜感强弱有关”，

依据小概率值 $\alpha=0.005$ 的独立性检验，可以在犯错误的概率不超过 0.5%的前提下认为“阅读量多少与华蜜感强弱有关”，

$\therefore$ A 错，C 对，

$\because m+36=90, 18+n=60$,

$\therefore m=54, n=42$,

$\therefore$ B 对，D 错.

■ 探究 核心题型

题型一 成对数据的相关性

例 1 (1) 对变量 x, y 有观测数据 $(x_i, y_i) (i=1, 2, \dots, 10)$ ，得散点图如图 1，对变量 u, v 有观测数据 $(u_i, v_i) (i=1, 2, \dots, 10)$ ，得散点图如图 2. 由这两个散点图可以推断()

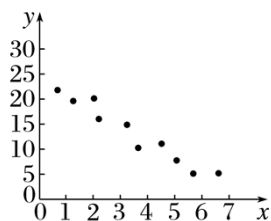


图 1

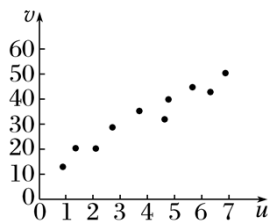


图 2

- A. 变量 x 与 y 正相关， u 与 v 正相关
- B. 变量 x 与 y 正相关， u 与 v 负相关
- C. 变量 x 与 y 负相关， u 与 v 正相关
- D. 变量 x 与 y 负相关， u 与 v 负相关

答案 C

解析 由题图可得两组数据均线性相关，且图 1 的阅历回来直线的斜率为负，图 2 的阅历

回来直线的斜率为正, 则由散点图可推断变量 x 与 y 负相关, u 与 v 正相关.

(2) (多选) 下列有关阅历回来分析的说法中正确的有()

- A. 阅历回来直线必过点 $(\bar{x}, \bar{y})$
- B. 阅历回来直线就是散点图中经过样本数据点最多的那条直线
- C. 当样本相关系数 $r > 0$ 时, 两个变量正相关
- D. 假如两个变量的相关性越弱, 则 $|r|$ 就越接近于 0

答案 ACD

解析 对于 A, 阅历回来直线必过点 $(\bar{x}, \bar{y})$, 故 A 正确;

对于 B, 阅历回来直线在散点图中可能不经过任一样本数据点, 故 B 不正确;

对于 C, 当样本相关系数 $r > 0$ 时, 则两个变量正相关, 故 C 正确;

对于 D, 假如两个变量的相关性越弱, 则 $|r|$ 就越接近于 0, 故 D 正确.

【老师备选】

1. 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, 若全部样本点 (x_i, y_i) ($i=1, 2, \dots, n$) 都在直线 $y = \frac{1}{2}x + 1$ 上, 则这组样本数据的样本相关系数为()

- A. -1
- B. 0
- C. $\frac{1}{2}$
- D. 1

答案 D

解析 全部样本点均在同一条斜率为正数的直线上, 则样本相关系数最大, 为 1.

2. (多选) 下列选项中正确的是()

- A. 阅历回来分析中, R^2 的值越大, 说明残差平方和越小
- B. 若一组观测数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 满意 $y_i = bx_i + a + e_i$ ($i=1, 2, \dots, n$), 若 e_i 恒为 0, 则 $R^2 = 1$
- C. 阅历回来分析是对具有相关关系的两个变量进行统计分析的一种常用方法
- D. 画残差图时, 纵坐标为残差, 横坐标肯定是编号

答案 ABC

解析 对于 A, 阅历回来分析中, R^2 的值越大, 说明模型的拟合效果越好, 则残差平方和越小, A 对;

对于 B, 若一组观测数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 满意 $y_i = bx_i + a + e_i$ ($i=1, 2, \dots, n$), 若 e_i 恒为 0, 则 $R^2 = 1$, B 对;

对于 C, 阅历回来分析是对具有相关关系的两个变量进行统计分析的一种常用方法, C 对;

对于 D, 残差图中横坐标可以是样本编号, 也可以是身高数据, 还可以是体重的估计值等, D 错.

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。

如要下载或阅读全文，请访问：

<https://d.book118.com/548051042024006110>