

§11.3 变量间的相关关系、统计案例

最新考纲	考情考向分析
1.会作两个相关变量的数据的散点图，会利用散点图认识变量间的相关关系. 2.了解最小二乘法的思想，能根据给出的线性回归方程系数公式建立线性回归方程. 3.了解独立性检验的基本思想、方法及其初步应用. 4.了解回归分析的基本思想、方法及简单应用.	回归分析，独立性检验是全国卷高考重点考查的内容，必考一个解答题，选择、填空题中也会出现. 主要考查回归方程，相关系数，利用回归方程进行预测，独立性检验的应用等.

基础知识 自主学习

回扣基础知识 训练基础题目

知识梳理

1. 两个变量的线性相关

(1)正相关

在散点图中，点散布在从左下角到右上角的区域，对于两个变量的这种相关关系，我们将其称为正相关.

(2)负相关

在散点图中，点散布在从左上角到右下角的区域，两个变量的这种相关关系称为负相关.

(3)线性相关关系、回归直线

如果散点图中点的分布从整体上看大致在一条直线附近，就称这两个变量之间具有线性相关关系，这条直线叫做回归直线.

2. 回归方程

(1)最小二乘法

求回归直线，使得样本数据的点到它的距离的平方和最小的方法叫做最小二乘法.

(2)回归方程

方程 $\hat{y} = \hat{b}x + \hat{a}$ 是两个具有线性相关关系的变量的一组数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

的回归方程，其中 $\hat{a}$ ， $\hat{b}$ 是待定参数.

$$\begin{cases} \hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, \\ \hat{a} = \bar{y} - \hat{b}\bar{x}. \end{cases}$$

3. 回归分析

(1)定义：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.

(2)样本点的中心

对于一组具有线性相关关系的数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ ，其中 $(\bar{x}, \bar{y})$ 称为样本点的中心.

(3)相关系数

当 $r > 0$ 时，表明两个变量正相关；

当 $r < 0$ 时，表明两个变量负相关.

r 的绝对值越接近于 1，表明两个变量的线性相关性越强. r 的绝对值越接近于 0，表明两个变量之间几乎不存在线性相关关系. 通常 $|r|$ 大于 0.75 时，认为两个变量有很强的线性相关性.

4. 独立性检验

(1)分类变量：变量的不同“值”表示个体所属的不同类别，像这样的变量称为分类变量.

(2)列联表：列出的两个分类变量的频数表，称为列联表. 假设有两个分类变量 X 和 Y ，它们的可能取值分别为 $\{x_1, x_2\}$ 和 $\{y_1, y_2\}$ ，其样本频数列联表(称为 2×2 列联表)为

2×2 列联表

	y_1	y_2	总计
x_1	a	b	$a+b$
x_2	c	d	$c+d$
总计	$a+c$	$b+d$	$a+b+c+d$

构造一个随机变量 $K^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ ，其中 $n = \underline{a+b+c+d}$ 为样本容量.

(3)独立性检验

利用随机变量 K^2 来判断“两个分类变量有关系”的方法称为独立性检验.

基础自测

题组一 思考辨析

1. 判断下列结论是否正确(请在括号中打“√”或“×”)

(1)相关关系与函数关系都是一种确定性的关系，也是一种因果关系. (×)

解析 由 $\bar{x}=30$, 得 $\bar{y}=0.67 \times 30+54.9=75$.

设表中的“模糊数字”为 a ,

则 $62+a+75+81+89=75 \times 5$, $\therefore a=68$.

题组三 易错自纠

5. 某医疗机构通过抽样调查(样本容量 $n=1\,000$), 利用 2×2 列联表和 K^2 统计量研究患肺病是否与吸烟有关. 计算得 $K^2=4.453$, 经查阅临界值表知 $P(K^2 \geq 3.841) \approx 0.05$, 现给出四个结论, 其中正确的是()

- A. 在100个吸烟的人中约有95个人患肺病
- B. 若某人吸烟, 那么他有95%的可能性患肺病
- C. 有95%的把握认为“患肺病与吸烟有关”
- D. 只有5%的把握认为“患肺病与吸烟有关”

答案 C

解析 由已知数据可得, 有 $1-0.05=95\%$ 的把握认为“患肺病与吸烟有关”.

6. 在一次考试中, 5名学生的数学和物理成绩如下表: (已知学生的数学和物理成绩具有线性相关关系)

学生的编号 <i>i</i>	1	2	3	4	5
数学成绩 <i>x</i>	80	75	70	65	60
物理成绩 <i>y</i>	70	66	68	64	62

现已知其线性回归方程为 $\hat{y}=0.36x+\hat{a}$, 则根据此线性回归方程估计数学得90分的同学的物理成绩为_____. (四舍五入到整数)

答案 73

解析 $\bar{x}=\frac{60+65+70+75+80}{5}=70$,

$\bar{y}=\frac{62+64+66+68+70}{5}=66$,

所以 $66=0.36 \times 70+\hat{a}$, $\hat{a}=40.8$,

即线性回归方程为 $\hat{y}=0.36x+40.8$.

当 $x=90$ 时, $\hat{y}=0.36 \times 90+40.8=73.2 \approx 73$.

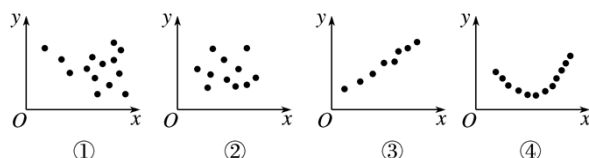
题型分类 深度剖析

真题典题深度剖析 重点难点多维探究

题型一 相关关系的判断

自主演练

1. 观察下列各图形,



其中两个变量 x, y 具有相关关系的图是()

A. ①②

B. ①④

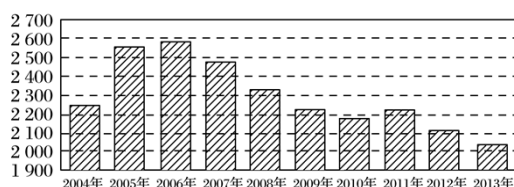
C. ③④

D. ②③

答案 C

解析 由散点图知③中的点都分布在一条直线附近, ④中的点都分布在一条曲线附近, 所以③④中的两个变量具有相关关系.

2. (2018·广州质检)根据下面给出的 2004 年至 2013 年我国二氧化硫排放量(单位: 万吨)的柱形图. 以下结论不正确的是()



A. 逐年比较, 2008 年减少二氧化硫排放量的效果最显著

B. 2007 年我国治理二氧化硫排放显现成效

C. 2006 年以来我国二氧化硫年排放量呈减少趋势

D. 2006 年以来我国二氧化硫年排放量与年份正相关

答案 D

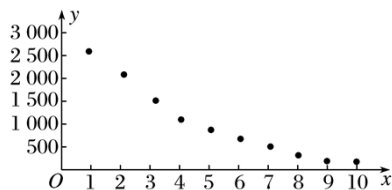
解析 从 2006 年, 将每年的二氧化硫排放量与前一年作差比较, 得到 2008 年二氧化硫排放量与 2007 年排放量的差最大, A 选项正确;

2007 年二氧化硫排放量较 2006 年降低了很多, B 选项正确;

虽然 2011 年二氧化硫排放量较 2010 年多一些, 但自 2006 年以来, 整体呈递减趋势, C 选项正确;

自 2006 年以来我国二氧化硫年排放量与年份负相关, D 选项错误, 故选 D.

3. x 和 y 的散点图如图所示, 则下列说法中所有正确命题的序号为_____.



① x, y 是负相关关系;

②在该相关关系中, 若用 $y = c_1 e^{c_2 x}$ 拟合时的相关指数为 R_1^2 , 用 $\hat{y} = \hat{b}x + \hat{a}$ 拟合时的相关指数为 R_2^2 , 则 $R_1^2 > R_2^2$;

③ x, y 之间不能建立线性回归方程.

答案 ①②

解析 在散点图中, 点散布在从左上角到右下角的区域, 因此 x, y 是负相关关系, 故①正确. 由散点图知用 $y = c_1 e^{c_2 x}$ 拟合比用 $\hat{y} = \hat{b}x + \hat{a}$ 拟合效果要好, 则 $R_1^2 > R_2^2$, 故②正确; x, y 之间可以建立线性回归方程, 但拟合效果不好, 故③错误.

思维升华 判定两个变量正, 负相关性的方法

(1)画散点图: 点的分布从左下角到右上角, 两个变量正相关; 点的分布从左上角到右下角, 两个变量负相关.

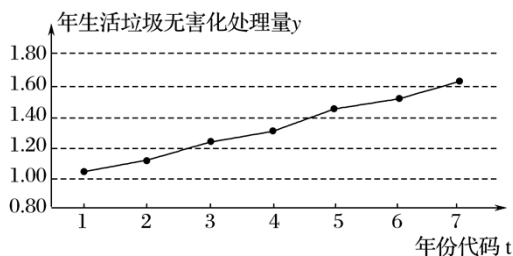
(2)相关系数: $r > 0$ 时, 正相关; $r < 0$ 时, 负相关.

(3)线性回归方程中: $\hat{b} > 0$ 时, 正相关; $\hat{b} < 0$ 时, 负相关.

题型二 线性回归分析

师生共研

典例 (2016·全国III)下图是我国 2008 年至 2014 年生活垃圾无害化处理量(单位: 亿吨)的折线图.



注: 年份代码 1~7 分别对应年份 2008~2014.

(1)由折线图看出, 可用线性回归模型拟合 y 与 t 的关系, 请用相关系数加以说明;

(2)建立 y 关于 t 的回归方程(系数精确到 0.01), 预测 2016 年我国生活垃圾无害化处理量.

附注:

参考数据: $\sum_{i=1}^7 y_i = 9.32$, $\sum_{i=1}^7 t y_i = 40.17$, $\sqrt{\sum_{i=1}^7 (y_i - \bar{y})^2} = 0.55$, $\sqrt{7} \approx 2.646$.

参考公式: 相关系数 $r = \frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (t_i - \bar{t})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$,

回归方程 $\hat{y} = \hat{a} + \hat{b}t$ 中斜率和截距的最小二乘估计公式分别为:

$$\hat{b} = \frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^n (t_i - \bar{t})^2}, \quad \hat{a} = \bar{y} - \hat{b} \bar{t}.$$

解 (1) 由折线图中数据和附注中参考数据得

$$\bar{t} = 4, \quad \sum_{i=1}^7 (t_i - \bar{t})^2 = 28, \quad \sqrt{\sum_{i=1}^7 (y_i - \bar{y})^2} = 0.55.$$

$$\begin{aligned} \sum_{i=1}^7 (t_i - \bar{t})(y_i - \bar{y}) &= \sum_{i=1}^7 t y_i - \bar{t} \sum_{i=1}^7 y_i \\ &= 40.17 - 4 \times 9.32 = 2.89, \end{aligned}$$

$$\text{所以 } r \approx \frac{2.89}{0.55 \times 2 \times 2.646} \approx 0.99.$$

因为 y 与 t 的相关系数近似为 0.99, 说明 y 与 t 的线性相关程度相当高, 从而可以用线性回归模型拟合 y 与 t 的关系.

(2) 由 $\bar{y} = \frac{9.32}{7} \approx 1.331$ 及 (1) 得

$$\hat{b} = \frac{\sum_{i=1}^7 (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^7 (t_i - \bar{t})^2} = \frac{2.89}{28} \approx 0.103,$$

$$\hat{a} = \bar{y} - \hat{b} \bar{t} \approx 1.331 - 0.103 \times 4 \approx 0.92.$$

所以 y 关于 t 的回归方程为 $\hat{y} = 0.92 + 0.10t$.

将 2016 年对应的 $t=9$ 代入回归方程得

$$\hat{y} = 0.92 + 0.10 \times 9 = 1.82.$$

所以预测 2016 年我国生活垃圾无害化处理量将约为 1.82 亿吨.

思维升华 线性回归分析问题的类型及解题方法

(1)求线性回归方程

①利用公式，求出回归系数 $\hat{b}$ ， $\hat{a}$.

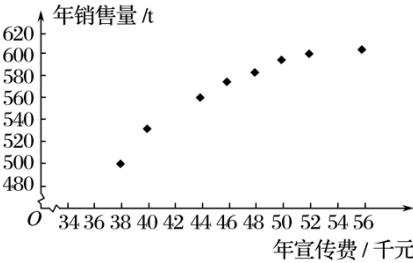
②待定系数法：利用回归直线过样本点的中心求系数.

(2)利用回归方程进行预测，把线性回归方程看作一次函数，求函数值.

(3)利用回归直线判断正、负相关；决定正相关还是负相关的是系数 $\hat{b}$.

(4)回归方程的拟合效果，可以利用相关系数判断，当 $|r|$ 越趋近于1时，两变量的线性相关性越强.

跟踪训练 某公司为确定下一年度投入某种产品的宣传费，需了解年宣传费 x (单位：千元)对年销售量 y (单位：t)和年利润 z (单位：千元)的影响，对近8年的年宣传费 x_i 和年销售量 y_i ($i=1,2,\cdots,8$)数据作了初步处理，得到下面的散点图及一些统计量的值.



$\bar{x}$	$\bar{y}$	$\bar{w}$	$\sum_{i=1}^8 (x_i - \bar{x})^2$	$\sum_{i=1}^8 (w_i - \bar{w})^2$	$\sum_{i=1}^8 (x_i - \bar{x}) \cdot (y_i - \bar{y})$	$\sum_{i=1}^8 (w_i - \bar{w}) \cdot (y_i - \bar{y})$
46.6	563	6.8	289.8	1.6	1 469	108.8

表中 $w_i = \sqrt{x_i}$, $\bar{w} = \frac{1}{8} \sum_{i=1}^8 w_i$.

(1)根据散点图判断， $y=a+bx$ 与 $y=c+d\sqrt{x}$ 哪一个适宜作为年销售量 y 关于年宣传费 x 的回归方程类型？(给出判断即可，不必说明理由)

(2)根据(1)的判断结果及表中数据，建立 y 关于 x 的回归方程；

(3)已知这种产品的年利润 z 与 x ， y 的关系为 $z=0.2y-x$.根据(2)的结果回答下列问题：

①年宣传费 $x=49$ 时，年销售量及年利润的预报值是多少？

②年宣传费 x 为何值时，年利润的预报值最大？

附：对于一组数据 $(u_1, v_1), (u_2, v_2), \cdots, (u_n, v_n)$ ，其回归直线 $\hat{v} = \hat{\alpha} + \hat{\beta} u$ 的斜率和截距的最小二乘估计分别为

$$\hat{\beta} = \frac{\sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})}{\sum_{i=1}^n (u_i - \bar{u})^2}, \quad \hat{\alpha} = \bar{v} - \hat{\beta} \bar{u}.$$

解 (1)由散点图可以判断, $y=c+d\sqrt{x}$ 适宜作为年销售量 y 关于年宣传费 x 的回归方程类型.

(2)令 $w=\sqrt{x}$, 先建立 y 关于 w 的线性回归方程, 由于

$$\hat{d} = \frac{\sum_{i=1}^8 (w_i - \bar{w}) \cdot (y_i - \bar{y})}{\sum_{i=1}^8 (w_i - \bar{w})^2} = \frac{108.8}{1.6} = 68,$$

$$\hat{c} = \bar{y} - \hat{d} \bar{w} = 563 - 68 \times 6.8 = 100.6,$$

所以 y 关于 w 的线性回归方程为 $\hat{y} = 100.6 + 68w$,

因此 y 关于 x 的回归方程为 $\hat{y} = 100.6 + 68\sqrt{x}$.

(3)①由(2)知, 当 $x=49$ 时,

年销售量 y 的预报值 $\hat{y} = 100.6 + 68\sqrt{49} = 576.6$,

年利润 z 的预报值 $\hat{z} = 576.6 \times 0.2 - 49 = 66.32$.

②根据(2)的结果知, 年利润 z 的预报值

$$\hat{z} = 0.2(100.6 + 68\sqrt{x}) - x = -x + 13.6\sqrt{x} + 20.12.$$

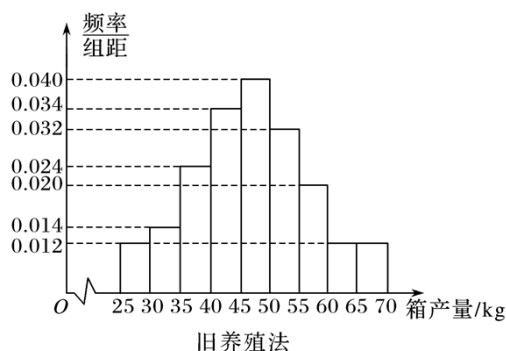
所以当 $\sqrt{x} = \frac{13.6}{2} = 6.8$, 即 $x=46.24$ 时, $\hat{z}$ 取得最大值.

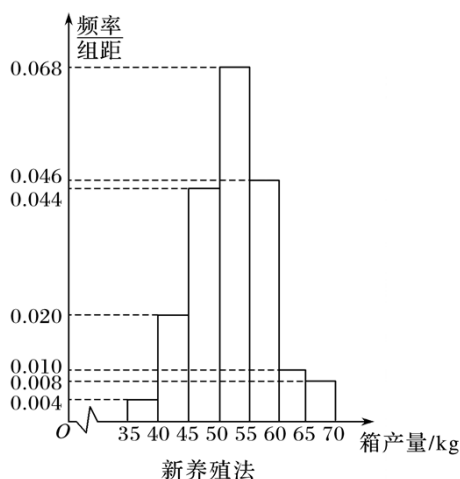
故年宣传费为 46.24 千元时, 年利润的预报值最大.

题型三 独立性检验

师生共研

典例 (2017·全国 II)海水养殖场进行某水产品的新、旧网箱养殖方法的产量对比, 收获时各随机抽取了 100 个网箱, 测量各箱水产品的产量(单位: kg), 其频率分布直方图如下:





(1) 设两种养殖方法的箱产量相互独立，记 A 表示事件“旧养殖法的箱产量低于 50 kg，新养殖法的箱产量不低于 50 kg”，估计 A 的概率；

(2) 填写下面列联表，并根据列联表判断是否有 99% 的把握认为箱产量与养殖方法有关：

	箱产量 < 50 kg	箱产量 $\geq$ 50 kg
旧养殖法		
新养殖法		

(3) 根据箱产量的频率分布直方图，求新养殖法箱产量的中位数的估计值(精确到 0.01).

附：

$P(K^2 \geq k_0)$	0.050	0.010	0.001
k_0	3.841	6.635	10.828

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

解 (1) 记 B 表示事件“旧养殖法的箱产量低于 50 kg”， C 表示事件“新养殖法的箱产量不低于 50 kg”.

由题意知， $P(A) = P(BC) = P(B)P(C)$.

旧养殖法的箱产量低于 50 kg 的频率为

$$(0.012 + 0.014 + 0.024 + 0.034 + 0.040) \times 5 = 0.62,$$

故 $P(B)$ 的估计值为 0.62.

新养殖法的箱产量不低于 50 kg 的频率为

$$(0.068 + 0.046 + 0.010 + 0.008) \times 5 = 0.66,$$

故 $P(C)$ 的估计值为 0.66.

因此，事件 A 的概率估计值为 $0.62 \times 0.66 = 0.4092$.

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/378075005060007042>