知识点107 成对数据的统计相关性（P3-10）

知识点108 一元线性回归模型及其应用（P11-18）

知识点109 列联表与独立性检验（P19-25）

01

知识点107 成对数据的统计相关性

教材知识萃取

变量的相关关系	相关关系分为正相关和负相关,是一种非确定关系.一般地,如果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附近,我们就称这两个变量线性相关.
样本的相关系数	$r=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}=\frac{\sum_{i=1}^{n}x_iy_i-n\overline{x}\cdot\overline{y}}{\sqrt{(\sum_{i=1}^{n}x_i^2-n\overline{x}^2)(\sum_{i=1}^{n}y_i^2-n\overline{y}^2)}}$($\lvert r\rvert\leq 1$).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.当$\lvert r\rvert$越接近1时,成对样本数据的线性相关程度越强;当$\lvert r\rvert$越接近0时,成对样本数据的线性相关程度越弱.

教材素材变式

1. [链接人A选必三P98—P100知识] 对于样本相关系数r，下列说法正确的是(**D**)

A.样本相关系数$r \in (-1,1)$

B.样本相关系数r越小，成对样本数据的线性相关程度越弱

C.当$r = 0$时，成对样本数据没有任何相关关系

D.当$r = 1$时，成对样本数据正相关且两个变量之间满足一种线性关系

【解析】对于A，样本相关系数$r \in [-1,1]$，故A错误；对于B，样本相关系数$|r|$越小，成对样本数据的线性相关程度越弱，故B错误；对于C，当$r = 0$时，成对样本数据没有线性相关关系，但可能有其他相关关系，故C错误；易知D正确．故选D．

变式探究

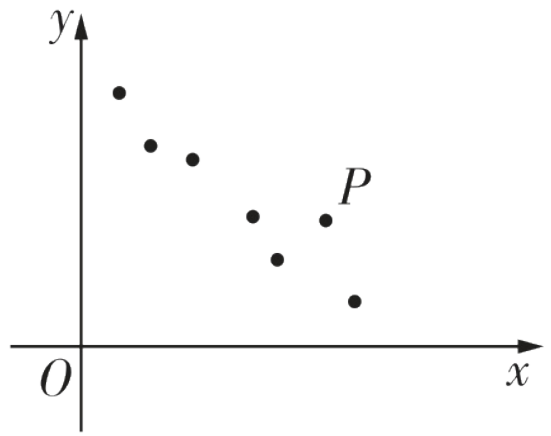
在一组样本数据(x_1, y_1)，(x_2, y_2)，$\cdots$，(x_n, y_n)（$n \geq 2$,x_1,x_2,$\cdots$,x_n不全相等）所对应的散点图中，若所有样本点$(x_i, y_i)(i = 1,2,3,\cdots,n)$都在直线$2x + y - 1 = 0$上，则这组样本数据的相关系数$r$为____$-1$.

【解析】 $\because$ 直线$2x + y - 1 = 0$的斜率$k = -2 < 0$，$\therefore$ 这两个变量负相关，$\therefore r < 0$，又所有样本点都在直线$2x + y - 1 = 0$上，$\therefore r = -1$.

2.[多选] [链接人A选必三P98—P100知识] 在如图所示的散点图中，若去掉点P，则下列说法错误的是(　**ACD**　)

A.样本相关系数r变大　　B.变量y与变量x的相关程度变强

C.变量y与变量x呈现正线性相关关系　　D.变量y与变量x的相关程度变弱



【解析】由散点图知，变量y与变量x呈现负线性相关关系，即$r<0$，故C错误.去掉点P后，变量y与变量x的线性相关程度变强，$|r|$更进一步接近1，所以r变小，故A错误,B正确，D错误．故选ACD．

3. [人A选必三P101例1变式] 某种机械设备随着使用年限的增加，它的使用功能逐渐减退，使用价值逐年减少，通常把它的使用价值逐年减少的“量”换算成费用，称为失效费.该种机械设备的使用年限x（单位：年）与失效费y（单位：万元）的统计数据如表所示.

使用年限x	2	4	5	6	8
失效费y	3	4	5	6	7

根据上表数据，计算y与x的相关系数r，并说明y与x的线性相关程度的强弱.（已知若$0.75 \leq |r| \leq 1$，则认为y与x线性相关程度较强；若$0.3 \leq |r| < 0.75$，则认为y与x线性相关程度一般；若$|r| < 0.3$，则认为y与x线性相关程度较弱）

附：$r = \dfrac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \overline{y})^2}}$，$\sqrt{2} \approx 1.41$.

【解析】 解法一 由题表知，$\overline{x}=\frac{1}{5}\times(2+4+5+6+8)=5$,

$\overline{y}=\frac{1}{5}\times(3+4+5+6+7)=5$,

$\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})=(2-5)\times(3-5)+(4-5)\times(4-5)+(5-5)\times(5-5)+(6-5)\times(6-5)+(8-5)\times($

$5)=14$,

$\sum_{i=1}^{5}(x_i-\overline{x})^2=(2-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(8-5)^2=20$,

$\sum_{i=1}^{5}(y_i-\overline{y})^2=(3-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(7-5)^2=10$,

所以$r=\frac{\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{5}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{5}(y_i-\overline{y})^2}}=\frac{14}{\sqrt{20}\times\sqrt{10}}=\frac{7\sqrt{2}}{10}\approx0.99$,

因为$0.75<r<1$，所以认为y与x的线性相关程度较强.

解法二 由题表知，$\overline{x}=\frac{1}{5}\times(2+4+5+6+8)=5$，$\overline{y}=\frac{1}{5}\times(3+4+5+6+7)=5$，

$\sum_{i=1}^{5}x_iy_i=2\times3+4\times4+5\times5+6\times6+8\times7=139$，

$\sum_{i=1}^{5}x_i^2=2^2+4^2+5^2+6^2+8^2=145$，

$\sum_{i=1}^{5}y_i^2=3^2+4^2+5^2+6^2+7^2=135$，

所以$r=\frac{\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{5}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{5}(y_i-\overline{y})^2}}=\frac{\sum_{i=1}^{5}x_iy_i-5\overline{x}\ \overline{y}}{\sqrt{\sum_{i=1}^{5}x_i^2-5\overline{x}^2}\sqrt{\sum_{i=1}^{5}y_i^2-5\overline{y}^2}}=\frac{139-5\times5\times5}{\sqrt{145-5\times5^2}\times\sqrt{135-5\times5^2}}=\frac{7\sqrt{2}}{10}\approx0.99$，

因为$0.75<r<1$，所以认为y与x的线性相关程度较强.

知识点108 一元线性回归模型及其应用

教材知识萃取

对于一组具有线性相关关系的数据$(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$,其回归直线$\hat{y}=\hat{b}x+\hat{a}$的斜率和截距的最小二乘估计公式分别为$\hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}$,$\hat{a}=\bar{y}-\hat{b}\bar{x}$.(注:观测值减去预测值称为残差)

常用结论

1.经验回归直线过点$(\bar{x},\bar{y})$.

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

3.$\hat{y}=\hat{b}x+\hat{a}$,若$\hat{b}>0$,两变量呈正相关;若$\hat{b}<0$,两变量呈负相关.

教材素材变式

1.[多选][链接人A选必三P118—P119知识]下列说法错误的是(**ACD**)

A.回归直线$\hat{y}=\hat{b}x+\hat{a}$至少经过点$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$中的一个

B.若变量y和x之间的相关系数$r=-0.936\ 2$，则变量y和x之间的负线性相关程度很强

C.在回归分析中，残差平方和越大，模型的拟合效果越好

D.已知回归方程为$\hat{y}=0.5x-8$，变量$x=2$时，变量y的值一定是-7

【解析】对于A，回归直线$\hat{y}=\hat{b}x+\hat{a}$是由最小二乘法计算出来的，它不一定经过样本数据点，但一定经过点$(\overline{x},\overline{y})$，所以A错误；对于B，由相关系数的意义知，当$|r|$越接近1时，变量$y$与$x$之间的线性相关程度越强，变量$y$和$x$之间的相关系数$r=-0.936\ 2$，则变量$y$和$x$之间具有很强的负线性相关关系，所以B正确；对于C，在回归分析中，残差平方和越小，模型的拟合效果越好，所以C错误；对于D，回归方程为$\hat{y}=0.5x-8$，变量$x=2$时，变量y的预测值是-7，但实际观测值可能不是-7，所以D错误.故选ACD.

2. [多选] [苏教选必二P187本章测试第6题变式] 已知变量x，y的取值情况如表所示，从散点图分析可知y与x线性相关，如果线性回归方程为$\hat{y} = 0.95x + 2.5$，则下列说法正确的是(　ABD　)

	0	1	2	3	4
	2.3	4.3	4.4	4.8	

A.m的值为6.2

B.回归直线必过点$(2,4.4)$

C.样本点$(4,m)$处的残差为0.1

D.将此表中的数据$(2,4.4)$去掉后，样本相关系数r不变

【解析】由题意可知，$\overline{x}=\frac{1}{5}\times(0+1+2+3+4)=2,\overline{y}=\frac{1}{5}\times(2.3+4.3+4.4+4.8+m)=\frac{1}{5}\times(15.8+m)$,所以样本点中心为$\left(2,\frac{15.8+m}{5}\right)$，将$\left(2,\frac{15.8+m}{5}\right)$代入$\hat{y}=0.95x+2.5$，可得$\frac{15.8+m}{5}=0.95\times2+2.5$，解得$m=6.2$，故A正确；由$m=6.2$，得样本点中心为$(2,4.4)$，所以回归直线必过点$(2,4.4)$，故B正确；当$x=4$时，$\hat{y}=0.95\times4+2.5=6.3$，由$m=6.2$，得样本点$(4,6.2)$处的残差为$6.2-6.3=-0.1$，故C错误；因为$\overline{x}=2$，$\overline{y}=4.4$,所以对于数据$(2,4.4)$，有$2-\overline{x}=2-2=0,4.4-\overline{y}=4.4-4.4=0$，由相关系数的公式知，$r=\frac{\sum_{i=1}^{5}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{5}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{5}(y_i-\overline{y})^2}}$，所以将此表中的数据$(2,4.4)$去掉后，样本相关系数$r$不变，故D正确.故选ABD.

3. [链接人A选必三P116知识] 近年来，我国云计算市场规模持续增长.某科技公司云计算市场规模y与年份代码x的关系可以用模型$y = c10^{tx}$拟合，设$z = \lg y$，2018年至2022年的数据统计如表所示：

年份	2018年	2019年	2020年	2021年	2022年
	1	2	3 年份代码x	4	5
	7	20	云计算市场规模y 71	200	510
	0.85	1.3	1.85 z = lg y	2.3	2.7

若根据上表得到回归方程$\hat{z} = 0.5x + \lg c$，则该科技公司2025年云计算市场规模约为(**B**)

A.$10^{3.8}$　　B.$10^{4.3}$　　C.$10^{4.8}$　　D.$10^{5.3}$

【解析】由题表知，$\overline{x} = \frac{1+2+3+4+5}{5} = 3$，$\overline{z} = \frac{0.85+1.3+1.85+2.3+2.7}{5} = 1.8$，将$\overline{x} = 3, \overline{z} = 1.8$代入回归方程$\hat{z} = 0.5x + \lg c$，可得$1.8 = 0.5 \times 3 + \lg c$，即$\lg c = 1.8 - 0.5 \times 3 = 0.3$，所以$z$关于$x$的回归方程为$\hat{z} = 0.5x + 0.3$，2025年时即当$x = 8$时，$\hat{z} = 0.5 \times 8 + 0.3 = 4.3 = \lg y$，此时$y = 10^{4.3}$.故选B.