普通高中课程标准实验教材

优质课堂1+1

高中数学

选修1-2

浙江教育出版社

普通高中课程标准实验教材

优质课堂

1+1

高中数学

浙江教育出版社

图书在版编目(CIP)数据

优质课堂1+1. 高中数学. 1-2：选修 / 郑日锋主编.
—杭州：浙江教育出版社，2009.5
配人教版教材
ISBN 978-7-5338-7921-1

Ⅰ. 优… Ⅱ. 郑… Ⅲ. 数学课－高中－教学参考资料
Ⅳ. G634

中国版本图书馆 CIP 数据核字(2009)第 044451 号

优质课堂 1+1　高中数学

选修 1-2

主　　编　郑日锋
出　　版　浙江教育出版社
　　　　　(杭州市天目山路 40 号　邮编：310013)
发　　行　浙江省新华书店集团有限公司
总 策 划　邱连根
责任编辑　金馥菊
装帧设计　韩　波
责任校对　戴正泉
责任印务　吴梦菁
图文制作　杭州富春电子印务有限公司
印刷装订　杭州广育多莉印刷有限公司

开　　本　850 × 1168　1/16
印　　张　5.75
字　　数　186 000
版　　次　2009 年 5 月第 1 版
印　　次　2009 年 5 月第 1 次印刷
印　　数　0 001—5 000
标准书号　ISBN 978-7-5338-7921-1
定　　价　9.00元

联系电话：0571-85170300-80928
e-mail：zjjy@zjcb.com
网　　址：www.zjeph.com

出版前言

为了更好地贯彻新课改的精神，为广大师生提供有较强针对性及操作性的辅导材料，我社组织省内部分优秀教师及教研员，依据《浙江省普通高中新课程实验学科教学指导意见》以及各学科现行使用教科书的要求，根据一轮新课程的教学实际，在原《随堂纠错超级练》的基础上，精心编写了《优质课堂 1+1》丛书。

这是一套涵盖高中各主要学科、包括课堂教学和阶段复习的同步实战型丛书。丛书的设计以帮助学生掌握基础知识、基本理论，提高学生的解题能力为目标，各栏目的设置注重对学生学习思路的拓展和学习方法的培养，适合课堂教学和课后训练。

《优质课堂 1+1》按章节编写，每节包括“课本解读”、“典例剖析”和“同步训练”等三个板块。其中，“课本解读”板块用简练的文字，从知识和能力的角度归纳整理了教科书的主要知识点，揭示了本章的重难点，为学生指点迷津。“典例剖析”选取每节典型例题，分析思路，点拨此类习题解答的基本策略和方法。“同步训练”按课时编写，从理解巩固、发展提高和高考链接三个层面，让学生在课堂学习之后，在对所学知识进行复习巩固的基础上，适当地拓展提升，同时对高考的命题特点有一个感性的认识。

本丛书的作者均为我省各学科的骨干教师和优秀教研员。他们不仅教学经验丰富，而且在习题的编制与选择方面有着深入的研究。在编写本丛书时，他们充分根据各学科的内容特点以及新课程的教学实际，为学生们提供了科学合理的训练素材，希望学生通过本丛书的学习，能在透彻理解教科书内容的基础上，循序渐进地提高自己的学习能力，掌握良好的学习方法，在高考中立于不败之地。

浙江教育出版社

2009 年 5 月

目录

MU LU

第一章 统计案例

1.1 回归分析的基本思想及其初步应用

课本解读

1. 随机误差

在线性回归模型 $y=bx+a+e$ 中，b 取估计值 $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}$，$a$ 取估计值 $\hat{a}=\overline{y}-\hat{b}\overline{x}$，$e$ 称为随机误差.

2. 线性相关关系

样本相关系数 $r=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}}$.

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

3. 总偏差平方和

把每个效应(观测值减去总的平均值)的平方相加，即用 $\sum\limits_{i=1}^{n}(y_i-\overline{y})^2$ 表示总的效应，称为总偏差平方和.

4. 残差、回归平方和

数据点和它在回归直线上相应位置的差异 $(y_i-\hat{y}_i)$ 是随机误差的效应，称 $\hat{e}_i=y_i-\hat{y}_i$ 为残差.

各个残差的平方和 $\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2$ 称为残差平方和，它代表了随机误差的效应.

总偏差平方和减去残差平方和所得的差称为回归平方和，它代表了解释变量的效应.

相关指数 R^2 的计算公式是 $R^2=1-\dfrac{\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i-\overline{y})^2}$.

R^2 的值越大，说明残差平方和越小，也就是说模型的拟合效果越好.

名师点拨

在"统计"一章中，学习了两个变量之间的相关关系，包括画散点图、求回归直线方程、利用回归直线方程进行预报等内容. 本节在此基础上进一步学习回归分析的基本思想及其初步应用. 从相关系数、相关指数和残差分析等角度探讨回归模型拟合的效果.

第1课时 回归分析(1)

典例剖析

例 1 若回归直线方程中的回归系数 $b=0$，则相关系数 r 为 ( )

(A) 1. (B) -1.

(C) 0. (D) 无法确定.

解题思路 $\because$ $\hat{b}=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}=0$,

$\therefore$ $\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=0$,

$\therefore$ $r=\dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2\ \sum\limits_{i=1}^{n}(y_i-\overline{y})^2}}=0$.

解 选 C.

例 2 为了考察两个变量 x 和 y 之间的线性相关性，甲、乙两位同学各自独立做了 10 次和 15 次试验，并且利用线性回归方法，求得回归直线分别为 l_1 和 l_2. 已知两人测得的变量 x 和 y 的平均值相等，且分别都是 s 和 t，则下列说法正确的是 ( )

(A) 直线 l_1 和直线 l_2 重合.

(B) 直线 l_1 和直线 l_2 平行.

(C) 直线 l_1 和直线 l_2 有公共点 (s,t).

(D) 直线 l_1 和直线 l_2 相交,但交点不一定是(s,t).

解 选C.

注 意

回归直线一定过点$(\overline{x},\overline{y})$.

例3 某产品的广告费支出 x(单位:百万元)与销售额 y(单位:百万元)之间的对应数据如下表:

x	2	4	5	6	8
y	30	40	60	50	70

(1) 画出散点图;

(2) 求回归直线方程;

(3) 在产品的广告费支出为 1 100 万元时,估计销售额是多少?

解 (1) 散点图如图 1-1 所示.

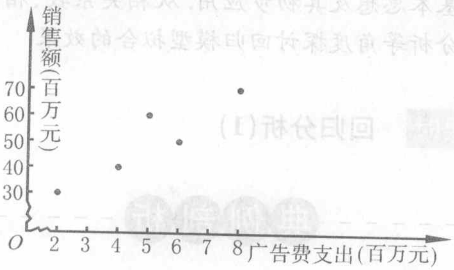


图 1-1

(2) 由已知,可得 $\overline{x}=5$, $\overline{y}=50$, $\sum\limits_{i=1}^{5}x_i^2=145$, $\sum\limits_{i=1}^{5}x_iy_i=1\ 380$,

则$\hat{b}=\dfrac{\sum\limits_{i=1}^{5}x_iy_i-5\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{5}x_i^2-5\overline{x}^2}=\dfrac{1\ 380-5\times5\times50}{145-5\times5^2}=6.5$,

$\hat{a}=\overline{y}-\hat{b}\overline{x}=50-6.5\times5=17.5$,

所求的回归直线方程为$\hat{y}=6.5x+17.5$.

(3) 当 $x=11$ 时,$\hat{y}=6.5\times11+17.5=89$,即销售额的估计值是 8 900 万元.

注 意

计算$\hat{b}$的公式可以改写成$\hat{b}=\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}$.

例4 已知某地单位面积菜地年平均使用氮肥量 x(kg)与单位面积蔬菜年平均产量 y(t)之间的对应数据如下表:

年份	1994	1995	1996	1997	1998	1999	2000	2001
x	70	74	80	78	85	92	90	95
y	5.1	6.0	6.8	7.8	9.0	10.2	10.0	12.0
年份	2002	2003	2004	2005	2006	2007	2008	
x	92	108	115	123	130	138	145	
y	11.5	11.0	11.8	12.2	12.5	12.8	13.0	

(1) 对变量 y 与 x 进行相关性检验;

(2) 若 y 与 x 之间具有线性相关关系,求回归直线方程;

(3) 估计单位面积施肥 150 kg 时,单位面积蔬菜的年平均产量.

解题思路 利用样本相关系数计算公式求出样本相关系数 r. 若 $|r|>0.75$,则线性相关;否则,不线性相关.

解 (1) $n=15$, $\overline{x}=101$, $\overline{y}=10.11$,

$\sum\limits_{i=1}^{15}x_i^2=161\ 125$, $\sum\limits_{i=1}^{15}y_i^2=1\ 628.55$,

$\sum\limits_{i=1}^{15}x_iy_i=16\ 076.8$,于是得

$r=\dfrac{\sum\limits_{i=1}^{15}x_iy_i-15\overline{x}\,\overline{y}}{\sqrt{(\sum\limits_{i=1}^{15}x_i^2-15\overline{x}^2)(\sum\limits_{i=1}^{15}y_i^2-15\overline{y}^2)}}$

$=\dfrac{16\ 076.8-15\times101\times10.11}{\sqrt{(161\ 125-15\times101^2)(1\ 628.55-15\times10.11^2)}}$

$\approx0.864\ 3>0.75$,

故 y 与 x 具有较强的线性相关关系.

(2) 设所求的回归直线方程为$\hat{y}=\hat{b}x+\hat{a}$,

则$\hat{b}=\dfrac{\sum\limits_{i=1}^{15}x_iy_i-15\overline{x}\,\overline{y}}{\sum\limits_{i=1}^{15}x_i^2-15\overline{x}^2}$

$=\dfrac{16\ 076.8-15\times101\times10.11}{161\ 125-15\times101^2}\approx0.093\ 7$,

$\hat{a}=\overline{y}-\hat{b}\overline{x}=10.11-0.093\ 7\times101\approx0.646\ 3$,

所以回归直线方程为$\hat{y}=0.093\ 7x+0.646\ 3$.

(3) 当 $x=150$ 时,$\hat{y}=0.093\ 7\times150+0.646\ 3$

≈14.701,

故单位面积蔬菜的年平均产量为 14.701 kg.

同步训练

理解巩固

1. 设两个变量 x 和 y 之间具有线性相关关系，它们的相关系数是 r，y 关于 x 的回归直线的斜率是 b，纵截距是 a，则（　　）

(A) b 与 r 的符号相同.

(B) a 与 r 的符号相同.

(C) b 与 r 的符号相反.

(D) a 与 r 的符号相反.

2. 设回归直线方程为 $\hat{y}=2-1.5x$，则变量 x 增加 1 个单位时（　　）

(A) y 平均增加 1.5 个单位.

(B) y 平均增加 2 个单位.

(C) y 平均减少 1.5 个单位.

(D) y 平均减少 2 个单位.

3. 下列两个变量关系是线性相关的是（　　）

(A) 人的身高与视力.

(B) 角的大小与所对的圆弧长.

(C) 收入水平与纳税水平.

(D) 人的年龄与身高.

4. 一位母亲记录了儿子 3 岁～9 岁的身高(数据略)，由此建立的身高 y(cm)与年龄 x(岁)的回归直线方程为 $\hat{y}=7.19x+73.93$，由此预测这个孩子 10 岁时的身高（　　）

(A) 一定是 145.83 cm.

(B) 在 145.83 cm 以上.

(C) 在 145.83 cm 左右.

(D) 在 145.83 cm 以下.

5. 作一个关于两个变量的散点图的主要目的是________________________.

6. 回归模型的基本步骤，不仅适用于线性回归模型，也适用于______________，比如用______曲线和________曲线来拟合.

7. 已知回归直线的斜率的估计值是 2.21，样本点的中心是(3,5)，则回归直线方程为________________.

8. 有一项调查，对 9 个不同的 x 值，测得 y 的 9 个对应值如下表：

x	1.5	1.8	2.4	3.0	3.5	3.9	4.4	4.8	5.0
y	4.8	5.7	7.0	8.3	10.9	12.4	13.1	13.6	15.3

试作出该数据的散点图，并由此判断是否存在回归直线. 若有，求出回归直线方程.

9. 要分析初中升学考试中的数学成绩对高一年级学生的数学学习有什么影响，在高一年级中随机抽取 10 名学生，统计他们入学时的数学成绩 x(分)和高一年级期末考试数学成绩 y(分)如下表：

学生编号	1	2	3	4	5	6	7	8	9	10
x	75	45	52	82	90	71	63	58	97	67
y	76	52	70	81	82	84	65	56	95	77

(1) 画出散点图；

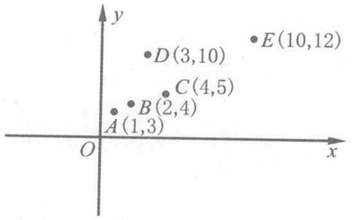
(2) 判断 y 与 x 是否具有线性相关关系；

(3) 若 y 与 x 具有线性相关关系，求回归直线方程；

(4) 若某学生的入学数学成绩为 70 分，试估计他高一期末考试的数学成绩.

发展提高

10. 如图所示的5组数据中，要去掉一组数据，使剩下的4组数据的线性相关系数最大，这组数据是 ( )



(A) B. (B) C. (C) E. (D) D.

11. 在某校高三年级学生学习数学的时间 x(h)与其考试成绩 y(分)之间建立线性回归方程，经计算方程为$\hat{y}=20-0.8x$，该方程中的参数计算 ( )
(A) $\hat{a}$的值是明显错误的.
(B) $\hat{b}$的值是明显错误的.
(C) $\hat{a}$和$\hat{b}$的值都是错误的.
(D) $\hat{a}$和$\hat{b}$的值都是正确的.

12. 某化工厂为预测某产品的回收率 y，需研究 y 和原料有效成分含量 x 之间的相关关系. 现抽取了8对观察值，计算得 $\sum_{i=1}^{8}x_i=52$，$\sum_{i=1}^{8}y_i=228$，$\sum_{i=1}^{8}x_i^2=478$，$\sum_{i=1}^{8}x_iy_i=1\ 849$，则 y 与 x 的回归直线方程为________.

13. 根据下列给出的条件，回答各题中的两个变量的关系.
(1) 某学生的期中考试数学成绩与复习时间投入量的关系；
(2) 各家商场进行促销活动，商品的销售额与广告费的关系.

14. 两个变量 x 和 y 之间的对应数据如下表：

x	1	2	3	4	5	6	7	8
y	1	4	9	16	25	36	49	64

试求出这两个变量之间的回归直线方程，并与 $y=x^2$ 进行比较，考察当 x 越来越大时，此回归直线方程有没有意义？为什么？

第2课时 回归分析(2)

典例剖析

例1 在回归分析中，相关指数 R^2 的值越大，说明残差平方和 ( )
(A) 越小. (B) 越大.
(C) 可能大也可能小. (D) 以上都不正确.

解题思路 利用公式 $R^2=1-\dfrac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$.

解 选A.

例2 炼钢是一个氧化降碳的过程，钢水含碳量的多少直接影响冶炼时间的长短，必须掌握钢水含碳量和冶炼时间的关系. 已测得炉料熔化完毕时钢水的含碳量 x(0.01%)与冶炼时间 y(min)(从炉料熔化完毕到出钢的时间)的对应数据如下表：

x	104	180	190	177	147	134	150	191	204	121
y	100	200	210	185	155	135	170	205	235	125

(1) 作出 y 与 x 的散点图，根据该图猜想它们之间的关系；
(2) 建立以 x 为解释变量，y 为预报变量的回归模型，并计算残差；
(3) 你认为这个模型能较好地刻画 y 与 x 的关系吗？请说明理由.

解 (1) 散点图如图1-2所示.

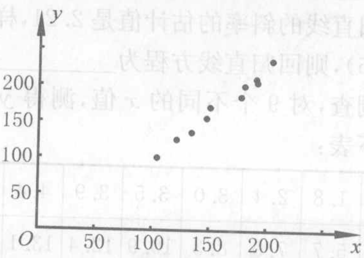


图1-2

从散点图中可以看出 y 与 x 近似呈线性关系.

(2) $\bar{x}=159.8$, $\bar{y}=172$, $\sum_{i=1}^{10} x_i^2=265\ 448$,

$\sum_{i=1}^{10} y_i^2=312\ 350$, $\sum_{i=1}^{10} x_i y_i=287\ 640$.

$$\therefore \hat{b}=\frac{\sum_{i=1}^{10} x_i y_i-10\bar{x}\bar{y}}{\sum_{i=1}^{10} x_i^2-10\bar{x}^2}\approx 1.267,$$

$\hat{a}=\bar{y}-\hat{b}\bar{x}=-30.47$,

回归直线方程为 $\hat{y}=1.267x-30.47$.

残差如下表:

x	104	180	190	177	147
y	100	200	210	185	155
$\hat{y}_i$	101.298	197.59	210.26	193.789	155.779
$\hat{e}_i=y_i-\hat{y}_i$	-1.298	2.41	-0.26	-8.789	-0.779
x	134	150	191	204	121
y	135	170	205	235	125
$\hat{y}_i$	139.308	159.58	211.527	227.998	122.837
$\hat{e}_i=y_i-\hat{y}_i$	-4.308	10.42	-6.527	7.002	2.163

(3) 上面建立的回归方程的相关指数 $R^2=0.962\ 1$,说明 x 能解释约 96.21% 的 y 值变化,因此所建立的模型能够很好地刻画 y 与 x 的关系.

注 意

在具有相关关系的两个变量的散点图中,样本点在一条直线附近,利用线性回归模型,通过计算相关指数判断线性回归模型是否较好地刻画两个变量的关系.

例 3 某城市 2003 年~2008 年人口总数 y(万)与年份 x 的关系如下表:

x	2003	2004	2005	2006	2007	2008
y	50	69	88	110	190	350

(1) 画出散点图,试建立 y 与 x 之间的回归方程;

(2) 据此估计该城市 2009 年人口总数;

(3) 计算相关指数 R^2、残差、残差平方和.

解 (1) 散点图如图 1-3 所示.

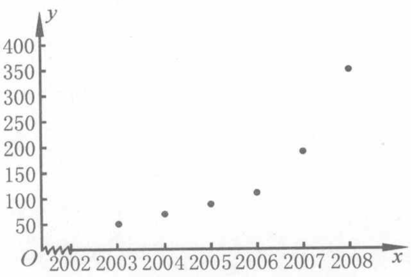


图 1-3

由图知,样本点分布在某一条指数函数曲线 $y=c_1 e^{c_2 x}$ 的周围.

令 $z=\ln y$,则 $z=\hat{b}x+\hat{a}$ ($\hat{a}=\ln c_1$, $\hat{b}=c_2$),得到变换后的数据如下表:

x	2003	2004	2005	2006	2007	2008
z	3.912	4.234	4.477	4.700	5.247	5.858

散点图如图 1-4 所示.

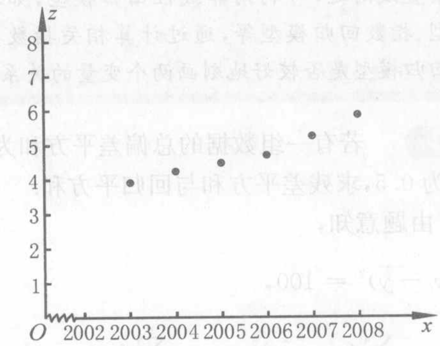


图 1-4

知变换后的样本分布在一条直线附近,故可以用线性回归方程来拟合.

计算得 $\bar{x}=2\ 005.5$, $\bar{z}=4.738$,

$\sum_{i=1}^{6} x_i^2=24\ 132\ 199$, $\sum_{i=1}^{6} z_i^2=137.211\ 202$,

$\sum_{i=1}^{6} x_i z_i=57\ 018.85$,

$$\text{于是}\hat{b}=\frac{\sum_{i=1}^{6} x_i z_i-6\bar{x}\bar{z}}{\sum_{i=1}^{6} x_i^2-6\bar{x}^2}$$

$$=\frac{57\ 018.85-6\times 2\ 005.5\times 4.738}{24\ 132\ 199-6\times 2\ 005.5^2}$$

$=0.371\ 2$, $\hat{a}=\bar{z}-\hat{b}\bar{x}=-739.71$,

故得线性回归方程为 $\hat{z}=0.371\ 2x-739.71$,

因此年份 x 与人口数 y(万)之间的非线性回归方程为 $\hat{y}=e^{0.371\ 2x-739.71}$.

(2) 令 $x=2\ 009$,则 $\hat{y}=e^{0.371\ 2\times 2\ 009-739.71}\approx 416.05$,于是估计 2009 年人口总数为 416.05 万.

(3) 列差如下表:

x	2003	2004	2005	2006	2007	2008
y	50	69	88	110	190	350
$\hat{y}_i$	44.862	65.027	94.255	136.620	198.026	287.034
$\hat{e}_i=y_i-\hat{y}_i$	5.138	3.973	-6.255	-26.620	-8.026	62.966

由表中数据,得残差平方和为 $\sum_{i=1}^{6}(y_i-\hat{y}_i)^2\approx$

4 819.067，又 $\sum_{i=1}^{6}(y_i-\bar{y})^2=63\,296.833$，故相关指数

$$R^2=1-\frac{\sum_{i=1}^{6}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{6}(y_i-\bar{y})^2}=1-\frac{4\,819.067}{63\,296.833}\approx 0.923\,9.$$

注　意

在具有相关关系的两个变量的散点图中，样本点不在一条直线附近，可利用非线性回归模型，如二次回归模型、指数回归模型等，通过计算相关指数判断非线性回归模型是否较好地刻画两个变量的关系．

例4　若有一组数据的总偏差平方和为100，相关指数为0.5，求残差平方和与回归平方和．

解　由题意知，

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=100,$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{100}=0.5,$$

则 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=50$，即残差平方和为50，

于是回归平方和为 $100-50=50$．

同步训练

理解巩固

1. 在对两个变量 x,y 进行线性回归分析时，有下列步骤：

①对所求出的回归方程作出解释；

②收集数据 (x_i,y_i)，$i=1,2,\cdots,n$；

③求线性回归方程；

④求相关系数；

⑤根据所搜集的数据绘制散点图．

若根据可靠性要求能够作出变量 x,y 具有线性相关结论，则下列操作顺序正确的是　　(　　)

(A) ①②⑤③④．　　(B) ③②④⑤①．

(C) ②④③①⑤．　　(D) ②⑤④③①．

2. 给出下列说法：

①回归方程适用于一切样本和总体；

②回归方程一般都有时间性；

③样本取值的范围会影响回归方程的适用范围；

④回归方程得到的预报值是预报变量的精确值．

其中正确的有　　(　　)

(A) ①②．　(B) ②③．　(C) ③④．　(D) ①③．

3. 若散点图中所有样本点都在一条直线上，则解释变量和预报变量之间的残差平方和为　　(　　)

(A) 1．　(B) 0.75．　(C) 0.5．　(D) 0．

4. 若某回归模型相对一组数据的残差平方和为90，相关指数为0.9，则其总偏差平方和为　　(　　)

(A) 9．　(B) 100．　(C) 810．　(D) 900．

5. 对于一组数据的两个回归模型，其残差平方和分别为153.4，200．若从中选取一个拟合程度较好的回归模型，则应选残差平方和为______的那个．

6. 在回归分析中，若残差图中存在异常点（个别数据对应残差过大，或残差呈现不随机的规律性等），则检查数据______，或模型______等．

7. 若某回归模型相对一组数据的总偏差平方和为1 780，其相关指数为0.95，则其残差平方和为______．

8. 某工厂的原料中含有两种有效成分 A 和 B，随机抽取10袋，测得原料中 A 和 B 的含量 x_i,y_i(%)如下表：

i	1	2	3	4	5	6	7	8	9	10
x_i	24	15	23	19	16	11	20	16	17	13
y_i	67	54	72	64	39	22	58	43	46	34

(1) 作出散点图；

(2) 求回归直线方程 $\hat{y}=\hat{b}x+\hat{a}$；

(3) 求残差平方和、总偏差平方和、相关指数 R^2．

9. 为了研究随时间 x(天)变化,某种细菌的繁殖个数 y 的情况,收集到的数据如下表:

x	1	2	3	4	5	6
y	6	12	25	49	95	190

(1) 用天数作解释变量,繁殖个数作预报变量,作出这些数据的散点图;

(2) 建立解释变量与预报变量之间的回归模型;

(3) 计算残差和相关指数;

(4) 你认为这个模型能较好地刻画解释变量与预报变量的关系吗?请说明理由.

发展提高

10. 在研究身高和体重的关系时,若身高解释了 64% 的体重变化,而随机误差贡献了剩余的 36%,则求得的相关指数 R^2 约为 ( )

(A) 0. (B) 0.36. (C) 0.64. (D) 1.

11. 一台机器使用的时间较长,但还可以使用,它按不同的转速生产出来的某机械零件有些会有缺陷,每小时生产有缺陷的零件数 y 随机器的运转速度 x(转/秒)而变化. 抽样试验的结果如下表:

x	16	14	12	8
y	11	9	8	5

(1) 画出散点图;

(2) 若 y 与 x 有线性相关关系,求回归直线方程;

(3) 计算相关指数 R^2;

(4) 若实际生产中,允许每小时生产的产品中有缺陷的零件最多为 10 件,则机器的运转速度应控制在什么范围?

1.2 独立性检验的基本思想及其初步应用

课本解读

1. 分类变量与定量变量

分类变量也称为定性变量,它的取值一定是离散的,不同的取值仅表示个体所属的类别. 如性别变量,只取男、女两个值.

定量变量的取值一定是实数,它们的取值大小有特定的含义,不同取值之间的运算也有特定的含义,如身高、体重等.

2. 列联表

列联表一般为两个或两个以上分类变量的汇总统计表,教科书中仅限于研究两个分类变量的列联表,并且每个分类变量只取两个值,这样的列联表称为 2×2 列联表.

3. 独立性检验

假设有两个分类变量 X 和 Y，它们的值域分别为 $\{x_1,x_2\}$ 和 $\{y_1,y_2\}$，其样本频数列联表（称为 2×2 列联表）为：

	y_1	y_2	总 计
x_1	a	b	$a+b$
x_2	c	d	$c+d$
总 计	$a+c$	$b+d$	$a+b+c+d$

若要推断的论述为 H_1："X 与 Y 有关系"，则可以按如下步骤判断 H_1 成立的可能性：

(1) 先假设 H_0："X 与 Y 没有关系"；

(2) 构造一个随机变量.

$K^2=\dfrac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$，其中 $n=a+b+c+d$ 为样本容量，计算 K^2 的观测值 k；

(3) 统计推断：当 $K^2>2.706$ 时，有 90% 的把握认为"X 与 Y 有关系"；当 $K^2>6.635$ 时，有 99% 的把握认为"X 与 Y 有关系".

4. 通过三维柱形图和二维条形图，可以粗略地判断两个分类变量是否有关系

在二维条形图中，可以估计满足条件 $X=x_1$ 的个体中具有 $Y=y_1$ 的个体所占的比例 $\dfrac{a}{a+b}$，也可以估计满足条件 $X=x_2$ 的个体中具有 $Y=y_1$ 的个体所占的比例 $\dfrac{c}{c+d}$. 两个比例的值相差越大，"X 与 Y 有关系"的可能就越大.

在三维柱形图中，主对角线上两个柱形高度的乘积 ad 与副对角线上两个柱形高度的乘积 bc 相差越大，"X 与 Y 有关系"的可能性越大.

名师点拨

在日常生活中，我们常常关心两个分类变量之间是否有关系. 如吸烟是否与患肺癌有关系，性别是否对于喜欢数学课程有影响等等. 通过三维柱形图和二维条形图，可以粗略地判断两个分类变量是否有关系，但是这种判断无法精确地给出所得结论的可靠程度. 本节介绍利用独立性检验来考察两个分类变量是否有关系，它能较精确地给出这种判断的可靠程度.

独立性检验的思想来自于统计上的假设检验思想，它与反证法类似. 假设检验和反证法都是先假设结论不成立，然后根据是否能够推出"矛盾"来断定结论是否成立. 但两者"矛盾"的含义不同，反证法中的"矛盾"是指不符合逻辑的事件的发生；而假设检验中的"矛盾"是指不符合逻辑的小概率事件的发生，即在结论不成立的假设下推出有利于结论成立的小概率事件的发生.

第1课时 独立性检验(1)

典例剖析

例1 通过三维柱形图和二维条形图（　　）

(A) 可以粗略地判断两个分类变量是否有关系，精确地给出结论的可靠程度.

(B) 可以粗略地判断两个分类变量是否有关系，不能给出结论的可靠程度.

(C) 可以精确地判断两个分类变量是否有关系.

(D) 可以给出结论的可靠程度.

解 选 B.

注 意

在三维柱形图中，主对角线上两个柱形高度的乘积 ad 与副对角线上两个柱形高度的乘积 bc 相差越大，两个分类变量有关系的可能性也越大. 在二维条形图中，满足条件 $X=x_1$ 的个体中具有 $Y=y_1$ 的个体所占比例 $\dfrac{a}{a+b}$，满足条件 $X=x_2$ 的个体中具有 $Y=y_1$ 的个体所占的比例 $\dfrac{c}{c+d}$，两个比例的值相差越大，两个分类变量有关系的可能性也越大.

例2 在独立性检验中，选用随机变量 K^2，用其取值大小来推断独立性是否成立. 当 K^2 满足条件________时，我们有 99% 的把握认为两个分类变量有关系.

解题思路 利用教科书中的临界值表：

$P(K^2\geqslant k)$	0.50	0.40	0.25	0.15	0.10	0.05	0.025	0.010	0.005	0.001
k	0.455	0.708	1.323	2.072	2.706	3.841	5.024	6.635	7.879	10.828

解 $K^2>6.635$.

例3 为了研究色盲与性别的关系，调查了1 000人. 结果男性中442人正常，38人色盲；女性中514人正常，6人色盲.

(1) 根据以上数据建立一个 2×2 列联表；

(2) 在多大程度上可以认为色盲与性别之间有关系？为什么？

解题思路 第(2)题先计算随机变量 K^2 的观测值,然后利用临界值表.

解 (1) 列联表如下:

人数	男	女	总计
正常	442	514	956
色盲	38	6	44
总计	480	520	1 000

(2) 由列联表,得 K^2 的观测值为

$$k=\frac{1\,000(442\times6-38\times514)^2}{956\times44\times480\times520}\approx27.139>10.828.$$

由临界值表知,$P(K^2\geqslant10.828)=0.001$,所以有 99.9%的把握认为色盲与性别有关系.

注 意

已知随机变量 K^2 的观测值,求有多少把握认为两个分类变量有关系(或者已知两个分类变量有关系的把握程度,求随机变量 K^2 的观测值 k 的范围),关键是熟悉临界值表.

例 4 为了检验某种新药的抗癌效果,进行了一项动物试验,其结果如下表:

人数	康复	死亡
用药组	28	32
非用药组	16	40

问:此药是否有效?

解题思路 判断药有效的标准是两个分类变量有关系这一结论的可信程度.

解 列联表如下:

人数	康复	死亡	总计
用药组	28	32	60
非用药组	16	40	56
总计	44	72	116

由列联表,得 K^2 的观测值为

$$k=\frac{116\times(28\times40-16\times32)^2}{60\times56\times44\times72}=4.028>3.841.$$

由临界值表知,$P(K^2\geqslant3.841)=0.05$,

故有 95%的把握认为此药有效.

注 意

独立性检验的应用,利用独立性检验的一般步骤解题.

同步训练

理解巩固

1. 利用独立性检验考察两个分类变量 X,Y 是否有关系,当随机变量 K^2 的观测值 ( )
(A) 越大,“X 与 Y 有关系”成立的可能性越大.
(B) 越大,“X 与 Y 有关系”成立的可能性越小.
(C) 越小,“X 与 Y 有关系”成立的可能性越大.
(D) 与“X 与 Y 有关系”成立的可能性无关.

2. 在作 2×2 列联表的独立性检验时,要求表中的 4 个观测数据 a,b,c,d 都 ( )
(A) 等于 5. (B) 不小于 5.
(C) 小于 5. (D) 大于 5.

3. 若有 95%的把握说明分类变量 X 与 Y 有关系,则 K^2 的观测值 k 应满足 ( )
(A) $k>3.841$. (B) $k\leqslant3.841$.
(C) $k>6.635$. (D) $k\leqslant6.635$.

4. 下面是一个 2×2 列联表:

	y_1	y_2	总计
x_1	a	21	73
x_2	6	25	31
总计	b	46	

则表中 a,b 处的值分别为 ( )
(A) 94,100. (B) 52,46.
(C) 52,58. (D) 58,52.

5. 有一项打鼾与患心脏病的调查中,共调查了 1 671 人,经过计算 $K^2=27.63$. 根据这一数据分析,有理由认为打鼾与患心脏病是________的(填“有关”或“无关”).

6. 某大学正在研究性别与职称(正教授、副教授)之间是否有关系,你认为应该收集的数据是________.

7. 给出下列列联表:

人数	再次发病	未发病
移植手术	39	157
未移植手术	29	167

则 $K^2=$________.

8. 为了研究患慢性气管炎与吸烟量的关系,调查了 228 人,其中每天的吸烟支数在 10 支以上 20 支以下的调查者中,患病者有 98 人,非患病者有 89 人;

每天吸烟支数在20支以上的调查者中，患病者有25人，非患病者有16人.

(1) 根据以上数据建立一个2×2列联表；

(2) 患慢性气管炎是否与吸烟量有关系？

9. 在一次恶劣气候的飞行航程中，调查男、女乘客在飞机上晕机的情况如下表：

人数	晕机	不晕机	总计
男性	24	31	55
女性	8	26	34
总计	32	57	89

据此资料你是否认为在恶劣气候飞行中男性比女性更容易晕机？

发展提高

10. 给出下列关系：

①人的年龄与他(她)拥有的财富之间的关系；

②曲线方程与此曲线上的点的坐标之间的关系；

③苹果的产量与气候之间的关系；

④森林中的同一种树木，其断面直径与高度之间的关系；

⑤学生与其学号之间的关系.

其中有相关关系的是________.

11. 随机抽取某校七年级学生180人，按男、女不同性别分类，并将学生成绩分为中等以上及中等以下两类，结果如下表：

成绩＼人数＼性别	男	女	总计
中等以上	50	70	120
中等以下	19	41	60
总计	69	111	180

性别是否与学习成绩有关？

12. 某地区的一种传染病与饮用水的调查结果如下表：

人数	得病	不得病	总计
干净水	52	466	518
不干净水	94	218	312
总计	146	684	830

得传染病是否与饮用水有关系？

13. 在调查男、女乘客是否晕机的情况下，已知男乘客占总调查人数的$\frac{2}{5}$，其中一半会晕机，而女乘客只有$\frac{1}{3}$的人会晕机，经过测算：有95%以上的把握认为是否晕机与性别有关，那么被调查的人中最少有多少人会晕机？

第 2 课时 独立性检验(2)

典例剖析

例 1 检验双向分类列联表数据，两个分类变量之间是彼此相关还是相互独立的问题时，在常用的方法中，最为精确的做法是 ( )

(A) 三维柱形图. (B) 二维条形图.

(C) 等高条形图. (D) 独立性检验.

解题思路 三维柱形图、二维条形图、等高条形图都只能粗略地判断两个分类变量是否有关系.

解 选 D.

例 2 有人发现，多看电视容易使人近视，下表是一个调查机构对此现象的调查结果：

人数	近视	不近视	总计
多看电视	68	42	110
少看电视	20	38	58
总计	88	80	168

据此推断大约有多大把握认为多看电视与近视有关系.

解 假设 H_0：多看电视与近视无关.

根据列联表中的数据，可得到 K^2 的观测值为

$$k=\frac{168\times(68\times38-42\times20)^2}{110\times58\times88\times80}=11.3765>10.828,$$

而 $P(K^2\geqslant10.828)=0.001$，

即有 99.9% 的把握认为多看电视与近视有关系.

例 3 某高校调查了 56 名男、女大学生在课余时间参加运动的情况，得到的数据如下表：

人数	参加运动	不参加运动	总计
男大学生	20	8	28
女大学生	12	16	28
总计	32	24	56

从表中数据分析有多大把握认为大学生的性别与参加运动之间有关系？

解 假设 H_0：性别与参加体育运动无关.

由列联表中的数据，可得到 K^2 的观测值为

$$k=\frac{56\times(20\times16-12\times8)^2}{28\times28\times32\times24}\approx4.667>3.841,$$

而 $P(K^2\geqslant3.841)=0.05$，

即有 95% 的把握认为性别与参加运动有关.

例 4 对 196 个接受心脏搭桥手术的病人和 196 个接受血管清障手术的病人进行 3 年的跟踪研究，调查他们是否又复发心脏病，得到的数据如下表：

人数	又复发过	未复发过	总计
心脏搭桥手术	39	157	196
血管清障手术	29	167	196
总计	68	324	392

试比较这两种手术对病人又复发心脏病的影响有没有差别.

解 假设 H_0：这两种手术对病人又复发心脏病没有差别.

由列联表数据，可得到 K^2 的观测值为

$$k=\frac{392\times(39\times167-29\times157)^2}{196\times196\times68\times324}\approx1.7792>1.323,$$

而 $P(K^2\geqslant1.323)=0.25$，

即只有 75% 的把握认为两种手术对心脏病复发的影响有差别，故不能作出这两种手术对心脏病复发的影响有差别的结论.

注意

利用独立性判断两个分类变量是否有关系的一般解题步骤是：

①作出统计假设 H_0：两个分类变量没有关系；

②列出两个分类变量的列联表；

③计算 K^2 的观测值 k.

可以通过查阅下表：

$P(K^2\geqslant k)$	k
0.50	0.455
0.40	0.708
0.25	1.323
0.15	2.072
0.10	2.706
0.05	3.841
0.025	5.024
0.010	6.635
0.005	7.879
0.001	10.828

当 $k>2.706$ 时，有 90% 的把握认为两个分类变量有关系；当 $k\leqslant2.706$ 时，认为没有充分的证据显示两个分类变量有关系.