经人民教育出版社授权

配人教版®

总主编◎李朝东

君子曰：学不可以已。青，取之于蓝而青于蓝；冰，水为之而寒于水。木直中绳，揉以为轮，其曲中规；虽有槁暴，不复挺者，揉使之然也。故木受绳则直，金就砺则利，君子博学而日参省乎己，则知明而行无过矣。吾尝终日而思矣，不如须臾之所学也；吾尝跂而望矣，不如登高之博见也。登高而招，臂非加长也，而见者远；顺风而呼，声非加疾也，而闻者彰。假舆马者，非利足也，而致千里；假舟楫者，非能水也，而绝江河。君子生非异也，善假于物也。

积土成山，风雨兴焉；积水成渊，蛟龙生焉；积善成德，而神明自得，圣心备焉。故不积跬步，无以至千里；不积小流，无以成江海。骐骥一跃，不能十步；驽马十驾，功在不舍。锲而舍之，朽木不折；锲而不舍，金石可镂。蚓无爪牙之利，筋骨之

精讲精练

修订版

经典学子

本册主编：吴士刚　高志强

学生用书

选修1-2

高中数学

黄河出版传媒集团
宁夏人民教育出版社

图书在版编目(CIP)数据

精讲精练:人教A版.高中数学.1-2:选修/李朝东主编.
—银川:宁夏人民教育出版社,2009.10(2013.1重印)
ISBN 978-7-80764-211-4

Ⅰ.①精… Ⅱ.①李… Ⅲ.①数学课—高中—教学参考资料 Ⅳ.①G634

中国版本图书馆CIP数据核字(2009)第188519号

精讲精练——数学 选修1-2(人教A版) **李朝东 主编**

责任编辑 柳毅伟
封面设计 杭永鸿
责任印制 刘 丽

黄河出版传媒集团 宁夏人民教育出版社 出版发行

地 址 银川市北京东路139号出版大厦(750001)
网 址 www.yrpubm.com
网上书店 www.hh-book.com
电子信箱 jiaoyushe@yrpubm.com
邮购电话 0951-5014294
经 销 全国新华书店
印刷装订 宁夏雅昌彩色印务有限公司

开 本 880mm×1230mm 1/16 **印 张** 7.5
印刷委托书号 (宁)0010840 **字 数** 150千 **印 数** 2241册
版 次 2009年10月第1版 **印 次** 2013年1月第4次印刷
书 号 ISBN 978-7-80764-211-4/G·1148
定 价 10.02元

目 录

CONTENTS

第一章 统计案例

1.1 回归分析的基本思想及其初步应用

——自·主·探·究——

课标导学 kebiaodaoxue

1. 了解回归分析的基本思想、方法及初步应用.

2. 了解非线性回归问题的解决思路.

3. 提高对现代计算机技术中统计方法的应用认识.

基础梳理 jichushuli

1. 线性回归模型

(1) 函数关系是一种_____关系,相关关系是一种_____关系,回归分析是对具有_____关系的两个变量进行统计分析的一种常用方法,若两个变量之间具有线性相关关系,则称相应的回归分析为_____.

(2) 表示具有相关关系的两个变量组成一组数据,将各组数据在平面直角坐标系中用_____的方法得到的图形叫做散点图.

(3) 线性回归模型 $y=bx+a+e$ 中,e 称为_____,x 称为_____,y 称为_____.

(4) 在线性回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中,$\hat{b}=$_____,$\hat{a}=$_____(其中 $\bar{x}=$_____,$\bar{y}=$_____).

_____称为样本点的中心.

2. 样本相关系数 r 的应用

可用样本相关系数 r 来衡量两个变量之间的线性相关关系.

其中 $r=$_____.

(1) 当 $r>0$ 时,表明两个变量_____;

(2) 当 $r<0$ 时,表明两个变量_____;

(3) r 的绝对值越接近于1,表明两个变量的线性相关性越_____;

(4) r 的绝对值越接近于0,表示两个变量之间几乎_____线性相关关系.

通常当 $|r|$ 大于_____时,认为两个变量有很强的线性相关关系.

3. 刻画回归效果的方式

(1) 残差:把随机误差的估计值 $\hat{e}_i$ 称为相应于点 (x_i,y_i) 的残差.

(2) 残差图:作图时_____为残差,_____可以选为样本编号,或身高数据,或体重估计值等,这样作出的图形称为残差图.

(3) 残差图法

残差点_____地落在水平的带状区域内,说明选用的模型比较适合,这样的带状区域的宽度_____,说明模型拟合精度越高,回归方程的预报精度越高.

(4) 残差平方和法

残差平方和为_____,残差平方和_____,模型拟合效果越好.

(5) 利用 R^2 刻画回归效果

$R^2=1-$_____;R^2 表示_____变量对_____变量变化的贡献率.R^2 越接近于_____,表示回归的效果越好.

【参考答案】

1. (1) 确定性 非确定性 相关 线性回归分析

(2) 描点

(3) 随机误差 解释变量 预报变量

(4) $\dfrac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}$ $\bar{y}-\hat{b}\bar{x}$ $\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$ $\dfrac{1}{n}\sum\limits_{i=1}^{n}y_i$ $(\bar{x},\bar{y})$

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

(1) 正相关 (2) 负相关 (3) 强 (4) 不存在 0.75

3. (2) 纵坐标 横坐标

(3) 比较均匀 越窄

(4) $\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2$ 越小

(5) $\dfrac{\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}$ 解释 预报 1

重·难·点·突·破

疑难剖析
yinanpouxi

1. 线性回归模型问题

(1) 求回归直线方程的一般方法是: 作出散点图,将问题所给的数据在平面直角坐标系中描点,这样表示出的具有相关关系的两个变量的一组数据的图形就是散点图,从散点图中我们可以看出样本点是否呈条形分布,进而判断两个变量是否具有线性相关关系. 如果两个变量呈线性相关关系,那么利用回归系数公式求出$\hat{a},\hat{b}$,代入$\hat{y}=\hat{b}x+\hat{a}$中写出线性回归方程.

(2) 通常情况下,我们获得的样本点不会在一条直线上(这种情况是一种确定性关系——函数关系,而实际中获取的样本点,一般不会呈现这种一次函数关系),所以我们不能用一次函数$y=kx+b$来描述它们之间的关系,而是用线性回归模型$y=bx+a+e$(*)来表示.

其中a和b为模型的未知参数,e称为随机误差.

线性回归模型(*)中,自变量x称为解释变量,因变量y称为预报变量.

因变量y的值是由自变量x和随机误差e共同确定的.

对于线性回归模型(*)中的未知参数a和b,可以用最小二乘法估计$\hat{a}$和$\hat{b}$作为未知参数的最好估计,计算公式是:

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

其中$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$,$(\bar{x},\bar{y})$称为样本点的中心.

2. 线性回归分析

(1) 我们通常用样本相关系数r来描述线性相关关系的强弱,即r的绝对值越接近于1,表明两个变量的线性相关性越强.

(2) 残差分析与残差图

在研究两个变量的关系时,首先要根据散点图来粗略判断它们是否线性相关,是否可以用线性回归模型来拟合数据,然后通过残差: $\hat{e}_1,\hat{e}_2,\cdots,\hat{e}_n$来判断模型的拟合效果,判断原始数据中是否存在可疑数据,这方面的分析工作称为残差分析.

我们可以用残差图进行残差分析,在残差图中,纵坐标为残差,横坐标可以选为样本编号,或解释变量,或预报变量的值,一般选用样本数据的编号为横坐标.

在残差图中,观察出残差比较大的样本点,进一步确认所采集的异常样本点的错误原因,并予以纠正,然后重新利用线性回归模型拟合数据.

如果残差点比较均匀地落在水平的带状区域中,说明选用的模型比较合适,这样的带状区域的宽度越窄,说明模型拟合精度越高,回归方程的预报精度越高.

(3) 相关指数R^2

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. 而我们将$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$称为残差平方和,它代表了随机误差的效应.

我们可以用相关指数R^2来刻画回归的效果,其计算公式是: $R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$.

显然,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.

3. 两种特殊非线性回归模型的转化

(1) 将幂型函数$y=ax^m$($a>0$且为常数,m为常数,x、y取正值)化为线性函数.

将$y=ax^m$两边同取以10为底的对数,则有$\lg y=m\lg x+\lg a$. 令$u=\lg y,v=\lg x,\lg a=b$,代入上式,得$u=mv+b$,其中m、b是常数,这是u、v的线性函数. 如果以u为纵坐标,v为横坐标,则$u=mv+b$的图形就是一条直线.

(2) 将指数型函数$y=ca^x$($a>0$且$a\neq1$,$c>0$且为常数)化为线性函数.

将$y=ca^x$两边同取以10为底的对数,则有$\lg y=x\lg a+\lg c$,令$\lg y=u,\lg a=k,\lg c=b$,得$u=kx+b$,其中k、b是常数,与幂型函数不同的是x依然保持原来的样子,只是用y的对数$\lg y$代替了y.

典型题解
dianxingtijie

题型1 线性回归方程的求法

例1 某班5名学生的数学和物理成绩(单位: 分)如下:

学科成绩 \ 学生	A	B	C	D	E
数学成绩(x)	88	76	73	66	63
物理成绩(y)	78	65	71	64	61

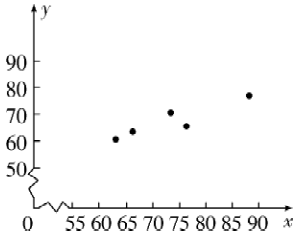
(1) 画出散点图;

(2) 求物理成绩y对数学成绩x的线性回归方程;

(3) 一名学生的数学成绩是96,试预测他的物理成绩.

[**解析**] 本例考查线性回归方程问题,正确掌握公式是解决本例的关键. 依次求出$\bar{x}$、$\bar{y}$、$\sum_{i=1}^{5}x_iy_i$、$\sum_{i=1}^{5}x_i^2$,然后代入公式便可以求出线性回归方程.

[**答案**] (1) 散点图如图所示.



(2) $\because \bar{x}=\frac{1}{5}\times(88+76+73+66+63)=73.2$,

$\bar{y}=\frac{1}{5}\times(78+65+71+64+61)=67.8$,

$\sum_{i=1}^{5}x_iy_i=88\times78+76\times65+73\times71+66\times64+63\times61$

$=25\,054$,

$\sum_{i=1}^{5}x_i^2=88^2+76^2+73^2+66^2+63^2=27\,174$,

$\therefore \hat{b}=\frac{\sum_{i=1}^{5}x_iy_i-5\bar{x}\bar{y}}{\sum_{i=1}^{5}x_i^2-5\bar{x}^2}=\frac{25\,054-5\times73.2\times67.8}{27\,174-5\times73.2^2}\approx0.625$,

$\hat{a}=\bar{y}-\hat{b}\bar{x}=67.8-0.625\times73.2=22.05$.

故 y 对 x 的线性回归方程为 $\hat{y}=0.625x+22.05$.

(3) 当 $x=96$ 时,$\hat{y}=0.625\times96+22.05\approx82$(分),

故可以预测他的物理成绩是82分.

[**点评**] (1) 解决线性回归问题一般先通过散点图来分析两变量间是否具有线性相关关系,然后利用回归直线方程的公式求出线性回归方程,在此基础上,借助线性回归方程对实际问题进行分析.

(2) 计算$\hat{b}$的公式有两个,即

$\hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$和$\hat{b}=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}$.

[**借题发挥1**] 假设某设备的使用年限 x(年)和所支出的维修费用 y(万元),有如下表所示的统计资料:

使用年限 x	2	3	4	5	6
维修费用 y	2.2	3.8	5.5	6.5	7.0

由资料知 y 与 x 具有线性相关关系,试求:

(1) 线性回归方程$\hat{y}=\hat{b}x+\hat{a}$的回归系数$\hat{b}$、$\hat{a}$;

(2) 估计使用年限为10年时的维修费用.

题型2 利用相关指数判断回归模型拟合效果问题

例2 在关于人体的脂肪含量(%)和年龄(岁)关系的研究中,研究人员获得了以下一组数据:

年龄 x	23	27	39	41	45	49	50
脂肪含量 y	9.5	17.8	21.2	25.9	27.5	26.3	28.2
年龄 x	53	54	56	57	58	60	61
脂肪含量 y	29.6	30.2	31.4	30.8	33.5	35.2	34.6

(1) 试问 y 与 x 之间是否有线性相关关系?若有,则求出线性回归方程;

(2) 求相关指数 R^2,并说明其含义;

(3) 给出37岁人的脂肪含量的预测值.

[**解析**] 本例考查回归直线方程及相关指数问题,相关指数运算正确是判断拟合效果好坏的关键.

[**答案**] (1) 由散点图(图略)可知 y 与 x 呈线性相关关系,

设线性回归方程为$\hat{y}=\hat{b}x+\hat{a}$,

由已知数据可求得$\hat{a}\approx-0.448$,$\hat{b}\approx0.576$.

$\therefore$ 线性回归方程为$\hat{y}=0.576x-0.448$.

(2) $R^2=1-\frac{\sum_{i=1}^{14}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{14}(y_i-\bar{y})^2}\approx0.942$,

R^2 为0.942,表明年龄解释了94.2%的脂肪含量变化.

(3) 当 $x=37$ 时,$\hat{y}=0.576\times37-0.448=20.864$.

[**点评**] (1) 数据繁杂的题目,可以采用分步计算的方法或借助计算器.(2) 相关指数 R^2 可以判断回归的拟合效果.

[**借题发挥2**] 为了研究三月下旬的平均气温 x(℃)与四月二十日前棉花害虫化蛹高峰日 y(日)的关系,某地观察了2005年至2010年间的情况,得到下面数据表:

年份	2005	2006	2007	2008	2009	2010
x	24.4	29.5	32.9	28.7	30.3	28.9
y	19	6	1	10	1	8

(1) 求 y 对 x 的线性回归方程,并说明线性回归模型的拟合效果;

(2) 根据规律推断,若该地区2012年三月下旬平均气温为27℃,试估计2012年四月化蛹高峰日为哪一天?

题型3 残差分析

例 3 一个车间为了规定工时定额,需要确定加工零件所花费的时间,为此进行了 10 次试验,测得的数据如下表所示:

零件数 x(个)	10	20	30	40	50	60	70	80	90	100
加工时间 y(min)	62	68	75	81	89	95	102	108	115	122

(1) 计算总偏差平方和,残差及残差平方和;

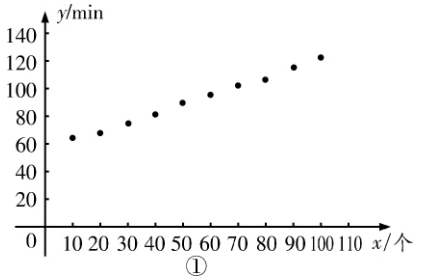
(2) 求出相关指数 R^2;

(3) 作出残差图;

(4) 进行残差分析.

[**解析**] 本例考查有关残差分析问题,涉及公式多而复杂,计算量也很大,只要求能了解有关公式的原理.

[**答案**] (1) 由表中数据,作散点图如图①所示:



①

由图①可以看出,样本点呈条状分布,零件数和加工时间有比较好的线性相关关系,因此可用线性回归方程来近似刻画它们之间的关系.

由表中数据,得加工时间对零件数的线性回归方程为 $\hat{y}=0.67x+54.9$.

列出残差表如下:

y_i	62	68	75	81	89	95	102	108	115	122
$\hat{y}_i$	61.6	68.3	75.0	81.7	88.4	95.1	101.8	108.5	115.2	121.9
$y_i-\bar{y}$	-29.7	-23.7	-16.7	-10.7	-2.7	3.3	10.3	16.3	23.3	30.3
$y_i-\hat{y}_i$	0.40	-0.30	0	-0.70	0.60	-0.1	0.20	-0.50	-0.20	0.10

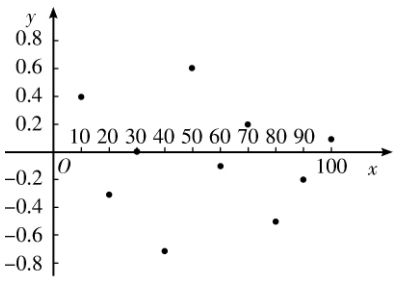
$$\therefore \sum_{i=1}^{10}(y_i-\bar{y})^2=(-29.7)^2+(-23.7)^2+\cdots+30.3^2$$
$$=3\,688.1.$$

$$\sum_{i=1}^{10}(y_i-\hat{y}_i)^2=0.40^2+(-0.30)^2+\cdots+0.10^2=1.45.$$

即总偏差平方和为 3 688.1,残差平方和为 1.45,残差值如表中第四行的值.

(2) $R^2=1-\dfrac{1.45}{3\,688.1}\approx 1-0.000\,39=0.999\,61$,相关指数 R^2 非常接近于 1,回归直线模型拟合效果较好. 由 R^2 得相关数 $r=0.999\,8$.

(3) 作出残差图,横坐标为零件个数,纵坐标为残差,如图②所示:



②

(4) 由 r 的值和散点图都说明 x 与 y 有很强的相关性,由 R^2 的值可以看出回归效果很好,也说明用线性回归模型拟合数据效果很好.

由残差图也可以观察到,第 4 个样本点和第 5 个样本点的残差比较大,需要确认在采集这两个样本点的过程中是否有人为的错误.

[**点评**] 残差图的效果与选用的横坐标无关,作残差分析时,一般从以下几个方面予以说明:(1) 散点图;(2) 相关系数 r;(3) 相关指数 R^2;(4) 残差图中的异常样本点和样本点的带状分布区域的宽窄.

[**借题发挥 3**] 某运动员训练次数(x) 与成绩(y) 之间的数据关系如下表所示:

编号	1	2	3	4	5	6	7	8
次数(x)	30	33	35	37	39	44	46	50
成绩(y)	30	34	37	39	42	46	48	51

(1) 作出散点图;

(2) 求出线性回归方程;

(3) 作出残差图;

(4) 计算 R^2;

(5) 试预测该运动员训练 47 次及 55 次的成绩.

题型4 有关非线性回归问题

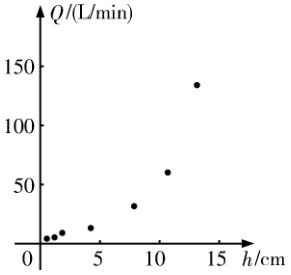
例 4 有一个测量水流量的实验装置,测得试验数据如下表所示:

i	1	2	3	4	5	6	7
水高 h(cm)	0.7	1.1	2.5	4.9	8.1	10.2	13.5
流量 Q(L/min)	0.082	0.25	1.8	11.2	37.5	66.5	134

根据表中数据,建立 Q 与 h 之间的回归方程.

[**解析**] 本例考查了非线性回归问题. 解答本例的关键是先画出散点图,再选择适当的回归模型.

[答案] 由表中测得的数据可以作出散点图,如图所示:



观察散点图中样本点的分布规律,可以判断出样本点分布在某一条曲线上,设表示该曲线的函数模型是

$Q=m\cdot h^n$(m、n是正的常数).

两边取常用对数,则$\lg Q=\lg m+n\cdot\lg h$,令$y=\lg Q$,$x=\lg h$,那么$y=nx+\lg m$,

即为线性函数模型$\hat{y}=\hat{b}x+\hat{a}$(其中$\hat{b}=n$,$\hat{a}=\lg m$).

i	h_i	Q_i	$x_i=\lg h_i$	$y_i=\lg Q_i$	x_i^2	x_iy_i
1	0.7	0.082	−0.154 9	−1.086 2	0.024 0	0.168 3
2	1.1	0.25	0.041 4	−0.602 1	0.001 7	−0.024 9
3	2.5	1.8	0.397 9	0.255 3	0.158 3	0.101 6
4	4.9	11.2	0.690 2	1.049 2	0.476 4	0.724 2
5	8.1	37.5	0.908 5	1.574 0	0.825 4	1.430 0
6	10.2	66.5	1.008 6	1.822 8	1.017 3	1.838 5
7	13.5	134	1.130 3	2.127 1	1.277 6	2.404 3
Σ			4.022	5.140 1	3.780 7	6.642

由上面的数据表,用最小二乘法可求得

$\hat{b}\approx 2.509\,7$,$\hat{a}\approx -0.707\,7$,

故Q对h的非线性回归方程为$Q=0.196\cdot h^{2.509\,7}$.

[点评] 非线性回归问题有时并不给出经验公式,这时我们可以画出已知数据的散点图,把它与《必修1》中学过的各种函数(幂函数、指数函数、对数函数等)图象作比较,挑选一种跟这些散点拟合得最好的函数,然后采用适当的变量转换,把问题转化为线性回归分析问题,使之得到解决.

[借题发挥4] 某城市理论预测2010年到2015年人口总数与年份的关系如下表所示:

年份x(年)	2010	2011	2012	2013	2014	2015
人口数y(万)	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2016年人口总数;

(3) 计算相关指数R^2、残差、残差平方和.

提·升·训·练

1. 在对两个变量x、y进行线性回归分析时一般有下列步骤:

①对所求出的线性回归方程作出解释;

②收集数据(x_i,y_i),$i=1,2,\cdots,n$;

③求线性回归方程;

④求相关指数;

⑤根据所搜集的数据绘制散点图.

如果根据可靠性要求能够判定变量x、y具有线性相关性,则在下列操作顺序中正确的是 ( )

A. ①②⑤③④

B. ③②④⑤①

C. ②④③①⑤

D. ②⑤③④①

2. 已知回归直线的斜率的估计值为1.23,样本点的中心为(4,5),则线性回归方程是 ( )

A. $\hat{y}=1.23x+4$ B. $\hat{y}=1.23x+5$

C. $\hat{y}=0.08x+1.23$ D. $\hat{y}=1.23x+0.08$

3. 由一组数据(x_1,y_1),(x_2,y_2),$\cdots$,(x_n,y_n)得到线性回归方程$\hat{y}=\hat{b}x+\hat{a}$,则下列说法不正确的是 ( )

A. 直线$\hat{y}=\hat{b}x+\hat{a}$必过点$(\bar{x},\bar{y})$

B. 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)中的一个点

C. 直线$\hat{y}=\hat{b}x+\hat{a}$的斜率为$\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n}x_i^2-n\bar{x}^2}$

D. 直线$\hat{y}=\hat{b}x+\hat{a}$与各点(x_1,y_1),(x_2,y_2),$\cdots$,(x_n,y_n)的偏差是该坐标平面上所有直线与这些点的偏差中最小的

4. 已知 y 与 x 的线性回归方程为 $y=2-1.5x$，则变量 x 增加一个单位时，下列说法正确的是 （　　）

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

5. 下表是降耗技术改造后生产甲产品过程中记录的产量 x(t) 与相应的生产能耗 y(吨标准煤) 的几组对应数据，根据表中提供的数据，求出 y 关于 x 的线性回归方程 $\hat{y}=0.7x+0.35$，那么表中 m 的值为 （　　）

x	3	4	5	6
y	2.5	m	4	4.5

A. 4　　B. 3.5　　C. 4.5　　D. 3

6. 在研究硝酸钠的可溶性程度时，对于不同的温度(℃) 下观测它在水中的溶解度，得到的观测结果如下表所示:

温度(x)	0	10	20	50	70
溶解度(y)	66.7	76.0	85.0	112.3	128.0

由此得到回归直线的斜率是_____.

7. 对于一组数据的两个函数模型，其残差平方和分别为 180.2 和 290.7，若从中选取一个拟合程度较好的函数模型，应选第_____种.

8. 在研究身高和体重的关系时，求得相关指数 $R^2\approx$_____，可以叙述为“身高解释了 64% 的体重变化，而随机误差贡献了剩余的 36%”，所以身高对体重的效应比随机误差的效应大得多.

9. 某电脑公司有 6 名产品推销员，其中 5 名推销员的工作年限 x 与推销金额 y 数据如下表所示:

推销员编号	1	2	3	4	5
工作年限 x(年)	3	5	6	7	9
年推销金额 y(万元)	2	3	3	4	5

(1) 求年推销金额 y 关于工作年限 x 的线性回归方程;

(2) 若第 6 名推销员的工作年限为 11 年，试估计他的年推销金额.

10. 已知某种商品价格 x(元) 与需求量 y(件) 之间的关系有如下一组数据:

x	14	16	18	20	22
y	12	10	7	5	3

求 y 关于 x 的线性回归方程，并说明回归模型拟合效果的好坏.

11. 假定小麦基本苗数 x 与成熟期有效穗 y 之间存在相关关系,今测得5组数据如下表所示:

x	15.0	25.8	30.0	36.6	44.4
y	39.4	42.9	42.9	43.1	49.2

(1) 以 x 为解释变量,y 为预报变量,作出散点图;

(2) 求 y 与 x 之间的线性回归方程,并对基本苗数56.7预报其有效穗;

(3) 计算各组残差,并计算残差平方和;

(4) 求 R^2,并说明随机误差对有效穗的效应.

12. 下表是1957年美国旧轿车价格的调查资料,以 x 表示轿车的使用年数,y 表示相应的平均价格,求 y 关于 x 的回归方程.

使用年数 x(年)	1	2	3	4	5	6	7	8	9	10
平均价格 y(美元)	2 651	1 943	1 494	1 087	765	538	484	290	226	204

1.2 独立性检验的基本思想及其初步应用

——自·主·探·究——

课标导学 kebiaodaoxue

1. 了解独立性检验的基本思想、方法,主要记住 K^2 的计算公式.

2. 了解实际推断原理和假设检验的基本思想、方法及初步应用.

基础梳理 jichushuli

1. 变量的不同"值"表示个体所属的不同类别,此类变量称为_____.

2. 列出两个分类变量的_____表,称为列联表.

3. 在独立性检验中,我们常用_____和_____来直观地反映数据情况.

4. 利用随机变量 K^2 来确定在多大程度上可以认为"两个分类变量有关系"的方法称为两个分类变量的_____,其中 $K^2=$_____($n=$_____).

【参考答案】

1. 分类变量

2. 频数

3. 二维条形图　三维柱形图

4. 独立性检验　$\dfrac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$

$a+b+c+d$

重·难·点·突·破

疑难剖析
yinanpouxi

1. 分类变量及列联表

样本中的变量可以取不同的值，例如：性别变量，取男女两个值，商品的等级变量可以取一级、二级、三级等等.

变量的不同值表示个体所属的不同类别，这样的变量称为分类变量，可以用列联表来表示.

列联表一般为两个以上分类变量的汇总统计表，教材中只研究两个分类变量，并且每个分类变量只取两个值的列联表，即 2×2 的列联表，通常列出 2×2 列联表，表中列出各类情况的频数.

一般地，假设两个分类变量 X 和 Y，它们的值域分别为 $\{x_1,x_2\}$ 和 $\{y_1,y_2\}$，其 2×2 列联表为下表所示：

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

由表中的数据构造一个随机变量

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

其中 $n=a+b+c+d$ 为样本容量，其作用是：使不同样本容量的数据使用统一的标准.

两个分类变量是否有关系呢？我们可以利用三维柱形图和二维条形图直观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.

在三维柱形图中，主对角线上两个柱形高度的乘积 ad 与副对角线上的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.

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 $\frac{a}{a+b}$，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 $\frac{c}{c+d}$，两个比例的值相差越大，X 与 Y 有关系的可能性就越大 .

拓展：由 2×2 列联表的频数，可以粗略地看出变量之间是否有关或是否存在差异；由三维柱形图和二维条形图能更直观地反映出相关数据的总体状况，直观地看出两个变量是否有关，但不能给出所得结论的可靠程度，因此还需要进行独立性检验.

2. 独立性检验

通过分析两个分类变量的列联表的数据和柱形（条）图，可以直观地得到两个变量是否有关，当得到的两个变量有关时，与事实是否一致呢？判断当然不是百分之百的正确，那么正确的可能性有多大呢？这就需要用统计观点来考察这个问题.

根据 K^2 的值可以考察两个分类变量是否有关系，并且能较精确地给出这种判断的可靠程度，具体做法是：

（1）根据观测数据（即 2×2 列联表中的频数）计算 $K^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ 的值.

（2）通过查阅下表中 k 的数值，确定 $K^2\geqslant k$ 的范围，从而由表中的数据得出 $P(K^2\geqslant k)$ 的值，即“X 与 Y 有关系”的可靠程度.

$P(K^2\geqslant k)$	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

例如：若 $k>10.828$，就有 99.9% 的把握认为“X 与 Y 有关系”；若 $k>2.706$，就有 90% 的把握认为“X 与 Y 有关系”；若 $k\leqslant2.706$，就认为没有充分的证据显示“X 与 Y 有关系”.

（3）表中结论的适用范围是：观测数据 a,b,c,d 都不小于 5，当观测数据 a,b,c,d 中有小于 5 的数据时，需采用很复杂的精确的检验方法.

独立性检验的基本思想是：独立性检验类似于间接证明中的反证法，要确认“两个分类变量有关系”这一结论成立的可信程度，首先假设该结论不成立，在假设下，我们构造的随机变量 K^2 应该很小. 如果由观测数据计算得到的 K^2 的观测值 k 很大，则在一定程度上说明假设不合理，不合理的程度可由上表中的给定的数值得出.

拓展：独立性检验与反证法的对比如下表所示：

反证法	独立性检验
要证明结论 A	备选假设 H_1
在 A 不成立的前提下进行推理	在 H_1 不成立的条件下进行推理
推出矛盾，意味着结论 A 成立	推出有利于 H_1 成立的小概率事件发生，意味着 H_1 成立的可能性较大
没有找到矛盾，不能对 A 下任何结论，即反证法不成功	推出有利于 H_1 成立的小概率事件不发生，接受原假设

典型题解
dianxingtijie

题型1　利用两种图形判断分类变量是否有关

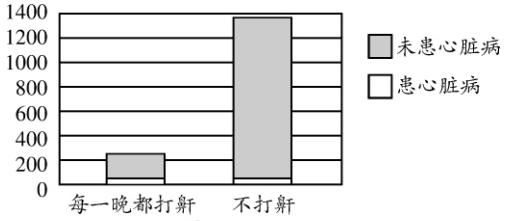
例 1　打鼾不仅影响别人休息，而且还可能与患某种疾病有关，下表是一次调查所得的数据，试求：每一晚都打鼾与患心脏病有关吗？

	患心脏病	未患心脏病	总计
每一晚都打鼾	30	224	254
不打鼾	24	1 355	1 379
总计	54	1 579	1 633

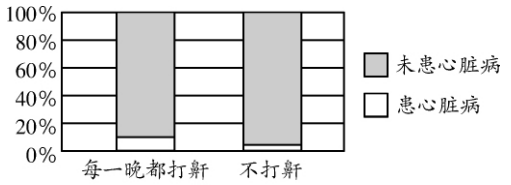
[**解析**]　本例考查利用二维条形图判断分类变量是否有关，掌握利用图形进行分析是解决本题的关键.

[**答案**]　二维条形图如下图①②所示，从图①②中可以粗略

地看出每一晚都打鼾与患心脏病有关.



①



②

[**点评**] (1) 通过二维条形图,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.

(2) 频率分析与图形分析都是对分类变量的一种定性分析.

[**借题发挥1**] 在500个用血清的人身上试验某种血清预防感冒的作用,把一年中的记录与另外500个未用血清的人作比较,得到如下的列联表:

	未感冒	感冒	总计
用血清	252	248	500
未用血清	224	276	500
总计	476	524	1 000

试用二维条形图分析血清是否能起到预防感冒的作用.

题型2 独立性检验

例2 为调查某地区老年人是否需要志愿者提供帮助,用简单随机抽样方法从该地区调查了500位老年人,结果如下:

性别 / 是否需要志愿者	男	女	总计
需要	40	30	70
不需要	160	270	430
总计	200	300	500

能否有99%的把握认为该地区的老年人需要志愿者提供帮助与性别有关?

[**解析**] 本例考查分类变量的独立性检验问题,掌握好K^2的计算公式及应用是解题关键.

[**答案**] 假设该地区的老年人需要志愿者提供帮助与性别无关.由公式,得K^2的观测值

$$k=\frac{500\times(40\times270-30\times160)^2}{200\times300\times70\times430}\approx9.967,$$

由于$9.967>6.635$,

故有99%的把握认为该地区的老年人是否需要志愿者提供帮助与性别有关.

[**点评**] 解决此类问题首先要确定好a,b,c,d,n的值,正确求出K^2的观测值后再与可信度分界值相比较,最后确定有多大把握认为两分类变量之间存在关系.

[**借题发挥2**] 某次航运中,海上出现恶劣气候,随机调查男、女乘客在船上晕船的情况如下表所示:

	晕船	不晕船	总计
男乘客	32	51	83
女乘客	8	24	32
总计	40	75	115

据此资料,你是否认为在恶劣气候中航行,男乘客比女乘客更容易晕船?

例3 为了调查胃病是否与生活规律有关,在某地对540名40岁以上的人进行了调查,结果是患胃病者生活不规律的共60人,患胃病者生活规律的共20人,未患胃病者生活不规律的共260人,未患胃病者生活规律的共200人.

(1) 根据以上的数据列出2×2列联表;

(2) 判断40岁以上的人患胃病与否和生活规律有关系吗?为什么?

[**解析**] 本例考查独立性检验,正确列出2×2列联表,代入K^2公式计算即可.

[**答案**] (1) 由已知可列2×2列联表,得

	患胃病	未患胃病	总计
生活规律	20	200	220
生活不规律	60	260	320
总计	80	460	540

(2) 由列联表可得$|20\times260-60\times200|=6\ 800$,相差较大,可以认为40岁以上的人患胃病与否与生活规律有关系.

再根据列联表中数据,得K^2的观测值

$k=\dfrac{540\times(20\times260-200\times60)^2}{80\times460\times220\times320}\approx9.638$,

$\because 9.638>6.635$.

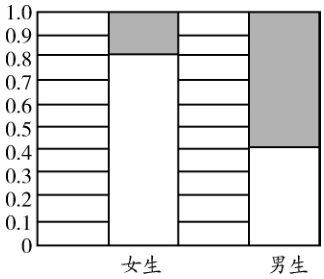
$\therefore$ 有99%的把握认为40岁以上的人患胃病与否和生活规律有关.

[**点评**] 解决独立性检验问题可以先用图形定性判断两个分类变量的关系,再用计算K^2的方法做出分类变量有关或无关的精确可信程度的判断.

[**借题发挥3**] 某企业为了更好地了解设备改造前后与生产合格品的关系,随机抽取了180件产品进行分析. 其中设备改造前生产的合格品有36件,不合格品有49件;设备改造后生产的合格品有65件,不合格品有30件. 根据上面的数据,你能得出什么结论呢?

提·升·训·练

1. 下图是调查某地区男女中学生喜欢理科的等高条形图,阴影部分表示喜欢理科的百分比,从图可以看出 ( )



(第1题)

A. 性别与喜欢理科无关

B. 女生中喜欢理科的比例为80%

C. 男生比女生喜欢理科的可能性大些

D. 男生不喜欢理科的比例为60%

2. 分类变量X和Y的2×2列联表如下,则 ( )

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

A. $ad-bc$越小,说明X与Y的关系越弱

B. $ad-bc$越大,说明X与Y的关系越强

C. $(ad-bc)^2$越大,说明X与Y的关系越强

D. $(ad-bc)^2$越接近于0,说明X与Y的关系越强

3. 某班主任对全班50名学生进行了作业量多少的调查,数据如下表所示:

	认为作业多	认为作业不多	总计
喜欢玩电脑游戏	18	9	27
不喜欢玩电脑游戏	8	15	23
总计	26	24	50

则认为喜欢玩电脑游戏与认为作业量的多少有关系的把握大约为 ( )

A. 99% B. 95%

C. 90% D. 无充分依据

4. 某医疗机构通过抽样调查(样本容量$n=1\,000$),利用2×2列联表和卡方统计量,研究患肺病是否与吸烟有关. 计算得$K^2\approx4.453$,经查临界值表知$P(K^2\geqslant3.841)\approx0.05$,则下列结论正确的是 ( )

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. 在犯错误的概率不超过0.05的前提下,认为"患肺病与吸烟有关"

D. 在犯错误的概率不超过0.95的前提下,认为"患肺病与吸烟有关"

5. 两个分类变量X和Y,可能的取值分别为$\{x_1,x_2\}$和$\{y_1,y_2\}$,其样本频数满足$a=10,b=21,c+d=35$,若X与Y有关系的可信程度为90%,则c的值可能为 ( )

A. 4 B. 5

C. 6 D. 7

6. 随机变量K^2的观测值k越大,说明两个分类变量间有关系的可能性_____.

7. 对于两个分类变量X与Y:

(1) 如果$k>6.635$,就约有_____的把握认为"X与Y有关系";

(2) 如果$k>3.841$,就约有_____的把握认为"X与Y有关系";

(3) 如果$k\leqslant2.706$,就认为_____显示"X与Y有关系".

8. 下面是一个2×2列联表:

	y_1	y_2	总计
x_1	a	21	73
x_2	2	25	27
总计	b	46	100

则表中a,b处的值分别为_____.

9. 有2×2列联表如下:

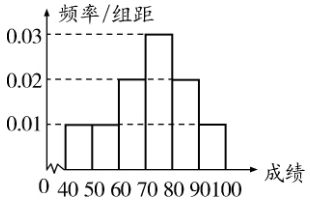
	B	$\overline{B}$	总计
A	54	40	94
$\overline{A}$	32	63	95
总计	86	103	189

由上表可计算K^2的观测值$k\approx$_____.

10. 服用某种维生素对婴儿头发稀疏或稠密的影响调查如下:服用维生素的婴儿有 60 人,头发稀疏的有 5 人;不服用维生素的婴儿有 60 人,头发稀疏的有 46 人,作出 2×2 列联表.

11. 研究人员选取 170 名男女大学生为样本,对他们进行一种心理测验. 发现有 60 名女生对该心理测验中的最后一个题目的反应是:持肯定态度的有 18 名,持否定态度的有 42 名;男生 110 名在相同的题目上持肯定态度的有 22 名,持否定态度的有 88 名. 问:性别与态度之间是否存在某种关系? 分别用二维条形图和独立性检验的方法判断.

12. 某校举办安全法规知识竞赛,从参赛的高一、高二学生中各抽出 100 人的成绩作为样本. 对高一年级的 100 名学生的成绩进行统计,并按 [40,50),[50,60),[60,70),[70,80),[80,90),[90,100] 分组,得到成绩分布的频率分布直方图(如图所示).



(第 12 题)

(1) 若规定 60 分以上(包括 60 分)为合格,计算高一年级这次竞赛的合格率;

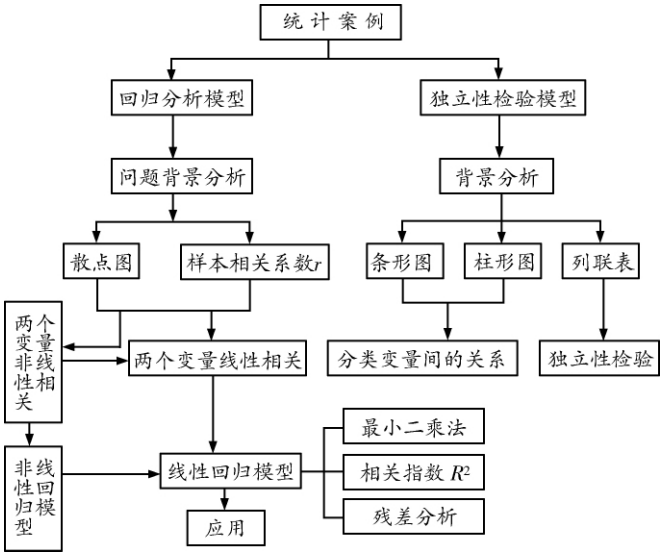
(2) 统计方法中,同一组数据常用该组区间的中点值作为代表,据此,估计高一年级这次知识竞赛的平均成绩;

(3) 若高二年级这次知识竞赛的合格率为 60%,由以上统计数据填写下面 2×2 列联表,并问是否有 99% 的把握认为“这次知识竞赛的成绩与年级有关系”?

	高一	高二	总计
合格人数			
不合格人数			
总计			

单元知识整合

知识网络
zhishiwangluo



专题整合
zhuantizhenghe

专题一　回归分析

1. 线性回归方程$\hat{y}=\hat{b}x+\hat{a}$中的斜率$\hat{b}$和截距$\hat{a}$的确定方法和回归直线过样本点的中心$(\bar{x},\bar{y})$.

2. 样本相关系数 r 的计算公式及 r 的几何意义，$|r|$越接近于 1，表明两个变量的线性相关性越强，当$|r|$大于 0.75 时，认为两个变量有很强的线性相关关系.

3. 残差分析，残差为$\hat{e}_i=y_i-\hat{y}_i$.

4. 样本相关指数 $R^2=1-\dfrac{\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}$，$R^2$ 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.

5. 建立回归模型的基本步骤：

(1) 确定研究对象，明确哪个变量是解释变量，哪个变量是预报变量.

(2) 画出解释变量和预报变量的散点图，观察它们之间的关系.

(3) 由经验确定回归方程的类型.

(4) 按照一定的规则估计回归方程中的参数.

(5) 得出结果后分析残差图是否有异常.

例 1　在 7 块形状、大小相同的试验田上进行施化肥量 x 对水稻产量 y 影响的试验，得到如下表所示的一些数据(单位：kg)：

编号	1	2	3	4	5	6	7
施化肥量 x	15	20	25	30	35	40	45
水稻产量 y	330	345	365	405	445	450	455

(1) 以施化肥量 x 为解释变量，水稻产量 y 为预报产量，作出散点图；

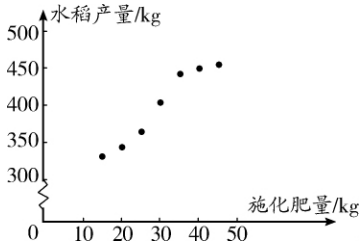
(2) 求 y 与 x 之间的回归方程，并求施化肥量为 28 kg 时的水稻产量的预报值；

(3) 计算残差及残差平方和；

(4) 求相关指数 R^2，并说明随机误差对预报变量的影响有多大.

[**解析**]　本例考查了回归分析，解决的关键是先通过散点图确定好线性相关关系，然后再求解其他问题进行回归分析.

[**答案**]　(1) 散点图如图所示：



(2) 由散点图可以看出，样本点呈条状分布，施化肥量和水稻产量有较好的线性相关关系，因此可以用线性回归方程近似刻画它们之间的关系.

设回归方程$\hat{y}=\hat{b}x+\hat{a}$，

由已知数据，得$\bar{x}=30$，$\bar{y}\approx 399.3$，$\sum\limits_{i=1}^{7}x_i^2=7\,000$，$\sum\limits_{i=1}^{7}x_iy_i=87\,175$，

于是$\hat{b}=\dfrac{\sum\limits_{i=1}^{7}x_iy_i-7\bar{x}\bar{y}}{\sum\limits_{i=1}^{7}x_i^2-7\bar{x}^2}=\dfrac{87\,175-7\times30\times399.3}{7\,000-7\times30^2}\approx4.75$，

$\hat{a}=\bar{y}-\hat{b}\bar{x}=399.3-4.75\times30=256.8$.

因此所求的线性回归方程是$\hat{y}=4.75x+256.8$.

当 $x=28$ 时,水稻产量的预报值是

$\hat{y}=4.75\times28+256.8=389.8(\text{kg})$.

(3) 由残差 $\hat{e}_i=y_i-\hat{y}_i$,得

$\hat{e}_1=1.95,\hat{e}_2=-6.8,\hat{e}_3=-10.55,\hat{e}_4=5.7,$

$\hat{e}_5=21.95,\hat{e}_6=3.2,\hat{e}_7=-15.55,$

故残差平方和 $\sum_{i=1}^{7}\hat{e}_i^2=927.68.$

(4) 计算得 $\sum_{i=1}^{7}(y_i-\bar{y})^2=16\,721.43,$

$\therefore$ 相关指数 $R^2=1-\frac{927.68}{16\,721.43}\approx0.945.$

$\therefore$ 解释变量施化肥量约解释了94.5%的水稻产量变化,

故随机误差约解释了 $1-94.5\%=5.5\%$.

[点评] 解回归分析问题的关键有两个:(1) 求回归方程;(2) 利用残差分析对拟合效果进行评价.

专题二 独立性检验

若两个分类变量 X 和 Y 的值域分别为 $\{x_1,x_2\}$ 和 $\{y_1,y_2\}$,其 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	$a+c$	$b+d$	$a+b+c+d$

则分析它们之间是否有关系的途径有:

(1) 通过计算它们发生的频率,来初步判断;

(2) 通过作三维柱形图和二维条形图,利用图形直观来判断;

(3) 利用独立性检验的基本思想来进行定量的判定.

独立性检验的基本思想类似于反证法,即要确定"两个分类变量 X 与 Y 有关系"这一结论成立的可信程度,首先假设结论不成立,即它们之间没有关系,也就是它们是相互独立的,利用概率的乘法公式可推知 $(ad-bc)^2$ 接近于零,也就是随机变量 $K^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ 应该很小,如果计算出来的它的观测值 k 不是很小,通过查表可知 $P(K^2\geqslant k)$ 的概率很小,又根据小概率事件几乎不可能发生,由此推断假设不成立,从而可以推断 X 与 Y 之间有关系.

例2 在调查的480名男性中有38名患有色盲,520名女性中有6名患有色盲,利用独立性检验的方法来判断色盲与性别是否有关. 你所得的结论在什么范围内有效?

[解析] 本例考查列联表及独立性检验问题,应首先作出调查数据的列联表,再根据列联表进行分析,最后利用独立性检验做出判断.

[答案] 根据题目所给的数据作出如下的 2×2 列联表:

	色盲	不色盲	合计
男	38	442	480
女	6	514	520
合计	44	956	1 000

根据表中所给的数据,得

$a=38,b=442,c=6,d=514,a+b=480,c+d=520,a+c=44,b+d=956,n=1\,000,$

代入公式 $K^2=\frac{n(ad-bc)^2}{(a+b)(b+d)(a+c)(c+d)}$,得 K^2 的观测值

$k=\frac{1\,000\times(38\times514-6\times442)^2}{480\times520\times44\times956}\approx27.139,$

由于 $27.139>10.828$,故我们有99.9%的把握认为色盲和性别有关系. 这个结论只对所调查的480名男性和520名女性有效.

[点评] 利用图形来判断两个变量之间是否有关系,可以画出三维柱形图,也可以画出二维条形图,但从图形上只可以粗略地估计两个分类变量的关系,可以结合所求的数值来进行比较,作图应注意单位统一,图形准确,但它不能对两个分类变量有关或无关做出精确的判断,若要做出精确的判断,可以进行独立性检验的相关计算.

高考鉴赏
gaokao jianshang

例1 (山东) 某产品的广告费用 x 与销售额 y 的统计数据如下表所示:

广告费用 x(万元)	4	2	3	5
销售额 y(万元)	49	26	39	54

根据上表可得回归方程 $\hat{y}=\hat{b}x+\hat{a}$ 中的 $\hat{b}$ 为9.4,据此模型预报广告费用为6万元时销售额为 ( )

A. 63.6万元 B. 65.5万元

C. 67.7万元 D. 72.0万元

[解析] 易求得 $\bar{x}=3.5,\bar{y}=42$,则将 $(3.5,42)$ 代入 $\hat{y}=9.4x+\hat{a}$ 中,得 $42=9.4\times3.5+\hat{a}$,即 $\hat{a}=9.1$,则 $\hat{y}=9.4x+9.1$,$\therefore$ 当广告费用为6万元时销售额为 $9.4\times6+9.1=65.5$(万元).

[答案] B

[点评] 本例考查回归方程等知识,解题时必须明确线性回归直线方程过定点 $(\bar{x},\bar{y})$.

例2 (江西) 为了解儿子身高与其父亲身高的关系,随机抽取5对父子身高数据如下:

父亲身高 x(cm)	174	176	176	176	178
儿子身高 y(cm)	175	175	176	177	177

则 y 对 x 的线性回归方程为 ( )

A. $y=x-1$

B. $y=x+1$

C. $y=88+\frac{1}{2}x$

D. $y=176$

[解析] 本题主要考查线性回归方程以及运算求解能力. 利用公式求系数.

[答案] $\bar{x}=\frac{174+176+176+176+178}{5}=176,$

$\bar{y}=\frac{175+175+176+177+177}{5}=176,$

$$\hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{1}{2},\hat{a}=\bar{y}-\hat{b}\bar{x}=88,$$

$\therefore y=88+\frac{1}{2}x.$

例 3 （广东）为了解篮球爱好者小李的投篮命中率与打篮球时间之间的关系，下表记录了小李某月 1 号到 5 号每天打篮球时间 x（单位：小时）与当天投篮命中率 x 之间的关系：

时间 x	1	2	3	4	5
命中率 y	0.4	0.5	0.6	0.6	0.4

小李这 5 天的平均投篮命中率为_____；用线性回归分析的方法，预测小李该月 6 号打 6 小时篮球的投篮命中率为_____.

［**解析**］ 本题主要考查线性回归方程以及运算求解能力.

小李这 5 天的平均投篮命中率 $\bar{y}=\frac{0.4+0.5+0.6+0.6+0.4}{5}=0.5$，可求得小李这 5 天的平均打篮球时间 $\bar{x}=3$. 根据已有数据可求得 $\hat{b}=0.01$，$\hat{a}=0.47$，故回归方程为 $\hat{y}=0.47+0.01x$，将 $x=6$ 代入得 6 号打 6 小时篮球的投篮命中率约为 0.53.

［**答案**］ 0.5　0.53

例 4 （湖南）通过随机询问 110 名性别不同的大学生是否爱好某项运动，得到如下的列联表：

	男	女	总计
爱好	40	20	60
不爱好	20	30	50
总计	60	50	110

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

$$K^2=\frac{110\times(40\times30-20\times20)^2}{60\times50\times60\times50}\approx7.8.$$

附表：

$P(K^2\geqslant k)$	0.050	0.010	0.001
k	3.841	6.635	10.828

参照附表，得到的正确结论是（　　）

A. 有 99% 以上的把握认为“爱好该项运动与性别有关”

B. 有 99% 以上的把握认为“爱好该项运动与性别无关”

C. 在犯错误的概率不超过 0.1% 的前提下，认为“爱好该项运动与性别有关”

D. 在犯错误的概率不超过 0.1% 的前提下，认为“爱好该项运动与性别无关”

［**解析**］ $\because 6.635<7.8<10.828$，$\therefore$ 选 A.

［**答案**］ A

［**点评**］ 本例考查了独立性检验、解决实际问题的能力.

例 5 （安徽）某地最近十年粮食需求量逐年上升，下表是部分统计数据：

年份	2002	2004	2006	2008	2010
需求量（万吨）	236	246	257	276	286

（1）利用所给数据求年需求量与年份之间的回归直线方程 $\hat{y}=bx+a$；

（2）利用（1）中所求的直线方程预测该地 2012 年的粮食需求量.

［**解析**］ （1）先对大数据进行处理，再求回归直线方程较为简单.（2）求出回归直线方程后，即求当 $x=2\,012$ 时 y 的值.

［**答案**］ 由所给数据分析，年需求量与年份之间具有线性相关关系，可用线性回归方程来近似刻画，为此对数据可作预处理如下表所示：

年份 -2006	-4	-2	0	2	4
需求量 -257	-21	-11	0	19	29

对处理后的数据，容易算得

$$\bar{x}=\frac{1}{5}\times(-4-2+0+2+4)=0,$$

$$\bar{y}=\frac{1}{5}\times(-21-11+0+19+29)=3.2,$$

$$\sum_{i=1}^{5}x_iy_i=-4\times(-21)+(-2)\times(-11)+2\times19+4\times29=260,$$

$$\sum_{i=1}^{5}x_i^2=16+4+0+4+16=40.$$

$$\therefore b=\frac{\sum_{i=1}^{5}x_iy_i-5\bar{x}\bar{y}}{\sum_{i=1}^{5}x_i^2-5\bar{x}^2}=\frac{260}{40}=6.5,$$

$\therefore a=\bar{y}-b\bar{x}=3.2.$

$\therefore$ 所求回归直线方程为 $\hat{y}-257=6.5(x-2\,006)+3.2$，即 $\hat{y}=6.5(x-2\,006)+260.2$.

（2）当 $x=2\,012$ 时，$\hat{y}=6.5\times(2\,012-2\,006)+260.2=299.2$（万吨），

故预测 2012 年粮食需求量为 299.2 万吨.

［**点评**］ 本例考查回归思想及初步应用，考查回归直线的求法与用法及数据处理的基本方法和能力，考查用统计知识解决实际应用问题的能力.

例 6 （辽宁）为了比较注射 A、B 两种药物后产生的皮肤疱疹的面积，选 200 只家兔做试验，将这 200 只家兔随机地分成两组，每组 100 只，其中一组注射药物 A，另一组注射药物 B. 下表 1 和表 2 分别是注射药物 A 和药物 B 后的试验结果.（疱疹面积单位：mm^2）

表 1　注射药物 A 后皮肤疱疹面积的频数分布表

疱疹面积	[60,65)	[65,70)	[70,75)	[75,80)
频数	30	40	20	10