**第二讲变量间的相关关系与统计案例**

id:2147491920;FounderCES

题组1回归分析

1*.*[2017山东,5,5分][理]为了研究某班学生的脚长*x*(单位:厘米)和身高*y*(单位:厘米)的关系,从该班随机抽取10名学生,根据测量数据的散点图可以看出*y*与*x*之间有线性相关关系,设其回归直线方程为*=x+*,已知*xi=*225,*yi=*1 600,*=*4*.*该班某学生的脚长为24,据此估计其身高为()

A.160 B.163 C.166 D.170

2*.*[2015湖北,4,5分]已知变量*x*和*y*满足关系*y=-*0*.*1*x+*1,变量*y*与*z*正相关*.*下列结论中正确的是()

A.*x*与*y*正相关,*x*与*z*负相关

B.*x*与*y*正相关,*x*与*z*正相关

C.*x*与*y*负相关,*x*与*z*负相关

D.*x*与*y*负相关,*x*与*z*正相关

3*.*[2014重庆,3,5分][理]已知变量*x*与*y*正相关,且由观测数据算得样本平均数*=*3,*=*3*.*5,则由该观测数据算得的线性回归方程可能是()

A.*=*0*.*4*x+*2*.*3 B*.=*2*x-*2*.*4 C*.=-*2*x+*9*.*5 D.*=-*0*.*3*x+*4*.*4

4*.*[2014湖北,4,5分][理]根据如下样本数据

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *x* | 3 | 4 | 5 | 6 | 7 | 8 |
| *y* | 4*.*0 | 2*.*5 | *-*0*.*5 | 0*.*5 | *-*2*.*0 | *-*3*.*0 |

得到的回归方程为*=bx+a*,则()

A*.a>*0,*b>*0 B*.a>*0,*b<*0 C*.a<*0,*b>*0 D*.a<*0,*b<*0

5*.*[2015 新课标全国Ⅰ,19,12分][理]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*xi*和年销售量*yi*(*i=*1,2,…,8)数据作了初步处理,得到图14*-*2*-*1所示的散点图及一些统计量的值*.*

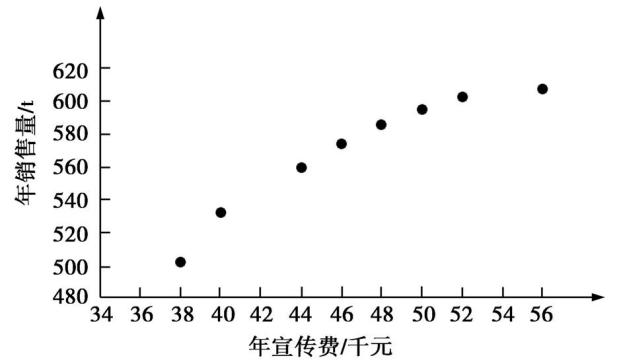


图14*-*2*-*1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | (*xi-*)2 | (*wi-*)2 | (*xi-*)(*yi-*) | (*wi-*)(*yi-*) |
| 46*.*6 | 563 | 6*.*8 | 289*.*8 | 1*.*6 | 1 469 | 108*.*8 |

表中*wi=*,*=wi.*

(Ⅰ)根据散点图判断,*y=a+bx*与*y=c+d*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)

(Ⅱ)根据(Ⅰ)的判断结果及表中数据,建立*y*关于*x*的回归方程;

(Ⅲ)已知这种产品的年利润*z*与*x*,*y*的关系为*z=*0*.*2*y-x.*根据(*Ⅱ*)的结果回答下列问题:

(i)年宣传费*x=*49时,年销售量及年利润的预报值是多少?

(ii)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),…,(*un*,*vn*),其回归直线*v=α+βu*的斜率和截距的最小二乘估计分别为*=*,*=-.*

题组2独立性检验在实际问题中的应用

6*.*[2014安徽,17,12分]某高校共有学生15 000人,其中男生10 500人,女生4 500人*.*为调查该校学生每周平均体育运动时间的情况,采用分层抽样的方法,收集300位学生每周平均体育运动时间的样本数据(单位:小时)*.*

(Ⅰ)应收集多少位女生的样本数据?

(Ⅱ)根据这300个样本数据,得到学生每周平均体育运动时间的频率分布直方图(如图14*-*2*-*2所示),其中样本数据的分组区间为:[0,2],(2,4],(4,6],(6,8],(8,10],(10,12]*.*估计该校学生每周平均体育运动时间超过4小时的概率;

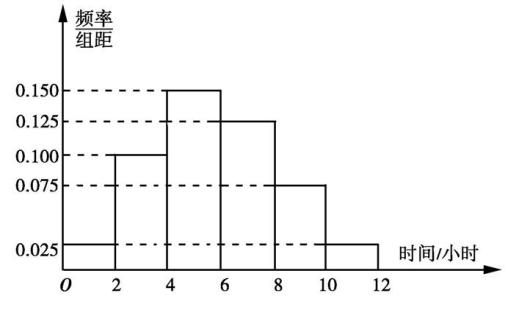


图14*-*2*-*2

(Ⅲ)在样本数据中,有60位女生的每周平均体育运动时间超过4小时,请完成每周平均体育运动时间与性别列联表,并判断是否有95*%*的把握认为“该校学生的每周平均体育运动时间与性别有关”*.*

附:*K*2*=*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *P*(*K*2≥*k*0) | 0*.*10 | 0*.*05 | 0*.*010 | 0*.*005 |
| *k*0 | 2*.*706 | 3*.*841 | 6*.*635 | 7*.*879 |

7*.*[2014辽宁,18,12分]某大学餐饮中心为了解新生的饮食习惯,在全校一年级学生中进行了抽样调查,调查结果如下表所示:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 喜欢甜品 | 不喜欢甜品 | 合计 |
| 南方学生 | 60 | 20 | 80 |
| 北方学生 | 10 | 10 | 20 |
| 合计 | 70 | 30 | 100 |

(*Ⅰ*)根据表中数据,问是否有95*%*的把握认为“南方学生和北方学生在选用甜品的饮食习惯方面有差异”;

(*Ⅱ*)已知在被调查的北方学生中有5名数学系的学生,其中2名喜欢甜品*.*现在从这5名学生中随机抽取3人,求至多有1人喜欢甜品的概率*.*

附: *χ*2*=*,

|  |  |
| --- | --- |
| *P*(*χ*2≥*k*) | 0*.*1000*.*0500*.*010 |
| *k* | 2*.*7063*.*8416*.*635 |

id:2147491995;FounderCES

**A组基础题**

1*.*[2018惠州市二调,7]某商场为了了解毛衣的月销售量*y*(单位:件)与月平均气温*x*(单位:℃)之间的关系,随机统计了某4个月的月销售量与当月平均气温,其数据如下表:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 月平均气温*x/*℃ | 17 | 13 | 8 | 2 |
| 月销售量*y/*件 | 24 | 33 | 40 | 55 |

由表中数据算出线性回归方程*=x+*中的*=-*2,气象部门预测下个月的平均气温为6 ℃,据此估计该商场下个月毛衣销售量为()

A.46件 B.40件 C.38件 D.58件

2*.*[2018湖南省五市十校联考,3]下表提供了某工厂节能降耗技术改造后,一种产品的产量*x*(单位:吨)与相应的生产能耗*y*(单位:吨)的几组对应数据:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *x/*吨 | 3 | 4 | 5 | 6 |
| *y/*吨 | 2*.*5 | *t* | 4 | 4*.*5 |

根据上表提供的数据,求得*y*关于*x*的线性回归方程为*=*0*.*7*x+*0*.*35,那么表格中*t*的值为()

A*.*3 B*.*3*.*15 C*.*3*.*25 D*.*3*.*5

3*.*[2018广东省百校联盟第二次联考,3]下表是我国某城市在2017年1月份至10月份期间各月最低温度与最高温度(单位:℃)的数据一览表*.*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 最高温  度*/*℃ | 5 | 9 | 9 | 11 | 17 | 24 | 27 | 30 | 31 | 21 |
| 最低温  度*/*℃ | *-*12 | *-*3 | 1 | *-*2 | 7 | 17 | 19 | 23 | 25 | 10 |

已知该城市的各月最低温与最高温具有相关关系,根据该一览表,则下列结论错误的是()

A.最低温度与最高温度为正相关

B.每月最高温度与最低温度的平均值在前8个月逐月增加

C.月温差(最高温度减最低温度)的最大值出现在1月

D.1月至4月的月温差(最高温度减最低温度)相对于7月至10月,波动性更大

4*.*[2017长沙市五月模拟,5]为了解某社区居民购买水果和牛奶的年支出费用与购买食品的年支出费用的关系,随机调查了该社区5户家庭,得到如下统计数据表:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 购买食品的年  支出费用*x/*万元 | 2*.*09 | 2*.*15 | 2*.*50 | 2*.*84 | 2*.*92 |
| 购买水果和牛奶的  年支出费用*y/*万元 | 1*.*25 | 1*.*30 | 1*.*50 | 1*.*70 | 1*.*75 |

根据上表可得回归直线方程*=x+*,其中*=*0*.*59,*=-*,据此估计该社区一户购买食品的年支出费用为3*.*00万元的家庭购买水果和牛奶的年支出费用为()

A.1*.*795万元 B.2*.*555万元 C.1*.*915万元 D.1*.*945万元

5*.*[2018成都市高三摸底测试,18]某医疗科研项目组对5只实验小白鼠体内的*A*,*B*两项指标数据进行收集和分析,得到的数据如下表:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 指标 | 1号小白鼠 | 2号小白鼠 | 3号小白鼠 | 4号小白鼠 | 5号小白鼠 |
| *A* | 5 | 7 | 6 | 9 | 8 |
| *B* | 2 | 2 | 3 | 4 | 4 |

(1)若通过数据分析,得知*A*项指标数据与*B*项指标数据具有线性相关关系*.*试根据上表,求*B*项指标数据*y*关于*A*项指标数据*x*的线性回归方程*=x+*;

(2)现要从这5只小白鼠中随机抽取3只,求其中至少有一只小白鼠的*B*项指标数据高于3的概率*.*

参考公式:*==*,*=-.*

**B组提升题**

6*.*[2018惠州市一调,19]近年来,随着我国在教育科研上的投入不断加大,科学技术得到迅猛发展,国内企业的国际竞争力得到大幅提升*.*伴随着国内市场增速放缓,国内有实力的企业纷纷进行海外布局,第二轮企业出海潮到来*.*如在智能手机行业,国产品牌已在赶超国外巨头,某品牌手机公司一直默默拓展海外市场,在海外共设30多个分支机构,需要国内公司外派大量70后、80后中青年员工*.*该企业为了解这两个年龄层员工是否愿意被外派工作的态度,按分层抽样的方法从70后和80后的员工中随机调查了100位,得到的数据如下表:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 愿意被外派 | 不愿意被外派 | 总计 |
| 70后 | 20 | 20 | 40 |
| 80后 | 40 | 20 | 60 |
| 总计 | 60 | 40 | 100 |

(1)根据调查的数据,是否有90*%*以上的把握认为“是否愿意被外派与年龄有关”,并说明理由;

(2)该公司欲举行参观驻海外分支机构的交流体验活动,拟安排6名参与调查的70后、80后员工参加*.*70后员工中有愿意被外派的3人和不愿意被外派的3人报名参加,从中随机选出3人,记选到愿意被外派的人数为*x*;80后员工中有愿意被外派的4人和不愿意被外派的2人报名参加,从中随机选出3人,记选到愿意被外派的人数为*y.*求*x<y*的概率*.*

参考数据:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *P*(*K*2≥*k*0) | 0*.*15 | 0*.*10 | 0*.*05 | 0*.*025 | 0*.*010 | 0*.*005 |
| *k*0 | 2*.*072 | 2*.*706 | 3*.*841 | 5*.*024 | 6*.*635 | 7*.*879 |

参考公式:*K*2 *=* ,其中*n=a+b+c+d.*

7*.*[2017桂林、百色、梧州、崇左、北海五市联考,18]某公司为了准确地把握市场,做好产品生产计划,对过去四年的数据进行整理得到了第*x*年与年销售量*y*(单位:万件)之间的关系如下表*.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *x* | 1 | 2 | 3 | 4 |
| *y* | 12 | 28 | 42 | 56 |

(1)在图14*-*2*-*3中画出表中数据的散点图;

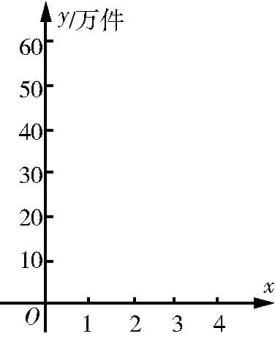


图14*-*2*-*3

(2)根据散点图选择合适的回归模型拟合*y*与*x*的关系(不必说明理由);

(3)建立*y*关于*x*的回归方程,预测第5年的销售量*.*

参考公式:回归直线的斜率和截距的最小二乘估计分别为

*==*,*=-.*

8*.*[2017四川省重点中学高三第二次学习情况评估,18] 某课题组对全班45名同学的饮食习惯进行了一次调查,并用如图14*-*2*-*4所示的茎叶图表示45名同学的饮食指数*.*说明:饮食指数低于70的人被认为喜食蔬菜,饮食指数不低于70的人被认为喜食肉类*.*

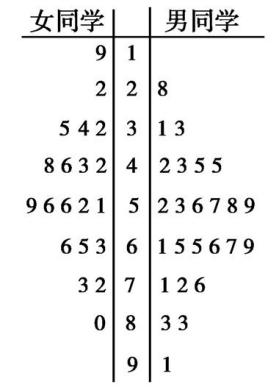


图14*-*2*-*4

(1)根据茎叶图,完成下面2*×*2列联表,并判断是否有90*%*的把握认为“喜食蔬菜还是喜食肉类与性别有关”,说明理由;

|  |  |  |  |
| --- | --- | --- | --- |
|  | 喜食蔬菜 | 喜食肉类 | 总计 |
| 男同学 |  |  |  |
| 女同学 |  |  |  |
| 总计 |  |  |  |

(2)用分层抽样的方法按照喜食蔬菜、喜食肉类从全班同学中随机抽取15名同学进行进一步调查,记抽到的喜食肉类的女同学的人数为*ξ*,求*ξ*的分布列和数学期望*E*(*ξ*)*.*

附:*K*2*=*

|  |  |  |  |
| --- | --- | --- | --- |
| *P*(*K*2≥*k*0) | 0*.*10 | 0*.*05 | 0*.*01 |
| *k*0 | 2*.*706 | 3*.*841 | 6*.*635 |

**答案**

id:2147495261;FounderCES

1*.*C由题意可知*=*4*x+*,又*=*22*.*5,*=*160,则160*=*22*.*5*×*4*+*,解得*=*70,所以*=*4*x+*70*.*当*x=*24时,*=*4*×*24*+*70*=*96*+*70*=*166*.*

2*.*C因为*y=-*0*.*1*x+*1,*x*的系数为负,所以*x*与*y*负相关;而*y*与*z*正相关,故*x*与*z*负相关*.*故选C*.*

3*.*A依题意知,相应的回归直线的斜率应为正,排除选项C,D*.*且直线必过点(3,3*.*5),代入选项A,B中的方程可得A正确*.*选A*.*

4*.*B根据题中表内数据画出散点图(图略),由散点图可知*b<*0,*a>*0,故选B*.*

5*.*(Ⅰ)由散点图可以判断,*y=c+d*适宜作为年销售量*y*关于年宣传费*x*的回归方程类型*.*

(Ⅱ)令*w=*,先建立*y*关于*w*的线性回归方程*.*由于*===*68,

*=-=*563*-*68*×*6*.*8*=*100*.*6,

所以*y*关于*w*的线性回归方程为*=*100*.*6*+*68*w*,因此*y*关于*x*的回归方程为*=*100*.*6*+*68*.*

(Ⅲ)(i)由(Ⅱ)知,当*x=*49时,年销售量*y*的预报值*=*100*.*6*+*68*×=*576*.*6,

年利润*z*的预报值*=*576*.*6*×*0*.*2*-*49*=*66*.*32*.*

(ii)根据(Ⅱ)的结果知,年利润*z*的预报值*=*0*.*2*×*(100*.*6*+*68)*-x=-x+*13*.*6*+*20*.*12*.*

所以当*==*6*.*8,即*x=*46*.*24时,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*.*

6*.*(Ⅰ)300*×=*90,所以应收集90位女生的样本数据*.*

(Ⅱ)由频率分布直方图得1*-*2*×*(0*.*100*+*0*.*025)*=*0*.*75,所以该校学生每周平均体育运动时间超过4小时的概率的估计值为0*.*75*.*

(Ⅲ)由(Ⅱ)知,300位学生中有300*×*0*.*75*=*225人的每周平均体育运动时间超过4小时,75人的每周平均体育运动时间不超过4小时*.*又样本数据中有210份是关于男生的,90份是关于女生的*.*所以每周平均体育运动时间与性别列联表如下:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 男生 | 女生 | 总计 |
| 每周平均体育运动时间不超过4小时 | 45 | 30 | 75 |
| 每周平均体育运动时间超过4小时 | 165 | 60 | 225 |
| 总计 | 210 | 90 | 300 |

结合列联表可算得*K*2的观测值*k==*≈4*.*762*>*3*.*841*.*

所以有95*%*的把握认为“该校学生的每周平均体育运动时间与性别有关”*.*

7*.*(Ⅰ)将2*×*2列联表中的数据代入公式计算,得

*χ*2*===*≈4*.*762*.*

由于4*.*762*>*3*.*841,所以有95*%*的把握认为“南方学生和北方学生在选用甜品的饮食习惯方面有差异”*.*

(Ⅱ)从5名数学系学生中任取3人的一切可能结果所组成的基本事件空间*Ω=*{(*a*1,*a*2,*b*1),(*a*1,*a*2,*b*2),(*a*1,*a*2,*b*3),(*a*1,*b*1,*b*2),(*a*1,*b*2,*b*3),(*a*1,*b*1,*b*3),(*a*2,*b*1,*b*2),(*a*2,*b*2,*b*3),(*a*2,*b*1,*b*3),(*b*1,*b*2,*b*3)},

其中*ai*表示喜欢甜品的学生,*i=*1,2*.bj*表示不喜欢甜品的学生, *j=*1,2,3*.*

*Ω*由10个基本事件组成,且这些基本事件的出现是等可能的*.*

用*A*表示“3人中至多有1人喜欢甜品”这一事件,则

*A=*{(*a*1,*b*1,*b*2),(*a*1,*b*2,*b*3),(*a*1,*b*1,*b*3),(*a*2,*b*1,*b*2),(*a*2,*b*2,*b*3),(*a*2,*b*1,*b*3),(*b*1,*b*2,*b*3)}*.*

事件*A*由7个基本事件组成,因而*P*(*A*)*=.*

id:2147495276;FounderCES

**A组基础题**

1*.*A由题中数据,得*=*10,*=*38,回归直线*=x+*过点(*,*),且*=-*2,所以*=*58,则回归方程为*=-*2*x+*58,所以当*x=*6时,*=*46,故选A*.*

2*.*A*==*4*.*5,*==*,线性回归方程过样本点的中心(,),所以*=*0*.*7*×*4*.*5*+*0*.*35,解得*t=*3,选A*.*

3*.*B将最高温度、最低温度、温差列表如下,

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 月份 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 最高温  度*/*℃ | 5 | 9 | 9 | 11 | 17 | 24 | 27 | 30 | 31 | 21 |
| 最低温  度*/*℃ | *-*12 | *-*3 | 1 | *-*2 | 7 | 17 | 19 | 23 | 25 | 10 |
| 温差  度*/*℃ | 17 | 12 | 8 | 13 | 10 | 7 | 8 | 7 | 6 | 11 |

由表格可知,最低温度大致随最高温度的增大而增大,A正确;每月最高温度与最低温度的平均值在前8个月不是逐月增加,B错;月温差的最大值出现在1月,C正确;1月至4月的月温差相对于7月至10月,波动性更大,D正确*.*故选B*.*

4*.*A*==*2*.*50(万元),*==*1*.*50(万元),又*=*0*.*59,所以*=-=*0*.*025,*=*0*.*59*x+*0*.*025,故年支出费用为3*.*00万元的家庭购买水果和牛奶的年支出费用约为*=*0*.*59*×*3*.*00*+*0*.*025*=*1*.*795万元*.*

5*.*(1)由题意,可得*=*7,*=*3,*xiyi=*110,*=*255,*==.*

*∵=-*,∴*=-.*

∴所求线性回归方程为*=x-.*

(2)设1号至5号小白鼠依次为*a*1,*a*2,*a*3,*a*4,*a*5,则在这5只小白鼠中随机抽取3只的抽取情况有*a*1*a*2*a*3,*a*1*a*2*a*4,*a*1*a*2*a*5,*a*1*a*3*a*4,*a*1*a*3*a*5,*a*1*a*4*a*5,*a*2*a*3*a*4,*a*2*a*3*a*5,*a*2*a*4*a*5,*a*3*a*4*a*5,共10种*.*

随机抽取的3只小白鼠中至少有一只的*B*项指标数据高于3的情况有*a*1*a*2*a*4,*a*1*a*2*a*5,*a*1*a*3*a*4,*a*1*a*3*a*5,*a*1*a*4*a*5,*a*2*a*3*a*4,*a*2*a*3*a*5,*a*2*a*4*a*5,*a*3*a*4*a*5,共9种*.*

∴从这5只小白鼠中随机抽取3只,其中至少有一只小白鼠的*B*项指标数据高于3的概率为*.*

**B组提升题**

6*.*(1)有90*%*以上的把握认为“是否愿意被外派与年龄有关”,理由如下:

由题意可知,*K*2的观测值*k==*≈2*.*778*>*2*.*706,

所以有90*%*以上的把握认为“是否愿意被外派与年龄有关”*.*

(2)“*x<y*”包含“*x=*0,*y=*1”“*x=*0,*y=*2”“*x=*0,*y=*3”“*x=*1,*y=*2”“*x=*1,*y=*3”“*x=*2,*y=*3”六个事件,

且*P*(*x=*0,*y=*1)*=*×*=*,

*P*(*x=*0,*y=*2)*=*×*=*,

*P*(*x=*0,*y=*3)*=*×=,

*P*(*x=*1,*y=*2)*=*×=,

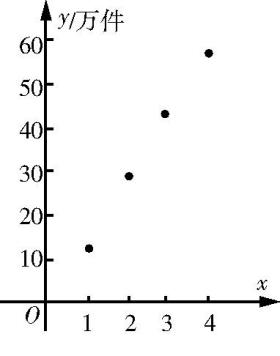
*P*(*x=*1,*y=*3)*=*×=,

*P*(*x=*2,*y=*3)*=*×=,

所以*P*(*x<y*)*===.*

即*x<y*的概率为*.*

7*.*(1)作出的散点图如图D 14*-*2*-*2所示:



图D 14*-*2*-*2

(2)根据散点图可知,可以用线性回归模型拟合*y*与*x*的关系*.*

(3)观察(1)中散点图可知各点大致分布在一条直线附近,列出表格:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *i* | *xi* | *yi* |  | *xiyi* |
| 1 | 1 | 12 | 1 | 12 |
| 2 | 2 | 28 | 4 | 56 |
| 3 | 3 | 42 | 9 | 126 |
| 4 | 4 | 56 | 16 | 224 |
| ∑ | 10 | 138 | 30 | 418 |

可得*=*,*=*,

所以*===*,

*=-* *=-×=-*2*.*

故回归直线方程为*=x-*2*.*

当*x=*5时,*=×*5*-*2*=*71*.*

故预测第5年的销售量为71万件*.*

8*.*(1)根据茎叶图,完成的2*×*2列联表如下,

|  |  |  |  |
| --- | --- | --- | --- |
|  | 喜食蔬菜 | 喜食肉类 | 总计 |
| 男同学 | 19 | 6 | 25 |
| 女同学 | 17 | 3 | 20 |
| 总计 | 36 | 9 | 45 |

计算得*K*2的观测值为*k==*0*.*562 5*<*2*.*706,

所以没有90*%*的把握认为“喜食蔬菜还是喜食肉类与性别有关”*.*

(2)因为从喜食肉类的同学中抽取的人数为9*×=*3,所以*ξ*的可能取值有0,1,2,3*.*

*P*(*ξ=*0)*==*,

*P*(*ξ=*1)*==*,

*P*(*ξ=*2)*==*,

*P*(*ξ=*3)*==.*

所以*ξ*的分布列为

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *ξ* | 0 | 1 | 2 | 3 |
| *P* |  |  |  |  |

所以*ξ*的数学期望*E*(*ξ*)*=*0*×+*1*×+*2*×+*3*×=*1*.*