## 第1讲　统计与统计案例

[**考情分析**]　高考对本讲内容的考查往往以实际问题为背景，考查随机抽样与用样本估计总体，线性回归方程的求解与运用，独立性检验问题．常与概率综合考查，中等难度．

考点一　统计图表

核心提炼

1．频率分布直方图中横坐标表示组距，纵坐标表示，频率＝组距×.

2．频率分布直方图中各小长方形的面积之和为1.

3．利用频率分布直方图求众数、中位数与平均数．

频率分布直方图中：

(1)最高的小长方形底边中点的横坐标即众数．

(2)中位数左边和右边的小长方形的面积和相等．

(3)平均数是频率分布直方图的“重心”，等于频率分布直方图中每个小长方形的面积乘以小长方形底边中点的横坐标之和．

例1　(1)(多选)(2020·新高考全国Ⅱ)我国新冠肺炎疫情防控进入常态化，各地有序推进复工复产，下面是某地连续11天复工复产指数折线图，下列说法正确的是(　　)

A．这11天复工指数和复产指数均逐日增加

B．这11天期间，复产指数增量大于复工指数的增量

C．第3天至第11天复工复产指数均增大都超过80%

D．第9天至第11天复产指数增量大于复工指数的增量

答案　CD

(2)学校为了了解新课程标准提升阅读要求对学生阅读兴趣的影响情况，随机抽取了100名学生进行调查．根据调查结果绘制学生周末阅读时间的频率分布直方图如图所示：

将阅读时间不低于30分钟的学生称为“阅读霸”，则下列结论正确的是(　　)

A．抽样表明，该校约有一半学生为阅读霸

B．该校只有50名学生不喜欢阅读

C．该校只有50名学生喜欢阅读

D．抽样表明，该校有50名学生为阅读霸

答案　A

解析　根据频率分布直方图可列下表：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 阅读时间(分钟) | [0,10) | [10,20) | [20,30) | [30,40) | [40,50) | [50,60] |
| 抽样人数(名) | 10 | 18 | 22 | 25 | 20 | 5 |

抽样100名学生中有50名为阅读霸，占一半，据此可判断该校约有一半学生为阅读霸．

易错提醒　(1)对于给出的统计图表，一定要结合问题背景理解图表意义，不能似懂非懂．

(2)频率分布直方图中纵坐标不要误以为频率．

跟踪演练1　(1)某旅游城市为向游客介绍本地的气温情况，绘制了一年中各月平均最高气温和平均最低气温的雷达图．图中*A*点表示十月的平均最高气温约为15 ℃，*B*点表示四月的平均最低气温约为5 ℃.下面叙述不正确的是(　　)

A．各月的平均最低气温都在0 ℃以上

B．七月的平均温差比一月的平均温差大

C．三月和十一月的平均最高气温基本相同

D．平均最高气温不低于20 ℃的月份有5个

答案　D

解析　由题中雷达图易知A，C正确．七月份平均最高气温超过20 ℃，平均最低气温约为13 ℃；一月份平均最高气温约为6 ℃，平均最低气温约为2 ℃，所以七月的平均温差比一月平均温差大，故B正确．由题图知平均最高气温不低于20 ℃的月份为六、七、八月，有3个．

(2)(多选)(2020·重庆模拟)新高考方案规定，普通高中学业水平考试分为合格性考试(合格考)和选择性考试(选择考)，其中“选择考”成绩将计入高考总成绩，即将学生考试时的原始卷面分数由高到低进行排序，评定为*A*，*B*，*C*，*D*，*E*五个等级，再转换为分数计入高考总成绩．某试点高中2020年参加“选择考”总人数是2018年参加“选择考”总人数的2倍，为了更好地分析该校学生“选择考”的水平情况，统计了该校2018年和2020年“选择考”成绩等级结果，得到如图所示的统计图．

针对该校“选择考”情况，2020年与2018年比较，下列说法正确的是(　　)

A．获得*A*等级的人数增加了

B．获得*B*等级的人数增加了1.5倍

C．获得*D*等级的人数减少了一半

D．获得*E*等级的人数相同

答案　AB

解析　设2018年参加“选择考”的总人数为*x*，则2020年参加“选择考”的总人数为2*x*，根据图表得出2018年和2020年各个等级的人数如表所示．

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 等级年份　　 | *A* | *B* | *C* | *D* | *E* |
| 2018 | 0.28*x* | 0.32*x* | 0.30*x* | 0.08*x* | 0.02*x* |
| 2020 | 0.48*x* | 0.8*x* | 0.56*x* | 0.12*x* | 0.04*x* |

由表可知，获得*A*等级的人数增加了，故A正确；获得*B*等级的人数增加了＝1.5倍，故B正确；获得*D*等级的人数增加了，故C错误；获得*E*等级的人数不相同，故D错误．

考点二　回归分析

核心提炼

在分析两个变量的相关关系时，可根据样本数据作出散点图来判断两个变量之间是否具有相关关系．若具有线性相关关系，则回归直线过样本点的中心(，)，并且可通过线性回归方程估计预报变量的值．

例2　(2020·全国Ⅱ)某沙漠地区经过治理，生态系统得到很大改善，野生动物数量有所增加．为调查该地区某种野生动物的数量，将其分成面积相近的200个地块，从这些地块中用简单随机抽样的方法抽取20个作为样区，调查得到样本数据(*xi*，*yi*)(*i*＝1,2，…，20)，其中*xi*和*yi*分别表示第*i*个样区的植物覆盖面积(单位：公顷)和这种野生动物的数量，并计算得

*i*＝60，*i*＝1 200，(*xi*－)2＝80，

(*yi*－)2＝9 000，(*xi*－)(*yi*－)＝800.

(1)求该地区这种野生动物数量的估计值(这种野生动物数量的估计值等于样区这种野生动物数量的平均数乘以地块数)；

(2)求样本(*xi*，*yi*)(*i*＝1,2，…，20)的相关系数(精确到0.01)；

(3)根据现有统计资料，各地块间植物覆盖面积差异很大．为提高样本的代表性以获得该地区这种野生动物数量更准确的估计，请给出一种你认为更合理的抽样方法，并说明理由．

附：相关系数*r*＝，≈1.414.

解　(1)由已知得样本平均数＝*i*＝60，

从而该地区这种野生动物数量的估计值为

60×200＝12 000.

(2)样本(*xi*，*yi*)(*i*＝1,2，…，20)的相关系数

*r*＝

＝＝≈0.94.

(3)分层抽样：根据植物覆盖面积的大小对地块分层，再对200个地块进行分层抽样．

理由如下：由(2)知各样区的这种野生动物数量与植物覆盖面积有很强的正相关关系．由于各地块间植物覆盖面积差异很大，从而各地块间这种野生动物数量差异也很大，采用分层抽样的方法较好地保持了样本结构与总体结构的一致性，提高了样本的代表性，从而可以获得该地区这种野生动物数量更准确的估计．

规律方法　样本数据的相关系数

*r*＝，

反映样本数据的相关程度，|*r*|越大，则相关性越强．

跟踪演练2　(1)已知某产品的销售额*y*与广告费用*x*之间的关系如下表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *x*(单位：万元) | 0 | 1 | 2 | 3 | 4 |
| *y*(单位：万元) | 10 | 15 | 20 | 30 | 35 |

若求得其线性回归方程为＝6.5*x*＋，则预计当广告费用为6万元时的销售额为(　　)

A．42万元 B．45万元 C．48万元 D．51万元

答案　C

解析　由题意，根据上表中的数据，

可得＝2，＝22，

即样本点的中心为(2,22)，

又线性回归方程＝6.5*x*＋经过样本点的中心，

所以22＝6.5×2＋，解得＝9，

所以＝6.5*x*＋9，当*x*＝6时，＝48.

(2)(2020·河北衡水中学月考)有一散点图如图所示，在5个(*x*，*y*)数据中去掉*D*(3,10)后，下列说法正确的是(　　)

A．残差平方和变小

B．相关系数*r*变小

C．相关指数*R*2变小

D．解释变量*x*与预报变量*y*的相关性变弱

答案　A

解析　∵从散点图可分析得出：

只有*D*点偏离直线远，去掉*D*点，解释变量*x*与预报变量*y*的线性相关性变强，

∴相关系数变大，相关指数变大，残差平方和变小，故选A.

考点三　独立性检验

核心提炼

假设有两个分类变量*X*和*Y*，它们的取值分别为{*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：

|  |  |  |  |
| --- | --- | --- | --- |
|  | *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* |

*K*2＝(其中*n*＝*a*＋*b*＋*c*＋*d*为样本容量)．

例3　(2020·新高考全国Ⅰ)为加强环境保护，治理空气污染，环境监测部门对某市空气质量进行调研，随机抽查了100天空气中的PM2.5和SO2浓度(单位：μg/m3)，得下表：

|  |  |  |  |
| --- | --- | --- | --- |
| SO2PM2.5　　 | [0,50] | (50,150] | (150,475] |
| [0,35] | 32 | 18 | 4 |
| (35,75] | 6 | 8 | 12 |
| (75,115] | 3 | 7 | 10 |

(1)估计事件“该市一天空气中PM2.5浓度不超过75，且SO2浓度不超过150”的概率；

(2)根据所给数据，完成下面的2×2列联表：

|  |  |  |
| --- | --- | --- |
| SO2PM2.5　　 | [0,150] | (150,475] |
| [0,75] |  |  |
| (75,115] |  |  |

(3)根据(2)中的列联表，判断是否有99%的把握认为该市一天空气中PM2.5浓度与SO2浓度有关？

附：*K*2＝，

|  |  |  |  |
| --- | --- | --- | --- |
| *P*(*K*2≥*k*0) | 0.050 | 0.010 | 0.001 |
| *k*0 | 3.841 | 6.635 | 10.828 |

解　(1)由表格可知，该市100天中，空气中的PM2.5浓度不超过75，且SO2浓度不超过150的天数为32＋6＋18＋8＝64，

所以该市一天中，空气中的PM2.5浓度不超过75，且SO2浓度不超过150的概率的估计值为＝0.64.

(2)由所给数据，可得2×2列联表：

|  |  |  |
| --- | --- | --- |
| SO2PM2.5　　 | [0,150] | (150,475] |
| [0,75] | 64 | 16 |
| (75,115] | 10 | 10 |

(3)根据2×2列联表中的数据可得

*K*2＝

＝

≈7.484>6.635，

故有99%的把握认为该市一天空气中PM2.5浓度与SO2浓度有关．

规律方法　独立性检验的关键

(1)根据2×2列联表准确计算*K*2，若2×2列联表没有列出来，要先列出此表．

(2)*K*2的观测值*k*越大，对应的假设*H*0成立的概率越小，*H*0不成立的概率越大．

跟踪演练3　(1)随着国家二胎政策的全面放开，为了调查一线城市和非一线城市的二胎生育意愿，某机构用简单随机抽样方法从不同地区调查了100位育龄妇女，结果如下表．

|  |  |  |  |
| --- | --- | --- | --- |
|  | 非一线城市 | 一线城市 | 总计 |
| 愿生 | 45 | 20 | 65 |
| 不愿生 | 13 | 22 | 35 |
| 总计 | 58 | 42 | 100 |

附表：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *P*(*K*2≥*k*0) | 0.100 | 0.050 | 0.010 | 0.001 |
| *k*0 | 2.706 | 3.841 | 6.635 | 10.828 |

由*K*2＝计算得，

*K*2的观测值*k*＝≈9.616，

参照附表，得到的正确结论是(　　)

A．在犯错误的概率不超过0.1%的前提下，认为“生育意愿与城市级别有关”

B．在犯错误的概率不超过0.1%的前提下，认为“生育意愿与城市级别无关”

C．有99%以上的把握认为“生育意愿与城市级别有关”

D．有99%以上的把握认为“生育意愿与城市级别无关”

答案　C

解析　由题意知，

*K*2的观测值*k*≈9.616>6.635，

∴有99%以上的把握认为“生育意愿与城市级别有关”．

(2)某校团委对“学生性别和喜欢某视频APP是否有关”做了一次调查，其中被调查的女生人数是男生人数的一半，男生喜欢某视频APP的人数占男生人数的，女生喜欢某视频APP的人数占女生人数的，若有95%的把握认为喜欢某视频APP和性别有关，则男生至少有(　　)

附：

|  |  |  |
| --- | --- | --- |
| *P*(*K*2≥*k*0) | 0.050 | 0.010 |
| *k*0 | 3.841 | 6.635 |

*K*2＝.

A．12人 B．6人 C．10人 D．18人

答案　A

解析　设被调查的男生人数为*x*，则被调查的女生人数为，则2×2列联表为

|  |  |  |  |
| --- | --- | --- | --- |
|  | 喜欢某视频APP | 不喜欢某视频APP | 总计 |
| 男生 |  |  | *x* |
| 女生 |  |  |  |
| 总计 |  | *x* |  |

若有95%的把握认为喜欢某视频APP和性别有关，则*K*2≥3.841，即*K*2＝＝≥3.841，则*x*≥≈10.243，又，，均为整数，所以男生至少有12人．

## 专题强化练

一、单项选择题

1．某人5次上班途中所花的时间(单位：分钟)分别为*x*，*y,*10,11,9，已知这组数据的平均数为10，方差为2，则|*x*－*y*|的值为(　　)

A．4 B．3 C．2 D．1

答案　A

解析　依题意有＝10，(*x*－10)2＋(*y*－10)2＋(10－10)2＋(11－10)2＋(9－10)2＝5×2，解得*x*＝8，*y*＝12或*x*＝12，*y*＝8，故|*x*－*y*|＝4.

2．(2019·全国Ⅲ)《西游记》《三国演义》《水浒传》和《红楼梦》是中国古典文学瑰宝，并称为中国古典小说四大名著．某中学为了解本校学生阅读四大名著的情况，随机调查了100位学生，其中阅读过《西游记》或《红楼梦》的学生共有90位，阅读过《红楼梦》的学生共有80位，阅读过《西游记》且阅读过《红楼梦》的学生共有60位，则该校阅读过《西游记》的学生人数与该校学生总数比值的估计值为(　　)

A．0.5 B．0.6 C．0.7 D．0.8

答案　C

解析　根据题意阅读过《红楼梦》《西游记》的人数用韦恩图表示如下：

所以该校阅读过《西游记》的学生人数与该校学生总数比值的估计值为＝0.7.

3．(2020·全国Ⅰ)某校一个课外学习小组为研究某作物种子的发芽率*y*和温度*x*(单位：℃)的关系，在20个不同的温度条件下进行种子发芽实验，由实验数据(*xi*，*yi*)(*i*＝1,2，…，20)得到下面的散点图：

由此散点图可以看出，在10℃至40℃之间，下面四个回归方程类型中最适宜作为发芽率*y*和温度*x*的回归方程类型的是(　　)

A．*y*＝*a*＋*bx* B．*y*＝*a*＋*bx*2

C．*y*＝*a*＋*b*e*x* D．*y*＝*a*＋*b*ln *x*

答案　D

解析　由散点图可以看出，点大致分布在对数型函数的图象附近．

4.某生产车间的甲、乙两位工人生产同一种零件，这种零件的标准尺寸为85 mm，现分别从他们生产的零件中各随机抽取8件进行检测，其尺寸(单位：mm)用茎叶图表示如图所示，则估计(　　)

A．甲、乙生产的零件尺寸的中位数相等

B．甲、乙生产的零件质量相当

C．甲生产的零件质量比乙生产的零件质量好

D．乙生产的零件质量比甲生产的零件质量好

答案　D

解析　甲生产的零件尺寸是93,89,88,85,84,82,79,78；乙生产的零件尺寸是90,88,86,85,85,84,84,78.故甲生产的零件尺寸的中位数是＝84.5，乙生产的零件尺寸的中位数是＝85，故A错误；根据数据分析，乙的数据较稳定，故乙生产的零件质量比甲生产的零件质量好，故B，C错误．

5．某校进行了一次创新作文大赛，共有100名同学参赛，经过评判，这100名参赛者的得分都在[40,90]之间，其得分的频率分布直方图如图所示，则下列结论错误的是(　　)

A．得分在[40,60)之间的共有40人

B．从这100名参赛者中随机选取1人，其得分在[60,80)之间的概率为0.5

C．估计得分的众数为55

D．这100名参赛者得分的中位数为65

答案　D

解析　根据频率和为1，计算(*a*＋0.035＋0.030＋0.020＋0.010)×10＝1，解得*a*＝0.005，

得分在[40,60)之间的频率是0.4，估计得分在[40,60)之间的有100×0.4＝40(人)，A正确；

得分在[60,80)之间的频率为0.5，可得从这100名参赛者中随机选取1人，得分在[60,80)之间的概率为0.5，B正确；

根据频率分布直方图知，最高的小矩形对应的底边中点为＝55，即估计众数为55，C正确；

根据频率分布直方图知，得分低于60分的直方图面积为(0.005＋0.035)×10＝0.4<0.5，而得分低于70分的直方图面积为(0.005＋0.035＋0.030)×10＝0.7>0.5，所以100名参赛者得分的中位数估计为60＋≈63.3，D错误．

二、多项选择题

6．(2020·烟台模拟)某大学为了解学生对学校食堂服务的满意度，随机调查了50名男生和50名女生，每位学生对食堂的服务给出满意或不满意的评价，得到如下表所示的列联表，经计算*K*2的观测值*k*≈4.762，则可以推断出(　　)

|  |  |  |
| --- | --- | --- |
|  | 满意 | 不满意 |
| 男 | 30 | 20 |
| 女 | 40 | 10 |

|  |  |  |  |
| --- | --- | --- | --- |
| *P*(*K*2≥*k*0) | 0.100 | 0.050 | 0.010 |
| *k*0 | 2.706 | 3.841 | 6.635 |

A.该学校男生对食堂服务满意的概率的估计值为

B．调研结果显示，该学校男生比女生对食堂服务更满意

C．有95%的把握认为男、女生对该食堂服务的评价有差异

D．有99%的把握认为男、女生对该食堂服务的评价有差异

答案　AC

解析　对于选项A，该学校男生对食堂服务满意的概率的估计值为＝，故A正确；对于选项B，该学校女生对食堂服务满意的概率的估计值为＝>，故B错误；因为*k*≈4.762>3.841，所以有95%的把握认为男、女生对该食堂服务的评价有差异，故C正确，D错误．

7．(2020·河北衡水中学月考)5G时代已经到来，5G的发展将直接带动包括运营、制造、服务在内的通信行业经济的快速增长，进而对GDP增长产生直接贡献，并通过产业间的关联效应和波及效应，间接带动国民经济各行业的发展，创造出更多的经济增加值．如图是某单位结合近几年数据，对今后几年的5G经济产出所做的预测．

结合上图，下列说法正确的是(　　)

A．5G的发展带动今后几年的总经济产出逐年增加

B．设备制造商的经济产出前期增长较快，后期放缓

C．信息服务商与运营商的经济产出的差距有逐步拉大的趋势

D．设备制造商在各年的经济产出中一直处于领先地位

答案　ABC

解析　由图易知A，B，C正确，而设备制造商的经济产出在2029年和2030年将低于信息服务商的经济产出，故D错误．

8．(2020·青岛模拟)某调查机构对全国互联网行业进行调查统计，得到整个互联网行业从业者年龄分布饼状图、90后从事互联网行业岗位分布条形图，则下列结论正确的是(　　)

注：90后指1990年及以后出生，80后指1980－1989年之间出生，80前指1979年及以前出生．

A．互联网行业从业人员中从事技术和运营岗位的人数占总人数的三成以上

B．互联网行业中从事技术岗位的人数超过总人数的20%

C．互联网行业中从事运营岗位的人数90后比80前多

D．互联网行业中从事技术岗位的人数90后比80后多

答案　ABC

解析　选项A，因为互联网行业从业人员中，“90后”占比为56%，其中从事技术和运营岗位的人数占的比分别为39.6%和17%，则“90后”从事技术和运营岗位的人数占总人数的56%×(39.6%＋17%)≈31.7%.“80前”和“80后”中必然也有从事技术和运营岗位的人，则总的占比一定超过三成，故选项A正确；

选项B，因为互联网行业从业人员中，“90后”占比为56%，其中从事技术岗位的人数占的比为39.6%，则“90后”从事技术岗位的人数占总人数的56%×39.6%≈22.2%.“80前”和“80后”中必然也有从事技术岗位的人，则总的占比一定超过20%，故选项B正确；

选项C，“90后”从事运营岗位的人数占总人数的比为56%×17%≈9.5%，大于“80前”的总人数所占比3%，故选项C正确；

选项D，“90后”从事技术岗位的人数占总人数的56%×39.6%≈22.2%，“80后”的总人数所占比为41%，条件中未给出“80后”从事技术岗位的占比，故不能判断，所以选项D错误．

三、填空题

9．某企业的一种商品的产量与成本数据如下表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 产量*x*(万件) | 14 | 16 | 18 | 20 | 22 |
| 成本*y*(元/件) | 12 | 10 | 7 | *a* | 3 |

若根据表中提供的数据，求出*y*关于*x*的线性回归方程为＝－1.15*x*＋28.1，则*a*的值为\_\_\_\_\_\_\_\_．

答案　5

解析　由题意知＝＝＝18，

＝＝，

又＝－1.15×18＋28.1＝7.4，

所以＝7.4，解得*a*＝5.

10．已知某地区中小学生人数和近视情况分别如图甲和图乙所示．为了了解该地区中小学生的近视形成原因，用分层抽样的方法抽取2%的学生进行调查，则样本容量为\_\_\_\_\_\_\_\_，抽取的高中生近视人数为\_\_\_\_\_\_\_\_．

答案　200　20

解析　由题图甲知，总人数为3 500＋2 000＋4 500＝10 000，所以样本容量为10 000×2%＝200，抽样比例为，所以高中生抽取的学生数为40，所以抽取的高中生近视人数为40×50%＝20.

11．下面的折线图给出的是甲、乙两只股票在某年中每月的收盘价格，已知股票甲的极差是6.88元，标准差为2.04元；股票乙的极差为27.47元，标准差为9.63元，根据这两只股票在这一年中的波动程度，给出下列结论：①股票甲在这一年中波动相对较小，表现的更加稳定；②购买股票乙风险高但可能获得高回报；③股票甲的走势相对平稳，股票乙的收盘价格波动较大；④两只股票在全年都处于上升趋势．其中正确的结论是\_\_\_\_\_\_\_\_．(填序号)

答案　①②③

解析　由题意可知，甲的标准差为2.04元，乙的标准差为9.63元，可知股票甲在这一年中波动相对较小，表现的更加稳定，故①正确；甲的极差是6.88元，乙的极差为27.47元，可知购买股票乙风险高但可能获得高回报，故②正确；通过折线图可知股票甲的走势相对平稳，股票乙的收盘价格波动较大，故③正确；通过折线图可得乙在6月到8月明显是下降趋势，故④错误．

12．设某大学的女生体重*y*(单位：kg)与身高*x*(单位：cm)具有线性相关关系，根据一组样本数据(*xi*，*yi*)(*i*＝1,2，…，*n*)，用最小二乘法建立的回归方程为＝0.85*x*－85.71，则下列结论中不正确的是\_\_\_\_\_\_\_\_．(填序号)

①*y*与*x*具有正的线性相关关系；

②回归直线过样本点的中心(，)；

③若该大学某女生身高增加1 cm，则其体重约增加0.85 kg；

④若该大学某女生身高为170 cm，则可断定其体重必为58.79 kg.

答案　④

解析　由于线性回归方程中*x*的系数为0.85，因此*y*与*x*具有正的线性相关关系，故①正确；因为回归直线必过样本点的中心(，)，所以②正确；由线性回归方程的意义知，某女生的身高增加1 cm，其体重约增加0.85 kg，故③正确；当某女生的身高为170 cm时，其体重估计值是58.79 kg，这不是确定值，因此④不正确．

四、解答题

13．某公司为了了解广告投入对销售收益的影响，在若干地区各投入3.5万元广告费用，并将各地的销售收益绘制成频率分布直方图，如图所示，由于工作人员操作失误，横轴的数据丢失，但可以确定横轴是从0开始计数的．

(1)根据频率分布直方图计算图中各小长方形的宽度；

(2)估计该公司投入3.5万元广告费用之后，对应销售收益的平均值(以各组的区间中点值代表该组的取值)；

(3)该公司按照类似的研究方法，测得另外一些数据，并整理得到下表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 广告投入*x*(单位：万元) | 1 | 2 | 3 | 4 | 5 |
| 销售收益*y*(单位：万元) | 2 | 3 | 2 |  | 7 |

表中的数据显示，*x*与*y*之间存在线性相关关系，请将(2)中的结果填入空白栏，并计算*y*关于*x*的线性回归方程．

附：＝，＝－.

解　(1)设各小长方形的宽度为*m*，由频率分布直方图中各小长方形面积总和为1，可知(0.08＋0.10＋0.14＋0.12＋0.04＋0.02)·*m*＝0.5*m*＝1，故*m*＝2.

(2)由(1)知，各分组依次是[0,2)，[2,4)，[4,6)，[6,8)，[8,10)，[10,12]，其中点值分别为1,3,5,7,9,11，对应的频率分别为0.16,0.20,0.28,0.24,0.08,0.04，

故可估计平均值为1×0.16＋3×0.20＋5×0.28＋7×0.24＋9×0.08＋11×0.04＝5.

(3)空白栏中填5.由题意可知，

＝＝3，＝＝3.8，

*xiyi*＝1×2＋2×3＋3×2＋4×5＋5×7＝69，

*x*＝12＋22＋32＋42＋52＝55.

根据公式可求得

＝＝＝＝1.2，

＝3.8－1.2×3＝0.2，

即线性回归方程为＝1.2*x*＋0.2.

14．(2020·全国Ⅲ)某学生兴趣小组随机调查了某市100天中每天的空气质量等级和当天到某公园锻炼的人次，整理数据得到下表(单位：天)：

|  |  |  |  |
| --- | --- | --- | --- |
| 锻炼人次空气质量等级　　 | [0,200] | (200,400] | (400,600] |
| 1(优) | 2 | 16 | 25 |
| 2(良) | 5 | 10 | 12 |
| 3(轻度污染) | 6 | 7 | 8 |
| 4(中度污染) | 7 | 2 | 0 |

(1)分别估计该市一天的空气质量等级为1,2,3,4的概率；

(2)求一天中到该公园锻炼的平均人次的估计值(同一组中的数据用该组区间的中点值为代表)；

(3)若某天的空气质量等级为1或2，则称这天“空气质量好”；若某天的空气质量等级为3或4，则称这天“空气质量不好”．根据所给数据，完成下面的2×2列联表，并根据列联表，判断是否有95%的把握认为一天中到该公园锻炼的人次与该市当天的空气质量有关？

|  |  |  |
| --- | --- | --- |
|  | 人次≤400 | 人次>400 |
| 空气质量好 |  |  |
| 空气质量不好 |  |  |

附：*K*2＝.

|  |  |  |  |
| --- | --- | --- | --- |
| *P*(*K*2≥*k*0) | 0.050 | 0.010 | 0.001 |
| *k*0 | 3.841 | 6.635 | 10.828 |

解　(1)由频数分布表可知，该市一天的空气质量等级为1的概率为＝0.43，

等级为2的概率为＝0.27，

等级为3的概率为＝0.21，

等级为4的概率为＝0.09.

(2)由频数分布表可知，一天中到该公园锻炼的人次的平均数为＝350.

(3)2×2列联表如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 人次≤400 | 人次>400 | 总计 |
| 空气质量好 | 33 | 37 | 70 |
| 空气质量不好 | 22 | 8 | 30 |
| 总计 | 55 | 45 | 100 |

*K*2＝≈5.820>3.841，

因此，有95%的把握认为一天中到该公园锻炼的人次与该市当天的空气质量有关．