

专题五

概率与统计

第1讲 统计与统计案例

考向预测



1. 抽样方法、样本的数字特征、统计图表、回归分析与独立性检验主要以选择题、填空题形式命题, 难度较小;
2. 注重知识的交汇渗透, 统计与概率, 回归分析与概率是近年命题的热点.

知识与技巧的梳理



1. 抽样方法

抽样方法包括简单随机抽样、系统抽样、分层抽样, 三种抽样方法都是等概率抽样, 体现了抽样的公平性, 但又各有其特点和适用范围.

2. 统计中的四个数据特征

(1) 众数: 在样本数据中, 出现次数最多的那个数据.

(2) 中位数: 样本数据中, 将数据按大小排列, 位于最中间的数据. 如果数据的个数为偶数, 就取中间两个数据的平均数作为中位数.

(3) 平均数: 样本数据的算术平均数, 即 $\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$.

(4) 方差与标准差.

$$s^2 = \frac{1}{n}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2],$$

$$s = \sqrt{\frac{1}{n}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]}.$$

3. 直方图的两个结论

(1) 小长方形的面积 = 组距 $\times \frac{\text{频率}}{\text{组距}}$ = 频率.

(2) 各小长方形的面积之和等于 1.

4. 回归分析与独立性检验

(1) 回归直线 $\hat{y} = \hat{b}x + \hat{a}$ 经过样本点的中心点 $(\bar{x}, \bar{y})$, 若 x 取某一个值代入回归直线方程 $\hat{y} = \hat{b}x + \hat{a}$ 中, 可求出 y 的估计值.

(2) 独立性检验

对于取值分别是 $\{x_1, x_2\}$ 和 $\{y_1, y_2\}$ 的分类变量 X 和 Y , 其样本频数列联表是:

	y_1	y_2	总计
--	-------	-------	----

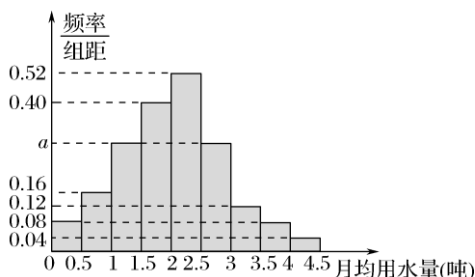
x_1	a	b	$a+b$
x_2	c	d	$c+d$
总计	$a+c$	$b+d$	n

则 $K^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ (其中 $n=a+b+c+d$ 为样本容量).

热点题型

热点一 用样本估计总体

【例 1】(2016 四川卷)我国是世界上严重缺水的国家,某市政府为了鼓励居民节约用水,计划调整居民生活用水收费方案,拟确定一个合理的月用水量标准 x (吨),一位居民的月用水量不超过 x 的部分按平价收费,超出 x 的部分按议价收费.为了了解居民用水情况,通过抽样,获得了某年 100 位居民每人的月均用水量(单位:吨),将数据按照 $[0, 0.5)$, $[0.5, 1)$, $\dots$, $[4, 4.5)$ 分成 9 组,制成了如图所示的频率分布直方图.



(1)求直方图中 a 的值;

(2)设该市有 30 万居民,估计全市居民中月均用水量不低于 3 吨的人数,并说明理由;

(3)若该市政府希望使 85%的居民每月的用水量不超过标准 x (吨),估计 x 的值,并说明理由.

解 (1)由频率分布直方图知,月均用水量在 $[0, 0.5)$ 中的频率为 $0.08 \times 0.5 = 0.04$.

同理,在 $[0.5, 1)$, $[1.5, 2)$, $[2, 2.5)$, $[3, 3.5)$, $[3.5, 4)$, $[4, 4.5)$ 中的频率分别为 0.08, 0.20, 0.26, 0.06, 0.04, 0.02.

由 $0.04 + 0.08 + 0.5 \times a + 0.20 + 0.26 + 0.5 \times a + 0.06 + 0.04 + 0.02 = 1$, 解得 $a = 0.30$.

(2)由(1)可知,100 位居民每人月均用水量不低于 3 吨的频率为 $0.06 + 0.04 + 0.02 = 0.12$.

由以上样本的频率,可以估计全市 30 万居民中月均用水量不低于 3 吨的人数为 $300\,000 \times 0.12 = 36\,000$.

(3)因为前 6 组的频率之和为 $0.04 + 0.08 + 0.15 + 0.20 + 0.26 + 0.15 = 0.88 > 0.85$,

而前 5 组的频率之和为 $0.04 + 0.08 + 0.15 + 0.20 + 0.26 = 0.73 < 0.85$.

所以 $2.5 \leq x < 3$.

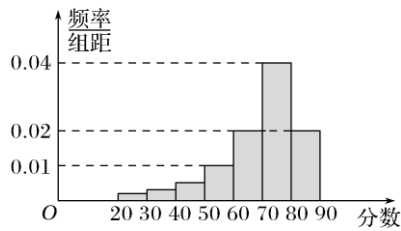
由 $0.3 \times (x - 2.5) = 0.85 - 0.73$, 解得 $x = 2.9$.

所以,估计月用水量标准为 2.9 吨时,85%的居民每月的用水量不超过标准.

探究提高 在本例中,抓住频率分布直方图各小长方形的面积之和为 1,这是求解的关键;本题易混淆频率分布条形图和频率分布直方图,误把频率分布直方图纵轴的几何意义当成频率,导致样本数据的频率求错.

【训练 1】(2017 北京卷)某大学艺术专业 400 名学生参加某次测评,根据男女学生人数比例,使用分层抽样的方法从中随机抽取了 100 名学生,记录他们的分数,将数据分成 7 组: $[20, 30)$, $[30, 40)$, $\dots$, $[80, 90]$, 并

整理得到如下频率分布直方图:



(1)从总体的 400 名学生中随机抽取一人, 估计其分数小于 70 的概率;

(2)已知样本中分数小于 40 的学生有 5 人, 试估计总体中分数在区间[40, 50)内的人数;

(3)已知样本中有一半男生的分数不小于 70, 且样本中分数不小于 70 的男女生人数相等.试估计总体中男生和女生人数的比例.

解 (1)根据频率分布直方图可知, 样本中分数不小于 70 的频率为 $(0.02+0.04) \times 10 = 0.6$,

所以样本中分数小于 70 的频率为 $1-0.6=0.4$.

所以从总体的 400 名学生中随机抽取一人, 其分数小于 70 的概率估计为 0.4.

(2)根据题意, 样本中分数不小于 50 的频率为

$$(0.01+0.02+0.04+0.02) \times 10 = 0.9,$$

分数在区间[40, 50)内的人数为 $100-100 \times 0.9-5=5$.

所以总体中分数在区间[40, 50)内的人数估计为 $400 \times \frac{5}{100} = 20$.

(3)由题意可知, 样本中分数不小于 70 的学生人数为

$$(0.02+0.04) \times 10 \times 100 = 60,$$

所以样本中分数不小于 70 的男生人数为 $60 \times \frac{1}{2} = 30$.

所以样本中的男生人数为 $30 \times 2 = 60$, 女生人数为 $100-60=40$, 男生和女生人数的比例为 $60:40=3:2$.

所以根据分层抽样原理, 总体中男生和女生人数的比例估计为 $3:2$.

热点二 回归分析与独立性检验

【例 2】 (1)某新闻媒体为了了解观众对央视《开门大吉》节目的喜爱与性别是否有关系, 随机调查了观看该节目的观众 110 名, 得到如下的列联表:

	女	男	总计
喜爱	40	20	60
不喜爱	20	30	50
总计	60	50	110

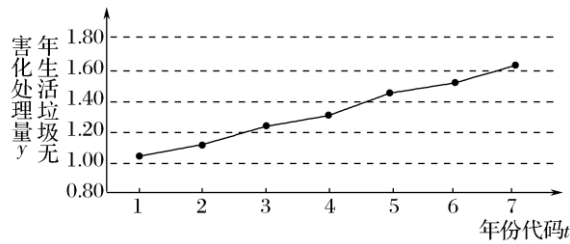
试根据样本估计总体的思想, 估计约有_____的把握认为“喜爱该节目与否和性别有关”.

参考附表:

$P(K^2 \geq k_0)$	0.050	0.010	0.001
k_0	3.841	6.635	10.828

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

(2)(2016·全国III卷)如图是我国 2008 年至 2014 年生活垃圾无害化处理量(单位: 亿吨)的折线图.



注: 年份代码 1~7 分别对应年份 2008~2014.

①由折线图看出, 可用线性回归模型拟合 y 与 t 的关系, 请用相关系数加以说明;

②建立 y 关于 t 的回归方程(系数精确到 0.01), 预测 2016 年我国生活垃圾无害化处理量.

附注:

参考数据: $\sum_{i=1}^7 y_i = 9.32$, $\sum_{i=1}^7 t_i y_i = 40.17$, $\sqrt{\sum_{i=1}^7 (y_i - \bar{y})^2} = 0.55$, $\sqrt{7} \approx 2.646$.

参考公式: 相关系数 $r = \frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (t_i - \bar{t})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$,

回归方程 $\hat{y} = \hat{a} + \hat{b}t$ 中斜率和截距的最小二乘估计公式分别为:

$$\hat{b} = \frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^n (t_i - \bar{t})^2}, \hat{a} = \bar{y} - \hat{b} \bar{t}.$$

(1)解析 分析列联表中数据, 可得 K^2 的一个观测值

$$k = \frac{110 \times (40 \times 30 - 20 \times 20)^2}{60 \times 50 \times 60 \times 50} \approx 7.822 > 6.635, \text{ 所以有 } 99\% \text{ 的把握认为 “喜爱《开门大吉》节目与否和性别有关”}.$$

答案 99%

(2)解 ①由折线图中的数据 and 附注中参考数据得

$$\bar{t} = 4, \sum_{i=1}^7 (t_i - \bar{t})^2 = 28, \sqrt{\sum_{i=1}^7 (y_i - \bar{y})^2} = 0.55.$$

$$\sum_{i=1}^7 (t_i - \bar{t})(y_i - \bar{y}) = \sum_{i=1}^7 t_i y_i - \bar{t} \sum_{i=1}^7 y_i = 40.17 - 4 \times 9.32 = 2.89, \text{ 所以 } r \approx \frac{2.89}{0.55 \times 2 \times 2.646} \approx 0.99.$$

因为 y 与 t 的相关系数近似为 0.99, 说明 y 与 t 的线性相关程度相当高, 从而可以用线性回归模型拟合 y 与 t 的关系.

$$\text{②由 } \bar{y} = \frac{9.32}{7} = 1.331 \text{ 及 (1) 得 } \hat{b} = \frac{\sum_{i=1}^7 (t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^7 (t_i - \bar{t})^2} = \frac{2.89}{28} \approx 0.103,$$

$$\hat{a} = \bar{y} - \hat{b} \bar{t} \approx 1.331 - 0.103 \times 4 \approx 0.92.$$

所以, y 关于 t 的回归方程为 $\hat{y} = 0.92 + 0.10t$.

将 2016 年对应的 $t=9$ 代入回归方程得: $\hat{y} = 0.92 + 0.10 \times 9 = 1.82$.

所以预测 2016 年我国生活垃圾无害化处理量约为 1.82 亿吨.

探究提高 1. 回归直线方程的关键: 正确理解计算 $\hat{b}$, $\hat{a}$ 的公式和准确地计算.

2. 独立性检验的关键: 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.

【训练 2】(1)(2017 贵阳调研)某医疗研究所为了检验某种血清能起到预防感冒的作用, 把 500 名使用血清的人与另外 500 名未使用血清的人一年中的感冒记录作比较, 利用 2×2 列联表计算得 K^2 的观测值 $k \approx 3.918$.

附表:

$P(K^2 \geq k_0)$	0.15	0.10	0.05	0.025	0.010	0.005	0.001
k_0	2.072	2.706	3.841	5.024	6.635	7.879	10.828

则作出“这种血清能起到预防感冒的作用”出错的可能性不超过()

A. 95%

B. 5%

C. 97.5%

D. 2.5%

(2)(2017 唐山一模)某市春节期间 7 家超市的广告费支出 x_i (万元)和销售额 y_i (万元)数据如下:

超市	A	B	C	D	E	F	G
广告费支出 x_i	1	2	4	6	11	13	19
销售额 y_i	19	32	40	44	52	53	54

①若用线性回归模型拟合 y 与 x 的关系, 求 y 关于 x 的线性回归方程;

②用对数回归模型拟合 y 与 x 的关系, 可得回归方程 $\hat{y} = 12 \ln x + 22$, 经计算得出线性回归模型和对数模型的 R^2 分别约为 0.75 和 0.97, 请用 R^2 说明选择哪个回归模型更合适, 并用此模型预测 A 超市广告费支出为 8 万元时的销售额.

参数数据及公式: $\bar{x} = 8$, $\bar{y} = 42$, $\sum_{i=1}^7 x_i y_i = 2794$, $\sum_{i=1}^7 x_i^2 = 708$,

$$\hat{b} = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}, \hat{a} = \bar{y} - \hat{b} \bar{x}, \ln 2 \approx 0.7.$$

(1)解析 $\because k \approx 3.918 > 3.841$, 且 $P(K^2 \geq k_0 = 3.841) = 0.05$, 根据独立性检验思想“这种血清能起到预防感冒的作用”出错的可能性不超过 5%.

答案 B

(2)解 ① $\because \bar{x} = 8$, $\bar{y} = 42$, $\sum_{i=1}^7 x_i y_i = 2794$, $\sum_{i=1}^7 x_i^2 = 708$.

$$\therefore \hat{b} = \frac{\sum_{i=1}^n x_i y_i - n \cdot \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{2794 - 7 \times 8 \times 42}{708 - 7 \times 8^2} = 1.7,$$

$$\text{因此 } \hat{a} = \bar{y} - \hat{b} \bar{x} = 42 - 1.7 \times 8 = 28.4.$$

所以, y 关于 x 的线性回归方程是 $\hat{y} = 1.7x + 28.4$.

② $\because 0.75 < 0.97$,

$\therefore$ 对数回归模型更合适.

当 $x = 8$ 时, $\hat{y} = 12 \ln 8 + 22 = 36 \ln 2 + 22 = 36 \times 0.7 + 22 = 47.2$ 万元.

∴广告费支出 8 万元时, 预测 A 超市销售额为 47.2 万元.

限时训练

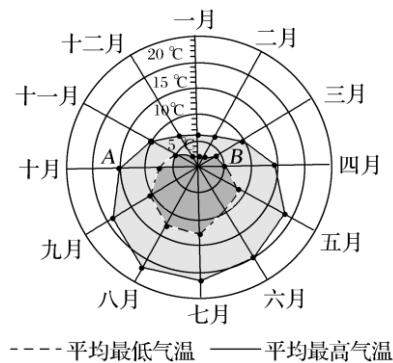
(45 分钟)

经典常规题

1.(2017 全国 I 卷)为评估一种农作物的种植效果,选了 n 块地作试验田.这 n 块地的亩产量(单位: kg)分别为 $x_1, x_2, \dots, x_n$, 下面给出的指标中可以用来评估这种农作物亩产量稳定程度的是()

- A. $x_1, x_2, \dots, x_n$ 的平均数
B. $x_1, x_2, \dots, x_n$ 的标准差
C. $x_1, x_2, \dots, x_n$ 的最大值
D. $x_1, x_2, \dots, x_n$ 的中位数

2.(2016 全国 III 卷)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°C , B 点表示四月的平均最低气温约为 5°C .下面叙述不正确的是()



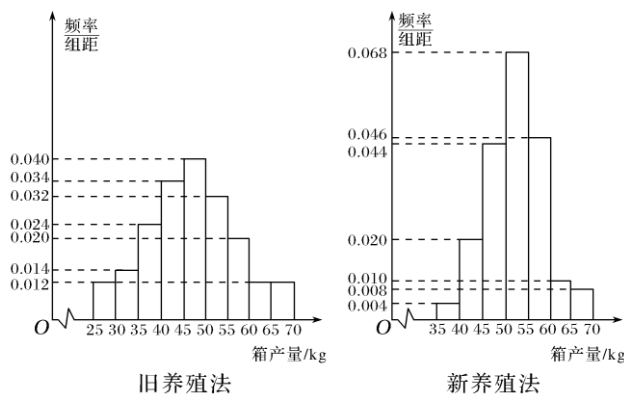
- A. 各月的平均最低气温都在 0°C 以上
B. 七月的平均温差比一月的平均温差大
C. 三月和十一月的平均最高气温基本相同
D. 平均最高气温高于 20°C 的月份有 5 个

3.(2017 山东卷)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 $\hat{y} = \hat{b}x + \hat{a}$.已知 $\sum_{i=1}^{10} x_i = 225$,

$\sum_{i=1}^{10} y_i = 1600$, $\hat{b} = 4$.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

4.(2017 全国 II 卷)淡水养殖场进行某水产品的新、旧网箱养殖方法的产量对比,收获时各随机抽取了 100 个网箱,测量各箱水产品的产量(单位: kg),其频率分布直方图如下:



(1)设两种养殖方法的箱产量相互独立,记 A 表示事件:旧养殖法的箱产量低于 50 kg,新养殖法的箱产量不低于 50 kg,估计 A 的概率;

(2)填写下面列联表,并根据列联表判断是否有 99%的把握认为箱产量与养殖方法有关:

	箱产量 < 50 kg	箱产量 $\geq$ 50 kg
旧养殖法		
新养殖法		

(3)根据箱产量的频率分布直方图,求新养殖法箱产量的中位数的估计值(精确到 0.01).

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

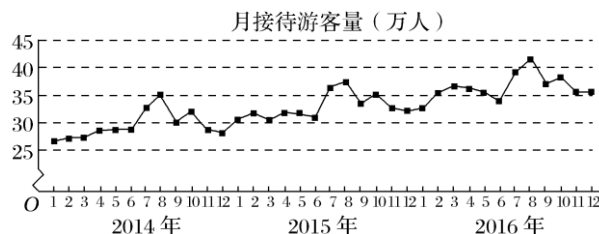
$P(K^2 \geq k)$	0.050	0.010	0.001
k	3.841	6.635	10.828

高频易错题

1.采用系统抽样方法从 960 人中抽取 32 人做问卷调查,为此将他们随机编号为 1, 2, ..., 960, 分组后在第一组采用简单随机抽样的方法抽到的号码为 9.抽到的 32 人中, 编号落入区间[1, 450]的人做问卷 A, 编号落入区间[451, 750]的人做问卷 B, 其余的人做问卷 C.则抽到的人中, 做问卷 B 的人数为()

A.7 B.9 C.10 D.15

2.(2017·全国III卷)某城市为了解游客人数的变化规律, 提高旅游服务质量, 收集并整理了 2014 年 1 月至 2016 年 12 月期间月接待游客量(单位: 万人)的数据, 绘制了下面的折线图.



根据该折线图, 下列结论错误的是()

- A.月接待游客量逐月增加
- B.年接待游客量逐年增加
- C.各年的月接待游客量高峰期大致在 7, 8 月
- D.各年 1 月至 6 月的月接待游客量相对于 7 月至 12 月, 波动性更小, 变化比较平稳

3.(2017·泉州模拟)某厂在生产甲产品的过程中, 产量 x (吨)与生产能耗 y (吨)的对应数据如表:

x	30	40	50	60
y	25	35	40	45

根据最小二乘法求得回归方程为 $\hat{y} = 0.65x + \hat{a}$, 当产量为 80 吨时, 预计需要生产能耗为_____吨.

4.(2017 全国III卷)某超市计划按月订购一种酸奶, 每天进货量相同, 进货成本每瓶 4 元, 售价每瓶 6 元, 未售出的酸奶降价处理, 以每瓶 2 元的价格当天全部处理完. 根据往年销售经验, 每天需求量与当天最高气温(单位: $^{\circ}\text{C}$)有关. 如果最高气温不低于 25, 需求量为 500 瓶; 如果最高气温位于区间 $[20, 25)$, 需求量为 300 瓶; 如果最高气温低于 20, 需求量为 200 瓶. 为了确定六月份的订购计划, 统计了前三年六月份各天的最高气温数据, 得下面的频数分布表:

最高气温	$[10, 15)$	$[15, 20)$	$[20, 25)$	$[25, 30)$	$[30, 35)$	$[35, 40)$
天数	2	16	36	25	7	4

以最高气温位于各区间的频率估计最高气温位于该区间的概率.

(1)估计六月份这种酸奶一天的需求量不超过 300 瓶的概率;

(2)设六月份一天销售这种酸奶的利润为 Y (单位: 元), 当六月份这种酸奶一天的进货量为 450 瓶时, 写出 Y 的所有可能值, 并估计 Y 大于零的概率.

精准预测题

1.(2017 汉中模拟)已知两个随机变量 x, y 之间的相关关系如表所示:

x	-4	-2	1	2	4
y	-5	-3	-1	-0.5	1

根据上述数据得到的回归方程为 $\hat{y} = \hat{b}x + \hat{a}$, 则大致可以判断()

- A. $\hat{a} > 0, \hat{b} > 0$ B. $\hat{a} > 0, \hat{b} < 0$ C. $\hat{a} < 0, \hat{b} > 0$ D. $\hat{a} < 0, \hat{b} < 0$

2.(2017 济南调研)2016 年济南地铁正式开工建设, 地铁时代的到来能否缓解济南的交通拥堵状况呢? 某社团进行社会调查, 得到的数据如下表:

	男性市民	女性市民
认为能缓解交通拥堵	48	30
认为不能缓解交通拥堵	12	20

则下列结论正确的是()

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

$P(K^2 \geq k)$	0.05	0.010	0.005	0.001
k	3.841	6.635	7.879	10.828

A.有 95%的把握认为“对能否缓解交通拥堵的认识与性别有关”

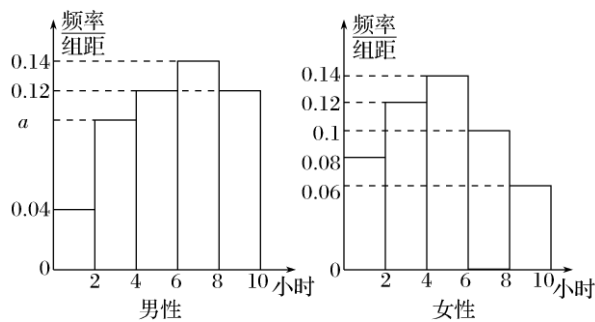
B.有 95%的把握认为“对能否缓解交通拥堵的认识与性别无关”

C.有 99%的把握认为“对能否缓解交通拥堵的认识与性别有关”

D.有 99%的把握认为“对能否缓解交通拥堵的认识与性别无关”

3.为了研究雾霾天气的治理情况,某课题组对部分城市进行空气质量调查,按地域特点把这些城市分成甲、乙、丙三组,已知三组城市的个数分别为 4, y , z , 依次构成等差数列,且 4, y , $z+4$ 成等比数列,若用分层抽样抽取 6 个城市,则乙组中应抽取的城市个数为_____.

4.(2017 赤峰二模)微信是腾讯公司推出的一种手机通讯软件,它支持发送语音短信、视频、图片和文字,一经推出便风靡全国,甚至涌现出一批在微信的朋友圈内销售商品的人(被称为微商).为了调查每天微信用户使用微信的时间,某经销化妆品的微商在一广场随机采访男性、女性用户各 50 名,将男性、女性使用微信的时间分成 5 组: $(0, 2]$, $(2, 4]$, $(4, 6]$, $(6, 8]$, $(8, 10]$ 分别加以统计,得到如图所示的频率分布直方图.



(1)根据女性频率分布直方图估计女性使用微信的平均时间;

(2)若每天玩微信超过 4 小时的用户列为“微信控”,否则称其为“非微信控”,请你根据已知条件完成 2×2 的列联表,并判断是否有 90%的把握认为“微信控”与“性别有关”?

参考答案

经典常规题

1. 【解题思路】刻画评估这种农作物亩产量稳定程度的指标是标准差.

【答案】B

2. 【解题思路】根据图示可得各月份的气温数据.

【答案】根据雷达图可知全年最低气温都在 0°C 以上, 故 A 正确; 一月平均最高气温是 6°C 左右, 平均最低气温 2°C 左右, 七月平均最高气温 22°C 左右, 平均最低气温 13°C 左右, 所以七月的平均温差比一月的平均温差大, B 正确; 三月和十一月的平均最高气温都是 10°C , 三月和十一月的平均最高气温基本相同, C 正确; 平均最高气温高于 20°C 的有七月和八月, D 项不正确. 故选 D.

3. 【解题思路】由回归直线方程过样本点中心可得 $\hat{a}$.

【答案】由已知得 $\bar{x} = 22.5$, $\bar{y} = 160$,

$\therefore$ 回归直线方程过样本点中心 $(\bar{x}, \bar{y})$, 且 $\hat{b} = 4$,

$\therefore 160 = 4 \times 22.5 + \hat{a}$, 解得 $\hat{a} = 70$.

$\therefore$ 回归直线方程为 $\hat{y} = 4x + 70$, 当 $x = 24$ 时, $\hat{y} = 166$. 故选 C.

4. 【解题思路】(1)以其频率代表概率; (2)完成 2×2 列联表, 并计算 K^2 ; (3)找出频率分布直方图中平方其面积的位置.

【答案】解 (1)记 B 表示事件“旧养殖法的箱产量低于 50 kg ”,

C 表示事件“新养殖法的箱产量不低于 50 kg ”.

由题意知, $P(A) = P(BC) = P(B)P(C)$.

旧养殖法的箱产量低于 50 kg 的频率为 $(0.012 + 0.014 + 0.024 + 0.034 + 0.040) \times 5 = 0.62$,

故 $P(B)$ 的估计值为 0.62 . 新养殖法的箱产量不低于 50 kg 的频率为 $(0.068 + 0.046 + 0.010 + 0.008) \times 5 = 0.66$,

故 $P(C)$ 的估计值为 0.66 .

因此, 事件 A 的概率估计值为 $0.62 \times 0.66 = 0.4092$.

(2)根据箱产量的频率分布直方图得列联表:

	箱产量 $< 50\text{ kg}$	箱产量 $\geq 50\text{ kg}$
旧养殖法	62	38
新养殖法	34	66

$$K^2 = \frac{200 \times (62 \times 66 - 34 \times 38)^2}{100 \times 100 \times 96 \times 104} \approx 15.705.$$

由于 $15.705 > 6.635$, 故有 99% 的把握认为箱产量与养殖方法有关.

(3)因为新养殖法的箱产量的频率分布直方图中, 箱产量低于 50 kg 的直方图面积为:

$$(0.004+0.020+0.044) \times 5 = 0.34 < 0.5,$$

箱产量低于 55 kg 的直方图面积为 $(0.004+0.020+0.044+0.068) \times 5 = 0.68 > 0.5$,

故新养殖法箱产量的中位数的估计值为 $50 + \frac{0.5-0.34}{0.068} \approx 52.35$ (kg).

高频易错题

1. 【解题思路】系统抽样也就是等距抽样, 找出间距与所分的段数, 再确定样本.

【答案】抽取号码的间隔为 $\frac{960}{32} = 30$, 从而区间 [451, 750] 包含的段数为 $\frac{750}{30} - \frac{450}{30} = 10$, 则编号落入区间 [451, 750] 的人数为 10 人, 即做问卷 B 的人数为 10. 故选 C.

2. 【解题思路】由题图可知, 2014 年 8 月到 9 月的月接待游客量在减少, 则 A 选项错误.

【答案】A

3. 【解题思路】由回归直线方程过样本点中心可得 $\hat{a}$.

【答案】由题意, $\bar{x} = 45$, $\bar{y} = 36.25$, 代入 $\hat{y} = 0.65x + \hat{a}$, 可得 $\hat{a} = 7$, $\therefore$ 当产量为 80 吨时, 预计需要生产能耗为 $0.65 \times 80 + 7 = 59$. 故填 59.

4. 【解题思路】(1) 以其频率代表概率; (2) 根据表格分别确定不同温度时所获利润, 再确定利润大于 0 时的频率 (也就是概率).

【答案】解 (1) 这种酸奶一天的需求量不超过 300 瓶, 当且仅当最高气温低于 25, 由表中数据可知, 最高气温低于 25 的频率为 $\frac{2+16+36}{90} = 0.6$.

所以这种酸奶一天的需求量不超过 300 瓶的概率的估计值为 0.6.

(2) 当这种酸奶一天的进货量为 450 瓶时,

若最高气温低于 20, 则 $Y = 200 \times 6 + (450 - 200) \times 2 - 450 \times 4 = -100$;

若最高气温位于区间 [20, 25), 则 $Y = 300 \times 6 + (450 - 300) \times 2 - 450 \times 4 = 300$;

若最高气温不低于 25, 则 $Y = 450 \times (6 - 4) = 900$,

所以, 利润 Y 的所有可能值为 -100, 300, 900.

Y 大于零当且仅当最高气温不低于 20, 由表格数据知, 最高气温不低于 20 的频率为 $\frac{36+25+7+4}{90} = 0.8$.

因此 Y 大于零的概率的估计值为 0.8.

精准预测题

1. 【解题思路】作出散点图, 画出回归直线直观判定 $\hat{b} > 0$, $\hat{a} < 0$.

【答案】C

2. 【解题思路】由 2×2 列联表, 可求 K^2 的观测值, $k = \frac{(48+30+12+20)(20 \times 48 - 12 \times 30)^2}{(48+30)(48+12)(12+20)(30+20)} \approx 5.288 > 3.841$.

由统计表 $P(K^2 \geq 3.841) = 0.05$, $\therefore$ 有 95% 的把握认为 “能否缓解交通拥堵的认识与性别有关”.

【答案】 A

3. 【解题思路】 根据等差数列和等比数列的定义列方程组解出 y, z .

【答案】 由题意可得 $\begin{cases} 2y = 4 + z, \\ y^2 = 4(z + 4), \end{cases}$ 即 $\begin{cases} y = 2 + \frac{z}{2}, \\ y^2 = 4z + 16, \end{cases}$ 解得 $z = 12$ 或 $z = -4$ (舍去), 故 $y = 8$.

所以甲、乙、丙三组城市的个数分别为 4, 8, 12.

因为一共要抽取 6 个城市, 所以抽样比为 $\frac{6}{4+8+12} = \frac{1}{4}$.

故乙组城市应抽取的个数为 $8 \times \frac{1}{4} = 2$. 故填 2.

4. 【解题思路】 (1) 取每组的中间值 x_n 代表这组, 平均数 $= x_1 p_1 + x_2 p_2 + \cdots + x_n p_n$; (2) 根据题意列出 2×2 列联表, 并计算 K^2 .

【答案】 解 (1) 女性平均使用微信的时间为: $0.16 \times 1 + 0.24 \times 3 + 0.28 \times 5 + 0.2 \times 7 + 0.12 \times 9 = 4.76$ (小时).

(2) 由已知得: $2(0.04 + a + 0.14 + 2 \times 0.12) = 1$, 解得 $a = 0.08$.

由题设条件得列联表

	微信控	非微信控	总计
男性	38	12	50
女性	30	20	50
总计	68	32	100

$$\therefore K^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{100(38 \times 20 - 30 \times 12)^2}{50 \times 50 \times 68 \times 32} \approx 2.941 > 2.706.$$

所以有 90% 的把握认为 “微信控” 与 “性别” 有关.