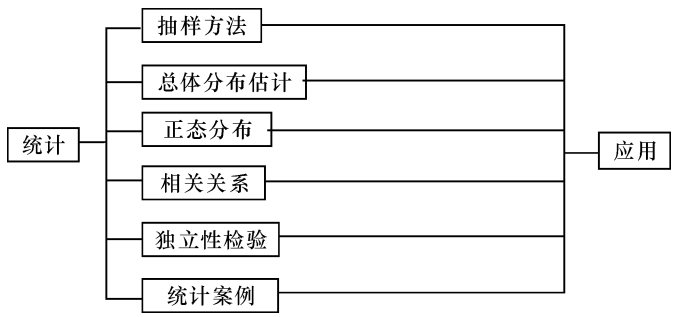
专题六　统计与概率

第 16 讲　统计与统计案例

知识网络



考情分析

年份	卷别	题号	考查内容	命题规律
2019	Ⅱ	5,13	统计的基本概念(平均数、中位数等)，用样本估计总体	如何抽取数据，如何从数据中提取信息，判断所得结论的可靠性，以及求随机变量分布的概率与特征. 一是根据实际条件选择抽样方法；二是提取统计图表中的有用信息(包括相关变量的关系判断)；三是求正态分布、二项分布(含两点分布)、超几何分布的概率情况和特征. 通过样本推断总体的过程，考查统计思想，重点是通过从已知数据和图表中提取有用的信息，以及个别事件的概率和整体随机变量的分布，或经过回归分析，解决实际问题.
	Ⅲ	3,17	用样本估计总体，直方图的有关计算	
2018	Ⅰ	3	统计图的识别与分析	
	Ⅱ	18	变量间的相关关系、利用回归直线方程进行估计	
	Ⅲ	18	茎叶图、中位数、列联表、独立性检验	
2017	Ⅰ	19	正态分布、数学期望、3σ 原则	
	Ⅱ	18	独立性检验、相互独立事件的概率、频率分布直方图	
	Ⅲ	3	折线图的识别	

备考建议

本节考点与实际问题联系紧密，复习中不能依赖记忆公式和简单的套用公式解题，应在充分认识统计方法特点的基础上，深刻理解回归分析和独立性检验的基本思想、方法及初步应用，提高阅读能力，找准数学模型，经历较为系统的数据处理的全过程，培养对数据的直观感觉，另外还要有意识地提高运算能力.

典例剖析

探究一　抽样方法

例 1 (1)现要完成下列 3 项抽样调查：

①从 10 盒酸奶中抽取 3 盒进行食品卫生检查；

②科技报告厅有 32 排，每排有 40 个座位，有一次报告会恰好坐满了听众，报告会结束后，为了听取意见，需要请 32 名听众进行座谈；

③高新中学共有 160 名教职工，其中一般教师

120名,行政人员16名,后勤人员24名,为了了解教职工对学校在校务公开方面的意见,拟抽取一个容量为20的样本.

较为合理的抽样方法是 ( )

A. ①简单随机抽样,②系统抽样,③分层抽样

B. ①简单随机抽样,②分层抽样,③系统抽样

C. ①系统抽样,②简单随机抽样,③分层抽样

D. ①分层抽样,②系统抽样,③简单随机抽样

(2)某工厂生产甲、乙、丙、丁四种不同型号的产品,产量分别为200,400,300,100件.为检验产品的质量,现用分层抽样的方法从以上所有的产品中抽取60件进行检验,则应从丙种型号的产品中抽取_____件.

(3)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

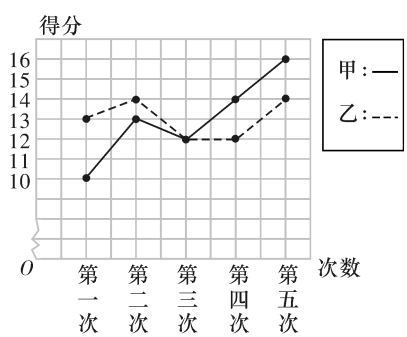
C. 10　　D. 15

【点评】(1)在系统抽样的过程中,要注意分段间隔,需要抽取几个个体,总体就需要分成几个组,则分段间隔即为$\frac{N}{n}$(n为样本容量),首先确定在第一组中抽取的个体的号码数,再从后面的每组中按规则抽取每个个体.

(2)分层抽样中要注意按比例抽取各层次的样本数据,样本容量与总体的个体数之比是分层抽样的比例常数,按这个比例可以确定各层应抽取的个体数与各层原有的人数,若各层应抽取的个体数不都是整数,则应当先剔除部分个体,调整总体个数.

探究二　用样本估计总体

例2 (1)甲、乙二人参加某体育项目训练,近期的五次测试成绩得分情况如图.

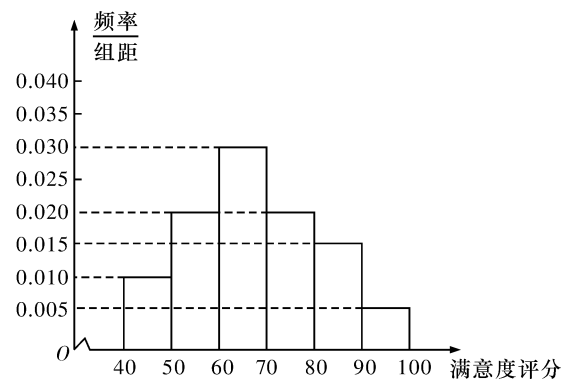


(Ⅰ)分别求出两人得分的平均数与方差;

(Ⅱ)根据上图和(Ⅰ)中算得的结果,对两人的训练成绩作出评价.

(2)某公司为了解用户对其产品的满意度,从A,B两地区分别随机调查了40名用户,根据用户对产品的满意度评分,得到A地区用户满意度评分的频率分布直方图和B地区用户满意度评分的频数分布表.

A地区用户满意度评分的频率分布直方图

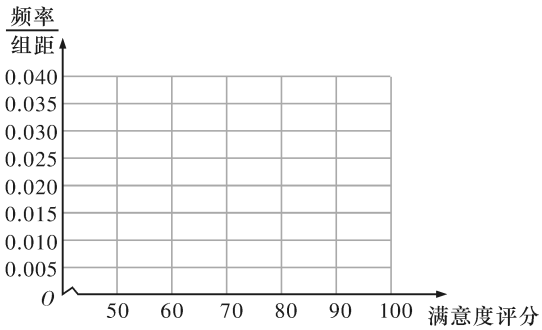


B地区用户满意度评分的频数分布表

满意度评分分组	[50,60)	[60,70)	[70,80)	[80,90)	[90,100]
频数	2	8	14	10	6

(Ⅰ)在下图中作出B地区用户满意度评分的频率分布直方图,并通过直方图比较两地区满意度评分的平均值及分散程度(不要求计算出具体值,给出结论即可).

B地区用户满意度评分的频率分布直方图



(Ⅱ)根据用户满意度评分,将用户的满意度分为三个等级:

满意度评分	低于70分	70分到89分	不低于90分
满意度等级	不满意	满意	非常满意

估计哪个地区用户的满意度等级为不满意的概率大?说明理由.

(3)某工厂36名工人的年龄数据如下表.

工人编号	年龄	工人编号	年龄	工人编号	年龄	工人编号	年龄
1	40	10	36	19	27	28	34
2	44	11	31	20	43	29	39
3	40	12	38	21	41	30	43
4	41	13	39	22	37	31	38
5	33	14	43	23	34	32	42
6	40	15	45	24	42	33	53
7	45	16	39	25	37	34	37
8	42	17	38	26	44	35	49
9	43	18	36	27	42	36	39

(Ⅰ)用系统抽样法从36名工人中抽取容量为9的样本,且在第一分段里用随机抽样法抽到的年龄数据为44,列出样本的年龄数据;

(Ⅱ)计算(Ⅰ)中样本数据的均值$\bar{x}$和方差s^2;

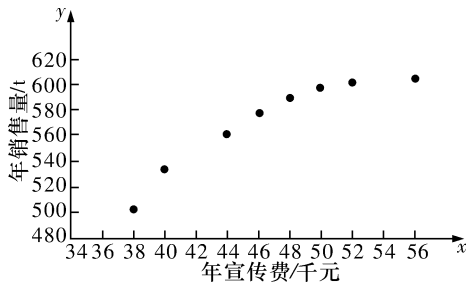
(Ⅲ)36名工人中年龄在$\bar{x}-s$与$\bar{x}+s$之间的有多少人?所占的百分比是多少(精确到0.01%)?

【点评】(1)在作茎叶图或读茎叶图时,首先要弄清楚"茎"和"叶"分别代表什么,根据茎叶图,我们可方便地求出数据的众数与中位数,大体上估计出两组数据平均数的大小与稳定性的高低.

(2)解决与频率分布直方图有关的问题时,应正确理解已知数据的含义,掌握图表中各个量的意义,通过图表对已知数据进行分类.

探究三　回归分析

例 3 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 x_i 和年销售量 y_i($i=1,2,\cdots,8$)数据作了初步处理,得到下面的散点图及一些统计量的值.



$\bar{x}$	$\bar{y}$	$\bar{w}$	$\sum_{i=1}^{8}(x_i-\bar{x})^2$	$\sum_{i=1}^{8}(w_i-\bar{w})^2$	$\sum_{i=1}^{8}(x_i-\bar{x})(y_i-\bar{y})$	$\sum_{i=1}^{8}(w_i-\bar{w})(y_i-\bar{y})$
46.6	563	6.8	289.8	1.6	1 469	108.8

表中 $w_i=\sqrt{x_i}$,$\bar{w}=\frac{1}{8}\sum_{i=1}^{8}w_i$.

(1)根据散点图判断,$y=a+bx$ 与 $y=c+d\sqrt{x}$ 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)

(2)根据(1)的判断结果及表中数据,建立 y 关于 x 的回归方程;

(3)已知这种产品的年利润 z 与 x、y 的关系为 $z=0.2y-x$.

根据(2)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 $(u_1,v_1),(u_2,v_2),\cdots\cdots,(u_n,v_n)$,其回归直线 $v=\hat{\alpha}+\hat{\beta}u$ 的斜率和截距的最小二乘估计分别为:

$$\hat{\beta}=\frac{\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{\sum_{i=1}^{n}(u_i-\bar{u})^2},\hat{\alpha}=\bar{v}-\beta\bar{u}.$$

【点评】已知变量的某个值去预测与其有线性相关关系的变量的值时,一般先求出回归直线方程 $\hat{y}=\hat{b}x+\hat{a}$,若 $\hat{a},\hat{b}$ 中有一个是已知的,常利用公式 $\hat{a}=\bar{y}-\hat{b}\bar{x}$ 求另一个量,再把 x 取值代入回归直线方程 $\hat{y}=\hat{b}x+\hat{a}$ 中,求出 $\hat{y}$ 的估计值.

探究四　独立性检验

例 4 某工厂为提高生产效率,开展技术创新活动,提出了完成某项生产任务的两种新的生产方式.为比较两种生产方式的效率,选取 40 名工人,将他们随机分成两组,每组 20 人,第一组工人用第一种生产方式,第二组工人用第二种生产方式.根据工人完成生产任务的工作时间(单位:min)绘制了如下茎叶图:

第一种生产方式		第二种生产方式
8	6	5 5 6 8 9
9 7 6 2	7	0 1 2 2 3 4 5 6 6 8
9 8 7 7 6 5 4 3 3 2	8	1 4 4 5
2 1 1 0 0	9	0

(1)根据茎叶图判断哪种生产方式的效率更高?并说明理由;

(2)求这 40 名工人完成生产任务所需时间的中位数 m,并将完成生产任务所需时间超过 m 和不超过 m 的工人数填入下面的列联表:

	超过 m	不超过 m
第一种生产方式		
第二种生产方式		

(3)根据(2)中的列联表,能否有 99%的把握认为两种生产方式的效率有差异?

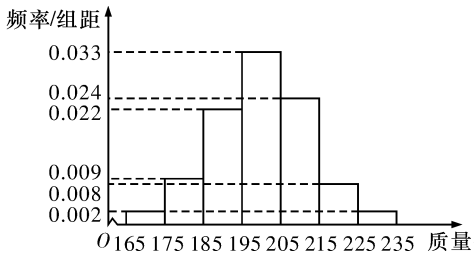
附:$K^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$.

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

【点评】独立性检验的具体步骤：第一步，根据题意确定临界值并作无关假设；第二步，找相关数据，列出 2×2 列联表；第三步，由公式 $K^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$（其中 $n=a+b+c+d$）计算出 K^2 的观测值；第四步，将 K^2 的观测值与临界值进行比较，进而作出推断.

探究五　正态分布

例 5 从某企业生产的某种产品中抽取 500 件，测量这些产品的一项质量指标值，由测量结果得如下频率分布直方图：



(1)求这 500 件产品质量指标值的样本平均数 $\bar{x}$ 和样本方差 s^2（同一组中的数据用该组区间的中点值作代表）；

(2)由直方图可以认为，这种产品的质量指标值 Z 服从正态分布 $N(\mu,\sigma^2)$，其中 μ 近似为样本平均数 $\bar{x}$，σ^2 近似为样本方差 s^2.

(ⅰ)利用该正态分布，求 $P(187.8<Z<212.2)$；

(ⅱ)某用户从该企业购买了 100 件这种产品，记 X 表示这 100 件产品中质量指标值位于区间(187.8,212.2)的产品件数. 利用(ⅰ)的结果，求 $E(X)$.

附：$\sqrt{150}\approx12.2$.

若 $Z\sim N(\mu,\sigma^2)$，

则 $P(\mu-\sigma<Z<\mu+\sigma)=0.6826$，

$P(\mu-2\sigma<Z<\mu+2\sigma)=0.9544$.

【点评】本题主要考查用样本估计总体、正态分布等基本知识和数据分析能力及运算求解能力.

规律总结

1. 进行系统抽样的关键是根据总体和样本的容量确定分段间隔，根据第一段确定编号. 如果总体不能被样本整除，即每段不能等分，应采用等可能剔除的方法剔除部分个体，以获得整数间隔.

2. 进行分层抽样时应注意以下几点：①分层抽样中分多少层、如何分层要视具体情况而定，总的原则是：层内样本的差异要小，两层之间的样本差异要大，且互不重叠；②为了保证每个个体等可能入样，所有层中每个个体被抽到的可能性要相同；③在每层抽样时，应采用简单随机抽样或系统抽样的方法进行抽样.

3. 进行线性回归分析时应注意的问题：

(1)正确理解计算 $\hat{b},\hat{a}$ 的公式和准确的计算是求线性回归方程的关键.

(2)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确定两个变量之间是否具有相关关系，若具有线性相关关系，则可通过线性回归方程估计和预测变量的值.

4. 独立性检验在实际应用中应注意的问题：

(1)独立性检验的关键是根据 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)复习独立性检验时，要根据实际问题，深刻体会独立性检验的思想.

5. 理解正态分布的概念与性质，它的密度曲线可以表示成一条钟形曲线，而且随着总体的均值与标准差的不同，曲线的形状产生相应的变化. 此外还要掌握好 $\mu-3\sigma$ 原则的应用.

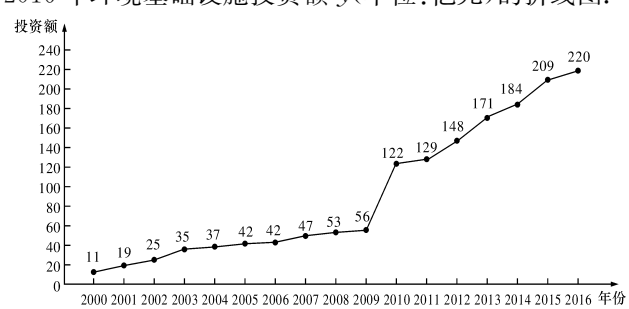
高考回眸

考题 1 [2019·全国卷Ⅱ]演讲比赛共有 9 位评委分别给出某选手的原始评分，评定该选手的成绩时，从 9 个原始评分中去掉 1 个最高分、1 个最低分，得到 7 个有效评分. 7 个有效评分与 9 个原始评分相比，不变的数字特征是　　(　　)

A. 中位数　　B. 平均数

C. 方差　　D. 极差

考题 2 [2018·全国卷Ⅱ]下图是某地区 2000 年至 2016 年环境基础设施投资额 y(单位：亿元)的折线图.

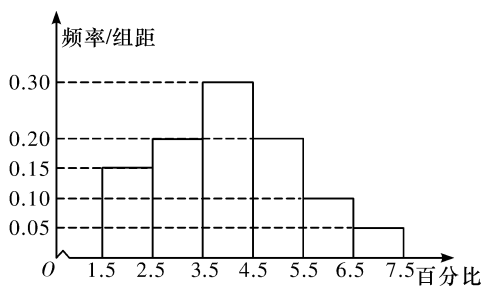


为了预测该地区2018年的环境基础设施投资额,建立了y与时间变量t的两个线性回归模型.根据2000年至2016年的数据(时间变量t的值依次为1,2,…,17)建立模型①:$\hat{y}=-30.4+13.5t$;根据2010年至2016年的数据(时间变量t的值依次为1,2,…,7)建立模型②:$\hat{y}=99+17.5t$.

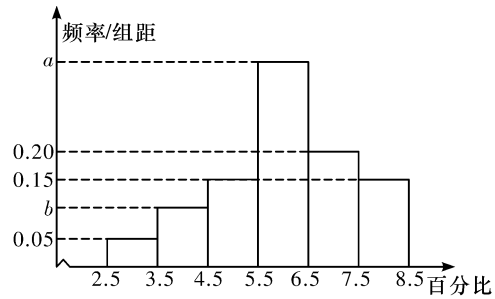
(1)分别利用这两个模型,求该地区2018年的环境基础设施投资额的预测值;

(2)你认为用哪个模型得到的预测值更可靠?并说明理由.

考题3 [2019·全国卷Ⅲ]为了解甲、乙两种离子在小鼠体内的残留程度,进行如下试验:将200只小鼠随机分成A,B两组,每组100只,其中A组小鼠给服甲离子溶液,B组小鼠给服乙离子溶液.每只小鼠给服的溶液体积相同、摩尔浓度相同.经过一段时间后用某种科学方法测算出残留在小鼠体内离子的百分比.根据试验数据分别得到如下直方图:



甲离子残留百分比直方图



乙离子残留百分比直方图

记C为事件:“乙离子残留在体内的百分比不低于5.5”,根据直方图得到$P(C)$的估计值为0.70.

(1)求乙离子残留百分比直方图中a,b的值;

(2)分别估计甲、乙离子残留百分比的平均值(同一组中的数据用该组区间的中点值为代表).

温馨提示:请完成考点限时训练(十六)p133