第 1 讲　统计与统计案例

高考定位　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，2018 年、2019 年和 2020 年在解答题中均有考查.

真 题 感 悟

1.(2019·全国Ⅱ卷)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

解析　中位数是将 9 个数据从小到大或从大到小排列后，处于中间位置的数据，因而去掉 1 个最高分和 1 个最低分，中位数是不变的，平均数、方差、极差均受影响.

答案　A

2.(2020·全国Ⅲ卷)在一组样本数据中，1，2，3，4 出现的频率分别为 p_1，p_2，p_3，p_4，且 $\sum_{i=1}^{4} p_i = 1$，则下面四种情形中，对应样本的标准差最大的一组是(　　)

A.$p_1 = p_4 = 0.1$，$p_2 = p_3 = 0.4$

B.$p_1 = p_4 = 0.4$，$p_2 = p_3 = 0.1$

C.$p_1 = p_4 = 0.2$，$p_2 = p_3 = 0.3$

D.$p_1 = p_4 = 0.3$，$p_2 = p_3 = 0.2$

解析　X 的可能取值为 1，2，3，4，四种情形的数学期望 $E(X) = 1\times p_1 + 2\times p_2 + 3\times p_3 + 4\times$

p_4 都为 2.5，方差 $D(X) = [1 - E(X)]_2 \times p_1 + [2 - E(X)]_2 \times p_2 + [3 - E(X)]_2 \times p_3 + [4 - E(X)]_2 \times p_4$，

标准差为$\sqrt{D(X)}$.

A 选项的方差 $D(X) = 0.65$；B 选项的方差 $D(X) = 1.85$；C 选项的方差 $D(X) = 1.05$；D 选项的

方差 $D(X) = 1.45$.

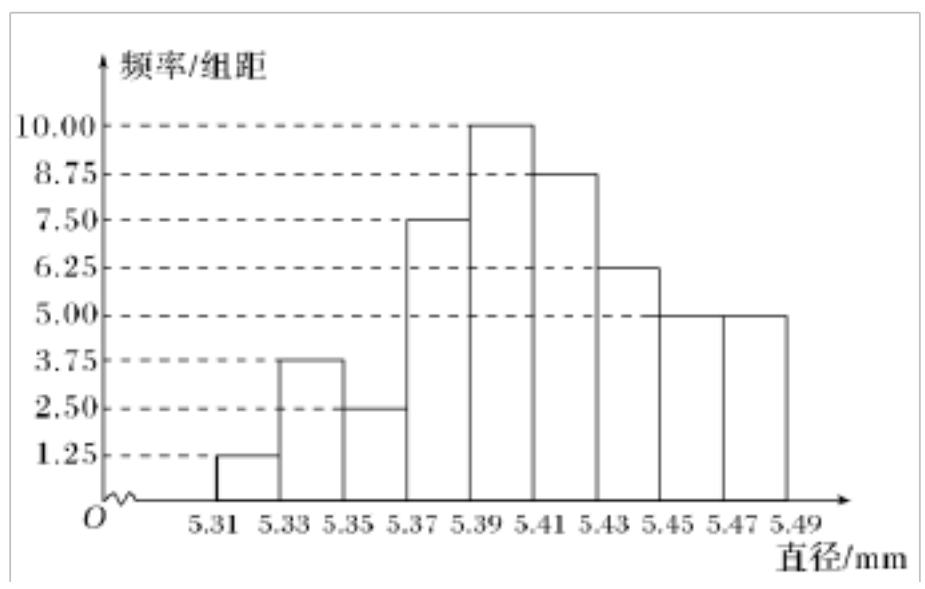
可知选项 B 的情形对应样本的标准差最大.故选 B.

答案　B

3.(2020·天津卷)从一批零件中抽取 80 个，测量其直径(单位：mm)，将所得数据分为 9 组：

[5.31，5.33)，[5.33，5.35)，…，[5.45，5.47)，[5.47，5.49]，并整理得到如下频率分布直

方图，则在被抽取的零件中，直径落在区间[5.43，5.47)内的个数为(　　)



A.10　　B.18　　C.20　　D.36

解析　因为直径落在区间[5.43，5.47)内的频率为 $0.02 \times (6.25 + 5.00) = 0.225$，所以个数为

$0.225 \times 80 = 18$.故选 B.

答案　B

4.(2020·全国Ⅱ卷)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$(x_i，y_i)(i = 1，2，…，20)$，其中 x_i 和

y_i 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 $\sum_{i=1}^{20} x_i =$

60，$\sum_{i=1}^{20} y_i = 1\,200$，$\sum_{i=1}^{20} (x_i - \bar{x})^2 = 80$，$\sum_{i=1}^{20} (y_i - \bar{y})^2 = 9\,000$，$\sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y}) = 800$.

(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$(x_i，y_i)(i = 1，2，…，20)$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 $r = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$，$\sqrt{2} \approx 1.414$.

解　(1)由已知得样本平均数 $\bar{y} = \dfrac{1}{20} \sum_{i=1}^{20} y_i = 60$，从而该地区这种野生动物数量的估计值为 $60 \times 200 = 12\,000$.

(2)样本$(x_i，y_i)(i = 1，2，…，20)$的相关系数

$$r = \frac{\sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{20} (x_i - \bar{x})^2 \sum_{i=1}^{20} (y_i - \bar{y})^2}} = \frac{800}{\sqrt{80 \times 9\,000}} = \frac{2\sqrt{2}}{3} \approx 0.94.$$

(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性.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.

考 点 整 合

1.抽样方法

抽样方法包括简单随机抽样、分层抽样，两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：在样本数据中，将数据按大小顺序排列，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$\bar{x}=\frac{1}{n}(x_1+x_2+\ldots+x_n)$.

(4)方差与标准差.

$s_2=\frac{1}{n}[(x_1-\bar{x})_2+(x_2-\bar{x})_2+\ldots+(x_n-\bar{x})_2]$，

$s=\sqrt{\frac{1}{n}[(x_1-\bar{x})_2+(x_2-\bar{x})_2+\ldots+(x_n-\bar{x})_2]}$.

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$(\bar{x}，\bar{y})$，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).

热点聚焦 | 分类突破 研热点 析角度

热点一 抽样方法

【例 1】 (1)总体由编号为 01，02，…，49，50 的 50 个个体组成，利用下面的随机数表选取 6 个个体 ,选取方法是从随机数表第 6 行的第 9 列和第 10 列数字开始从左到右依次选取两个数字，则选出的第 4 个个体的编号为(　　)

附：第 6 行至第 9 行的随机数表

2748　6198　7164　4148　7086　2888　8519　1620

7477　0111　1630　2404　2979　7991　9683　5125

3211　4919　7306　4916　7677　8733　9974　6732

2635　7900　3370　9160　1620　3882　7757　4950

A.3　　B.19　　C.38　　D.20

(2)(2020·百校大联考)在新冠肺炎疫情期间，大多数学生都进行网上上课.我校高一、高二、高三共有学生 1 800 名，为了了解同学们对“钉钉”授课软件的意见，计划采用分层抽样的方法从这 1 800 名学生中抽取一个容量为 72 的样本.若从高一、高二、高三抽取的人数恰好是从小到大排列的连续偶数，则我校高三年级的人数为(　　)

A.800　　B.750　　C.700　　D.650

解析　(1)由题意知，编号为 01～50 的个体才是需要的个体.由随机数表依次可得 41，48，28，19，16，20，……故第 4 个个体的编号为 19.故选 B.

(2)设从高三年级抽取的学生人数为 2x 人，则从高二、高一年级抽取的人数分别为 2x - 2，2x - 4.

由题意可得 $2x+(2x-2)+(2x-4)=72$，$\therefore x=13$.

设我校高三年级的学生人数为 N，且高三抽取 26 人，

由分层抽样，得$\frac{N}{1\,800}=\frac{26}{72}$，$\therefore N=650$(人).

答案　(1)B　(2)D

探究提高　解决此类题目的关键是深刻理解各种抽样方法的特点和适用范围.但无论哪种抽样方法，每一个个体被抽到的概率都是相等的，都等于样本容量与总体容量的比值.

【训练 1】 (1)总体由编号为 01，02，…，19，20 的 20 个个体组成.利用下面的随机数表选取 5 个个体，选取方法是从随机数表第 1 行第 6 列的数字开始，由左到右依次选取两个数字，则选出来的第 5 个个体的编号为_____.

附：第 1 行至第 2 行的随机数表

21 16 65 08　90 34 20 76　43 81 26 34　91 64 17 50　71 59 45 06

91 27 35 36　80 72 74 67　21 33 50 25　83 12 02 76　11 87 05 26

(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.

解析　(1)从随机数表的第 1 行第 6 列的数字开始，按规则得到的编号依次为 50，89，03，42，07，64，38，12，63，49，16，41，75，07，15，94，50，……其中编号在 01 至 20 之间的依次为 03，07，12，16，07，15，……按照编号重复的删除后一个的原则，可知选出来的第 5 个个体的编号为 15.

(2)因为样本容量 $n=60$，总体容量 $N=200+400+300+100=1\,000$，所以抽取比例为$\frac{n}{N}=\frac{60}{1000}=\frac{3}{50}$.

因此应从丙种型号的产品中抽取 $300\times\frac{3}{50}=18$(件).

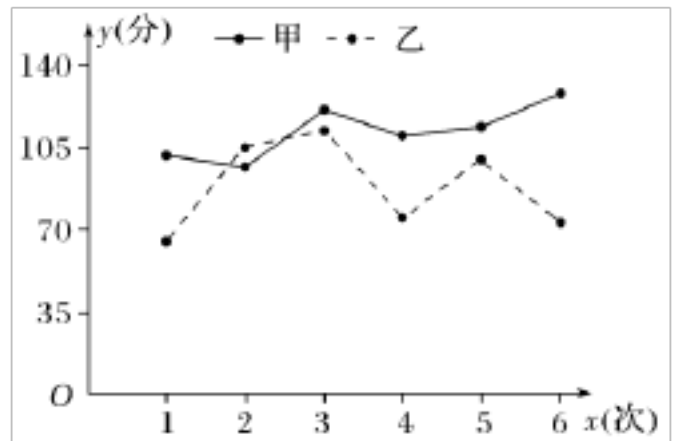
答案　(1)15　(2)18

热点二　用样本估计总体

角度 1　数字特征与统计图表的应用

【例 2】 (1)(2020·衡水检测)甲、乙两名同学高三以来 6 次数学模拟考试的成绩统计如下图，

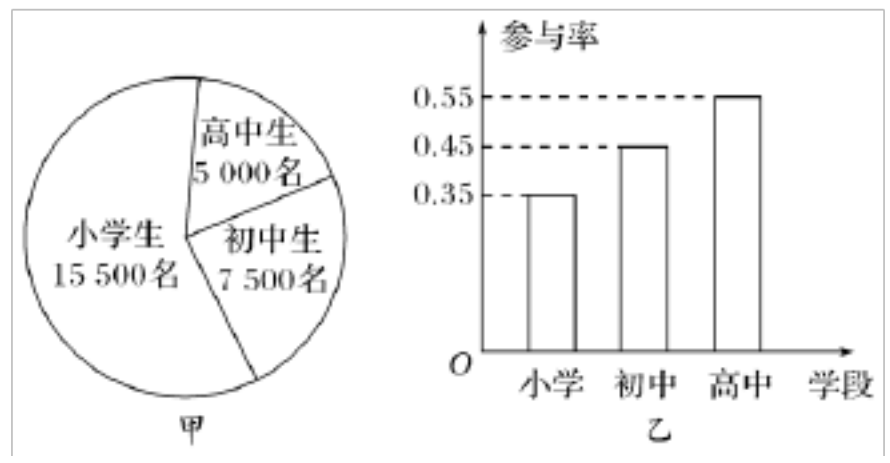
甲、乙两组数据的平均数分别为 $\bar{x}_{甲}$、$\bar{x}_{乙}$，标准差分别为 $s_{甲}$、$s_{乙}$，则(　　)



A.$\bar{x}_{甲}<\bar{x}_{乙}$，$s_{甲}<s_{乙}$　　B.$\bar{x}_{甲}<\bar{x}_{乙}$，$s_{甲}>s_{乙}$

C.$\bar{x}_{甲}>\bar{x}_{乙}$，$s_{甲}<s_{乙}$　　D.$\bar{x}_{甲}>\bar{x}_{乙}$，$s_{甲}>s_{乙}$

(2)2020 年初，我国突发新冠肺炎疫情，疫情期间中小学生“停课不停学”.已知某地区中小学生人数情况如甲图所示，各学段学生在疫情期间“家务劳动”的参与率如乙图所示.为了进一步了解该地区中小学生参与“家务劳动”的情况，现用分层抽样的方法抽取 4%的学生进行调查，则抽取的样本容量、抽取的高中生中参与“家务劳动”的人数分别为(　　)



A.2 750，200　　B.2 750，110

C.1 120，110　　D.1 120，200

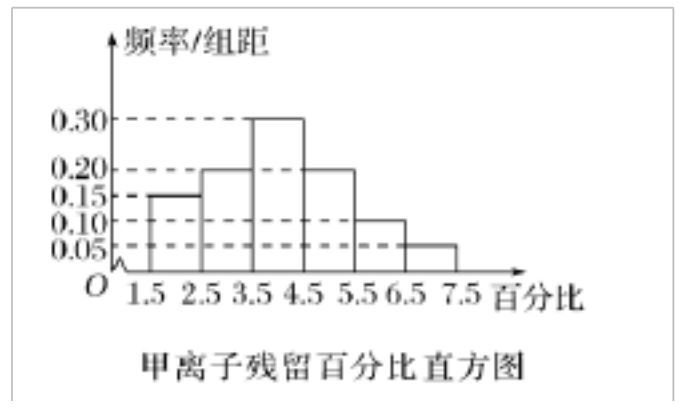
解析　(1)由统计图知，甲同学的总体成绩要好于乙同学的成绩，且乙同学的成绩波动较大，甲同学成绩较稳定.∴$\bar{x}_{甲} > \bar{x}_{乙}$，且 $s_{甲} < s_{乙}$.

(2)学生总数为 15 500 + 5 000 + 7 500 = 28 000 人，由于抽取 4%的学生进行调查，则抽取的样本容量为 28 000×4% = 1 120(人).故高中生应抽取的人数为 5 000×4% = 200(人)，而高中生中参与“家务劳动”的比率为 0.55，故高中生中参与“家务劳动”的人数为 200×0.55 = 110(人).

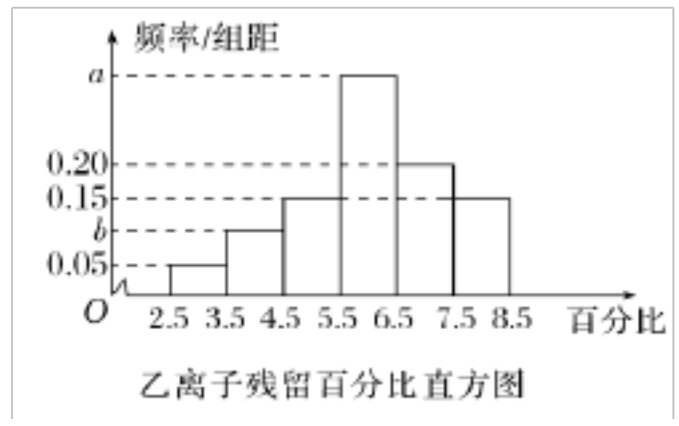
答案　(1)C　(2)C

角度 2　用样本的频率分布估计总体分布

【例 3】 (2019·全国Ⅲ卷)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).

解　(1)由已知得 0.70 = a + 0.20 + 0.15，

故 a = 0.35，

b = 1 - 0.05 - 0.15 - 0.70 = 0.10.

(2)甲离子残留百分比的平均值的估计值为

2×0.15 + 3×0.20 + 4×0.30 + 5×0.20 + 6×0.10 + 7×0.05 = 4.05.

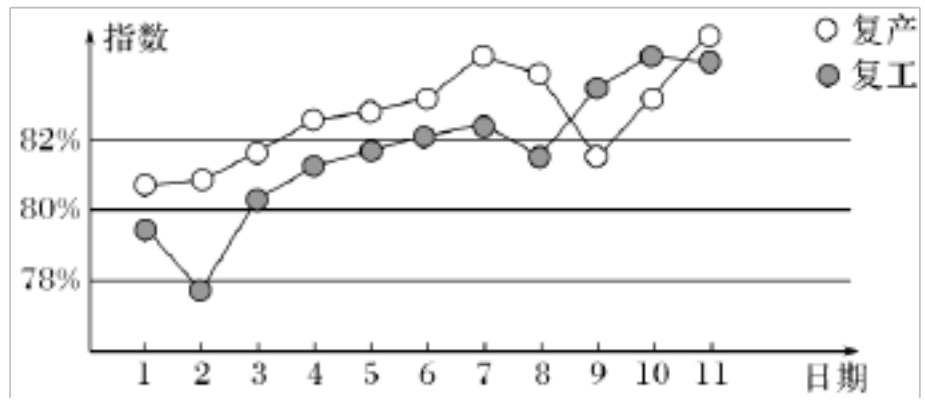
乙离子残留百分比的平均值的估计值为

3×0.05 + 4×0.10 + 5×0.15 + 6×0.35 + 7×0.20 + 8×0.15 = 6.00.

探究提高　1.平均数与方差都是重要的数字特征，是对数据的一种简明描述，它们所反映的情况有着重要的实际意义.平均数、中位数、众数描述数据的集中趋势，方差和标准差描述数据的波动大小.

2.在例 3 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 2】 (1)(2020·新高考海南卷)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

解析　由图可知，第 1 天到第 2 天复工指数减少，第 7 天到第 8 天复工指数减少，第 10 天到第 11 天复工指数减少，第 8 天到第 9 天复产指数减少，故 A 错误；由图可知，第一天的复产指数与复工指数的差大于第 11 天的复产指数与复工指数的差，所以这 11 天期间，复产指数增量小于复工指数的增量，故 B 错误；由图可知，第 3 天至第 11 天复工复产指数均超过 80%，故 C 正确；由图可知，第 9 天至第 11 天复产指数增量大于复工指数的增量，故 D 正确；故选 C、D.

答案　CD

(2)(2019·全国Ⅱ卷)某行业主管部门为了解本行业中小企业的生产情况，随机调查了 100 个企业，得到这些企业第一季度相对于前一年第一季度产值增长率 y 的频数分布表.

y 的分组	[- 0.20，0)	[0，0.20)	[0.20，0.40)	[0.40，0.60)	[0.60，0.80]
企业数	2	24	53	14	7

①分别估计这类企业中产值增长率不低于 40%的企业比例、产值负增长的企业比例；

②求这类企业产值增长率的平均数与标准差的估计值(同一组中的数据用该组区间的中点值为代表).(精确到 0.01)附：$\sqrt{74}\approx 8.602$.

解　①根据产值增长率频数分布表得，所调查的 100 个企业中产值增长率不低于 40%的企业频率为$\frac{14+7}{100}=0.21$.

产值负增长的企业频率为$\frac{2}{100}=0.02$.

所以用样本频率分布估计总体分布得这类企业中产值增长率不低于 40%的企业比例为 21%，产值负增长的企业比例为 2%.

②100 个企业的产值增长率平均数为

$$\bar{y}=\frac{1}{100}\times(-0.10\times2+0.10\times24+0.30\times53+0.50\times14+0.70\times7)=0.30,$$

$$s_2=\frac{1}{100}\sum_{i=1}^{5}n_i(y_i-\bar{y})_2=\frac{1}{100}\times[(-0.40)_2\times2+(-0.20)_2\times24+0_2\times53+0.20_2\times14+0.40_2\times7]$$

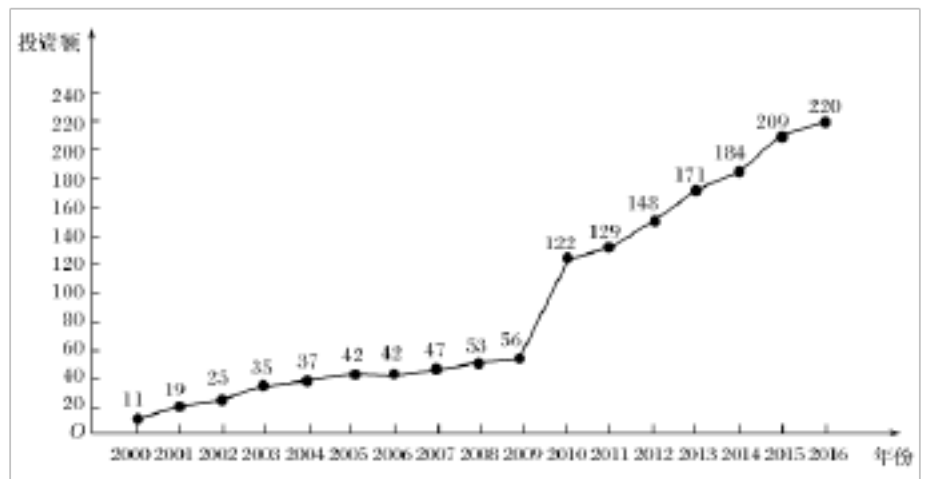
$$=0.029\ 6,$$

$$s=\sqrt{0.029\ 6}=0.02\times\sqrt{74}\approx0.17.$$

所以，这类企业产值增长率的平均数与标准差的估计值分别为 0.30，0.17.

热点三　回归分析在实际问题中的应用

【例 4】 如图是某地区 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.

解　(1)利用模型①，该地区 2018 年的环境基础设施投资额的预测值为$\hat{y}=-30.4+13.5\times19=226.1$(亿元).

利用模型②，该地区 2018 年的环境基础设施投资额的预测值为$\hat{y}=99+17.5\times9=256.5$(亿元).

(2)利用模型②得到的预测值更可靠.

理由如下：

(ⅰ)从折线图可以看出，2000 年至 2016 年的数据对应的点没有随机散布在直线 $y=-30.4+13.5t$ 上下，这说明利用 2000 年至 2016 年的数据建立的线性模型①不能很好地描述环境基础设施投资额的趋势.2010 年相对 2009 年的环境基础设施投资额有明显增加，2010 年至 2016 年的数据对应的点位于一条直线的附近，这说明从 2010 年开始环境基础设施投资额的变化规律呈线性增长趋势，利用 2010 年至 2016 年的数据建立的线性模型$\hat{y}=99+17.5t$ 可以较好地描述 2010 年以后的环境基础设施投资额的变化趋势，因此利用模型②得到的预测值更可靠.

(ⅱ)从计算结果看，相对于 2016 年的环境基础设施投资额 220 亿元，由模型①得到的预测值 226.1 亿元的增幅明显偏低，而利用模型②得到的预测值的增幅比较合理，说明利用模型②得到的预测值更可靠.

探究提高　1.求回归直线方程的关键及实际应用

(1)关键：正确理解$\hat{b}$，$\hat{a}$的计算公式和准确地计算.

(2)实际应用：在分析实际中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.

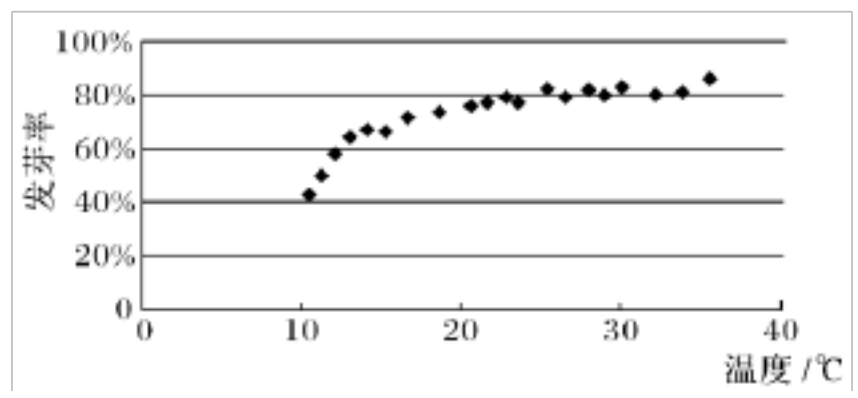
2.相关系数

(1)当 $r>0$ 时，表明两个变量正相关；当 $r<0$ 时，两个变量负相关.

(2)当$|r|>0.75$ 时，认为两个变量具有较强的线性相关关系.

【训练 3】 (1)(2020·全国Ⅰ卷)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$(x_i,y_i)(i=1,$

2，…，20)得到下面的散点图：



由此散点图，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.$y = a + be_x$　　D.$y = a + b\ln x$

(2)(2020·百强名校领军考试)已知变量 x，y 的关系可以用模型 $y = ce^{kx}$ 拟合，设 $z = \ln y$，其变换后得到一组数据如下：

x	16	17	18	19
z	50	34	41	31

由上表可得线性回归方程 $\hat{z} = -4x + \hat{a}$，则 $c = (\quad)$

A. - 4　　B.e^{-4}　　C.109　　D.e^{109}

解析　(1)由散点图可以看出，这些点大致分布在对数型函数的图象附近.故选 D.

(2)由数据表知 $\bar{x} = 17.5$，$\bar{z} = 39$.

∵样本点中心$(\bar{x}，\bar{z})$在回归直线上，

∴$\hat{a} = 39 + 4 \times 17.5 = 109$.

又 $z = \ln y = \ln(ce^{kx}) = kx + \ln c$，

∴$\ln c = \hat{a} = 109$，则 $c = e^{109}$.

答案　(1)D　(2)D

热点四　独立性检验