第 2 节 用样本估计总体（★★★）

内容提要

本节包含总体取值规律的估计，总体百分位数的估计，总体集中趋势估计，总体离散程度估计等内容，下面先把涉及到的一些基础知识进行梳理.

1．频率分布直方图：每个小矩形的面积表示数据落在该组的频率，各小矩形的面积之和为 1.

2．百分位数：一组数据的第 p 百分位数是这样一个值，它使得这组数据中至少有 $p\%$ 的数据小于或等于这个值，且至少有 $(100-p)\%$ 的数据大于或等于这个值，其计算步骤为：

①按从小到大排列原始数据；

②计算 $i=n\times p\%$，其中 n 为样本量；

③若 i 不是整数，而大于 i 的比邻整数为 j，则第 p 百分位数为第 j 项数据；若 i 是整数，则第 p 百分位数为第 i 项和第 $i+1$ 项数据的平均数.

3．平均数：可反映数据的集中趋势.

①一组数据 $x_1,x_2,\cdots,x_n$ 的平均数 $\overline{x}=\dfrac{x_1+x_2+\cdots+x_n}{n}$.

②由频率分布直方图估计样本平均数，常用每组区间中点值代表落在该区间的数据，若设各组区间中点为 $x_1,x_2,\cdots,x_n$，对应各组的频率为 $f_1,f_2,\cdots,f_n$，则可估计样本平均数 $\overline{x}=\sum\limits_{i=1}^{n}x_if_i$.

4．中位数：可反映数据的集中趋势.

①对于从小到大排列的一组数据，若数据个数为奇数，则中位数为最中间的一个数据；若数据个数为偶数，则中位数为中间两个数据的平均数.

②由频率分布直方图估计样本中位数，应在横轴上找到一个数，使其左右两侧频率各占 0.5.

5．众数：可反映数据的集中趋势.

①一组数据中出现次数最多的数据即为该组数据的众数，若有几个数据出现次数一样多，且都比其它数据多，则它们都是众数.

②由频率分布直方图估计样本众数，取最高的小矩形区间中点即可.

6．极差：一组数据的最大值与最小值之差，它可以一定程度反映数据的离散程度.

7．方差、标准差：刻画数据的离散程度. 方差、标准差越大，数据越分散，反之越集中.

①一组数据 $x_1,x_2,\cdots,x_n$ 的方差 $s^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i-\overline{x})^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i^2-\overline{x}^2$，方差公式的两种形式都需掌握，计算或证明时都可能用到. 方差的算数平方根 s 即为标准差. 若数据 $x_1,x_2,\cdots,x_n$ 有重复，设其不重复的值为 $y_1,y_2,\cdots,y_k$，对应的数据个数依次为 $f_1,f_2,\cdots,f_k$，则 $s^2=\dfrac{1}{n}\sum\limits_{i=1}^{k}f_i(y_i-\overline{x})^2$.

②由频率分布直方图估计样本方差，常用每组区间中点值代表落在该区间的数据，若设各组区间中点为 $x_1,x_2,\cdots,x_n$，对应各组的频率为 $f_1,f_2,\cdots,f_n$，则可估计样本方差 $s^2=\sum\limits_{i=1}^{n}(x_i-\overline{x})^2f_i$.

③方差、标准差的性质：设数据 $x_1,x_2,\cdots,x_n$ 的平均数为 $\overline{x}$，方差为 s^2，则数据 $y_i=ax_i+b(i=1,2,\cdots,n)$ 的平均数 $\overline{y}=a\overline{x}+b$，方差 $s_y^2=a^2s^2$（后面有证明），标准差 $s_y=as$.

证明：$s_y^2=\frac{1}{n}\sum_{i=1}^{n}y_i^2-\overline{y}^2=\frac{1}{n}\sum_{i=1}^{n}(ax_i+b)^2-(a\overline{x}+b)^2=\frac{1}{n}\sum_{i=1}^{n}(a^2x_i^2+2abx_i+b^2)-(a^2\overline{x}^2+2ab\overline{x}+b^2)$

$=\frac{1}{n}\sum_{i=1}^{n}a^2x_i^2+\frac{1}{n}\sum_{i=1}^{n}2abx_i+\frac{1}{n}\sum_{i=1}^{n}b^2-(a^2\overline{x}^2+2ab\overline{x}+b^2)$

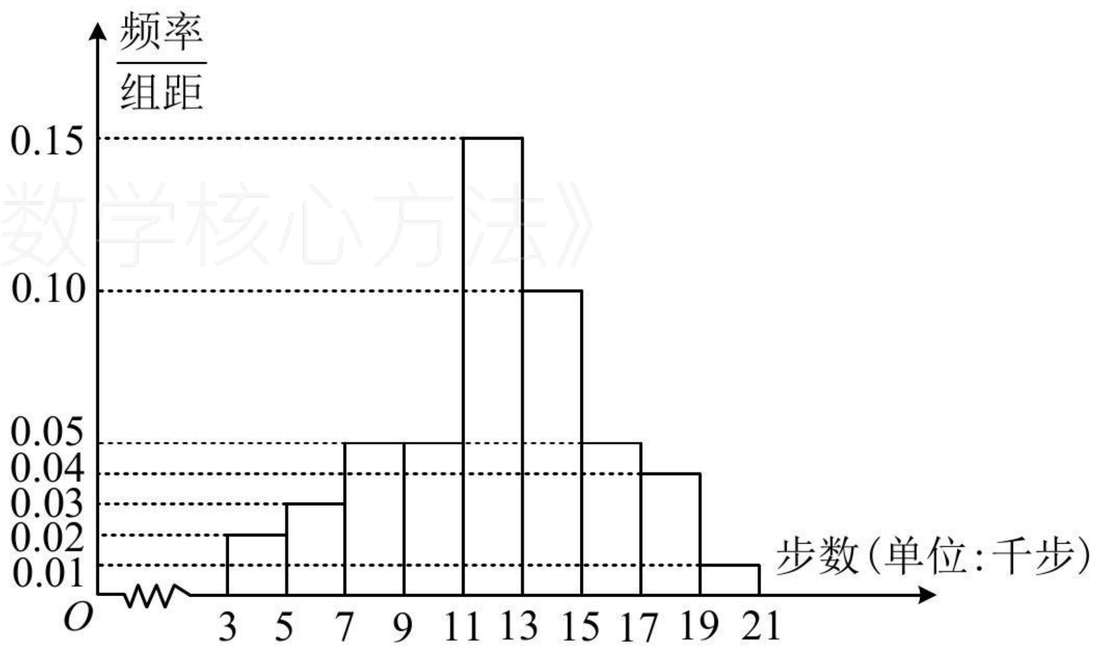
$=a^2\cdot\frac{1}{n}\sum_{i=1}^{n}x_i^2+2ab\overline{x}+\frac{1}{n}\cdot nb^2-(a^2\overline{x}^2+2ab\overline{x}+b^2)=a^2(\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2)=a^2s^2$.

典型例题

类型Ⅰ：由频率分布直方图计算频率、频数

【例 1】某地区工会利用“健步行 APP”开展健步走活动. 为了解会员的健步走情况，工会在某天从系统中抽取了 1000 名会员，统计了当天他们的步数（千步为单位），并将样本数据分为[3,5)，[5,7)，[7,9)，…，[17,19)，[19,21]九组，整理得到如图所示的频率分布直方图，则当天这 1000 名会员中步数少于 11 千步的人数为（　　）

（A）100　　（B）200　　（C）260　　（D）300

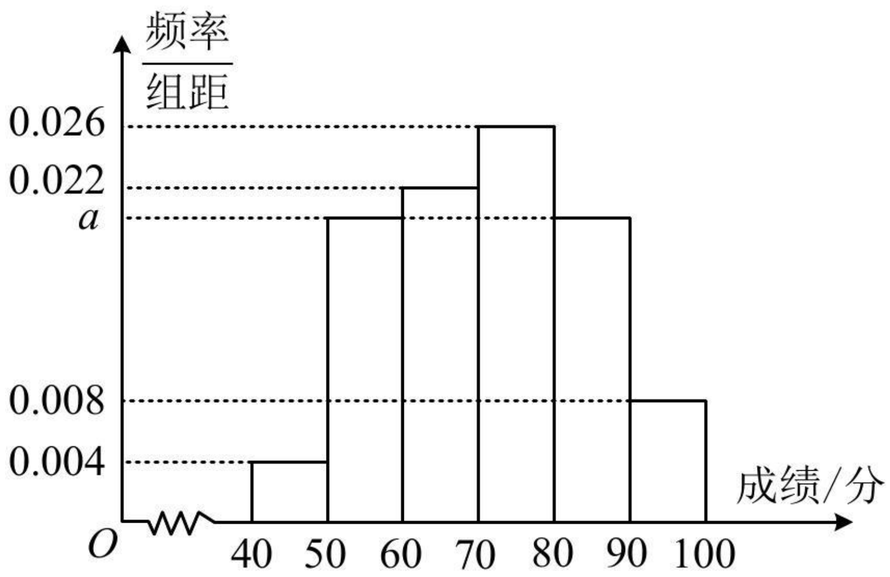


解析：要求人数，可由频率分布直方图求出对应的频率，再乘以样本量，

由图可知步数少于 11 千步的频率为 $2\times0.02+2\times0.03+2\times0.05+2\times0.05=0.3$，故所求人数为 $0.3\times1000=300$.

答案：D

【变式】如图是一学校期末考试中某班物理成绩的频率分布直方图，数据的分组依次为[40,50)，[50,60)，[60,70)，[70,80)，[80,90)，[90,100]，若成绩不低于 70 分的人数比成绩低于 70 分的人数多 4 人，则 $a=$_____；该班的学生人数为_____.



解析：先求 a，可用小矩形的面积和为 1 来建立方程，

由图可知，$10\times0.004+10\times a+10\times0.022+10\times0.026+10\times a+10\times0.008=1$，解得：$a=0.02$；

设该班的学生人数为 x，由图可知成绩不低于 70 分的频率为 $10\times0.026+10\times a+10\times0.008=0.54$，

所以成绩低于 70 分的频率为 $1-0.54=0.46$，由题意，$0.54x-0.46x=4$，解得：$x=50$.

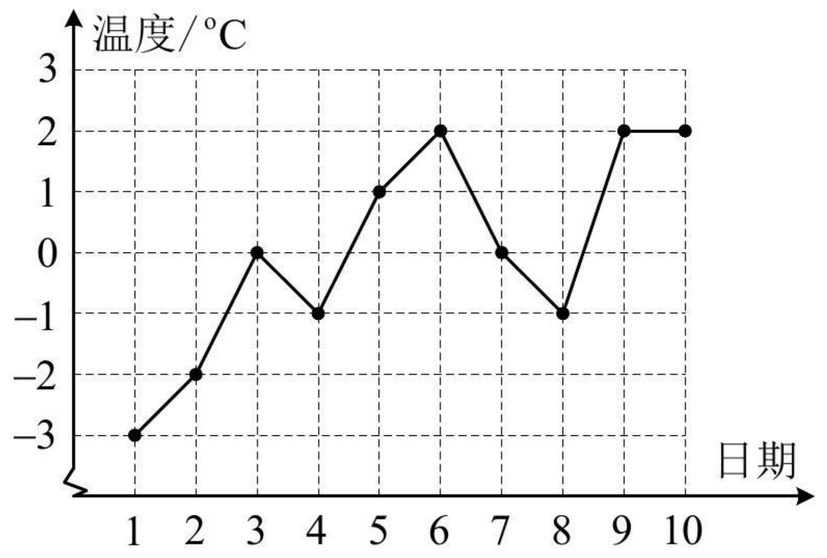
答案：0.02；50

【总结】在频率分布直方图中，需注意三点：①某个小矩形的面积代表数据落在该组的频率；②小矩形的面积之和为 1；③计算频数用频率×样本量.

类型Ⅱ：百分位数的计算

【例 2】如图是根据某市 1 月 1 日至 1 月 10 日的最低气温（单位：°C）的情况绘制的折线统计图，由图可知这 10 天的最低气温的第 40 百分位数是（　　）

（A）$2^{\circ}C$　　（B）$-1^{\circ}C$　　（C）$-0.5^{\circ}C$　　（D）$-2^{\circ}C$



解析：计算百分位数，先把数据按照由小到大的顺序排列，再计算 $i=n\times p\%$，看 i 是否为整数，

由折线图可知这 10 天的最低气温按照从低到高的顺序排列为 −3，−2，−1，−1，0，0，1，2，2，2，

而 $i=10\times40\%=4$，i 为整数，故取第 4 项和第 5 项数据的平均值作为第 40 百分位数，

所以这 10 天的最低气温的第 40 百分位数是 $\dfrac{-1+0}{2}=-0.5\ ^{\circ}C$.

答案：C

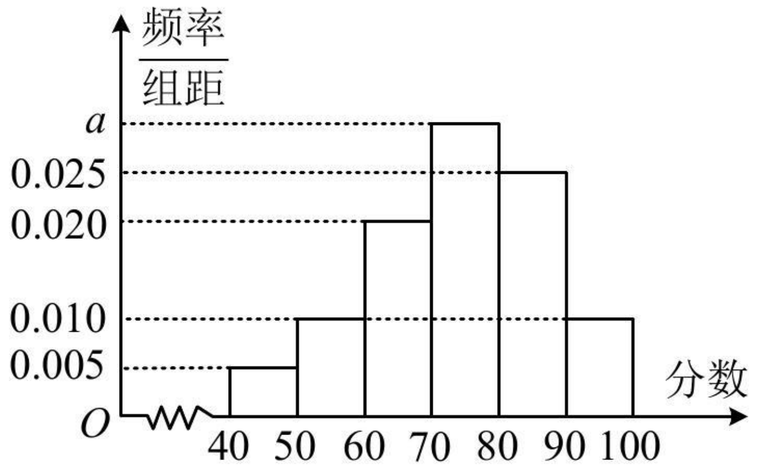
【变式 1】某学校组织班级知识竞赛，某班的 12 名学生的成绩（单位：分）分别是 58，67，73，74，76，82，82，87，90，92，93，98，则这 12 名学生成绩的第 70 百分位数是_____.

解析：所给数据已是从小到大排列，直接计算 $i=n\times p\%$，

由题意，$i=12\times70\%=8.4$，i 不是整数，所以取第 9 个数据为第 70 百分位数，故第 70 百分位数是 90.

答案：90

【变式 2】凯里市 2020 年被评为全国文明城市，为了巩文固卫，凯里一中某研究性学习小组举办了“文明城市”知识竞赛，从所有答卷中随机抽取 400 份试卷作为样本，将样本的成绩（满分 100 分，成绩均为不低于 40 分的整数）分成 6 段：$[40,50)$，$[50,60)$，…，$[90,100]$，得到如图所示的频率分布直方图，则 $a=$_____；由图可估计知识竞赛的第 80 百分位数为_____.



解析：由图可知，$10\times0.005+10\times0.01+10\times0.02+10\times a+10\times0.025+10\times0.01=1$，解得：$a=0.03$；

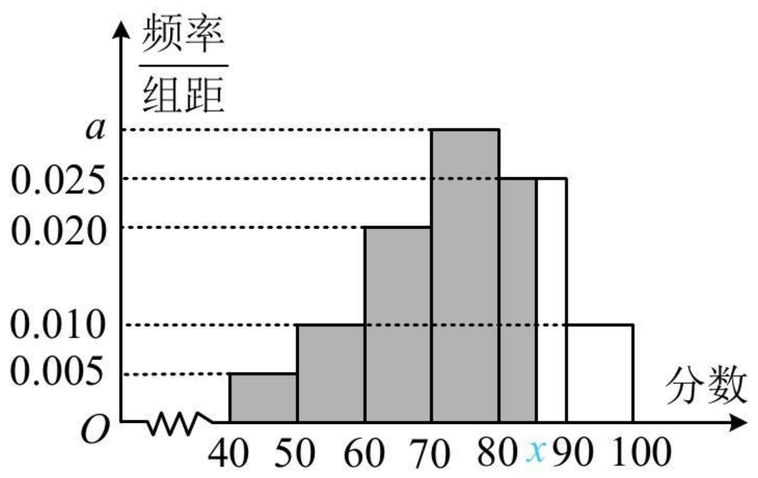
由频率分布直方图估计第 p 百分位数，就是要在频率分布直方图中找到左侧小矩形面积和为 $p\%$ 的位置，

由图可知前 4 组的面积之和 $1-10\times0.025-10\times0.01=0.65<0.8$，前 5 组的面积之和 $1-10\times0.01=0.9>0.8$，

所以第 80 百分位数必定在区间 $[80,90)$ 内，设为 x，

如图，由左侧阴影面积为 0.8 可得 $0.65+(x-80)\times0.025=0.8$，解得：$x=86$.

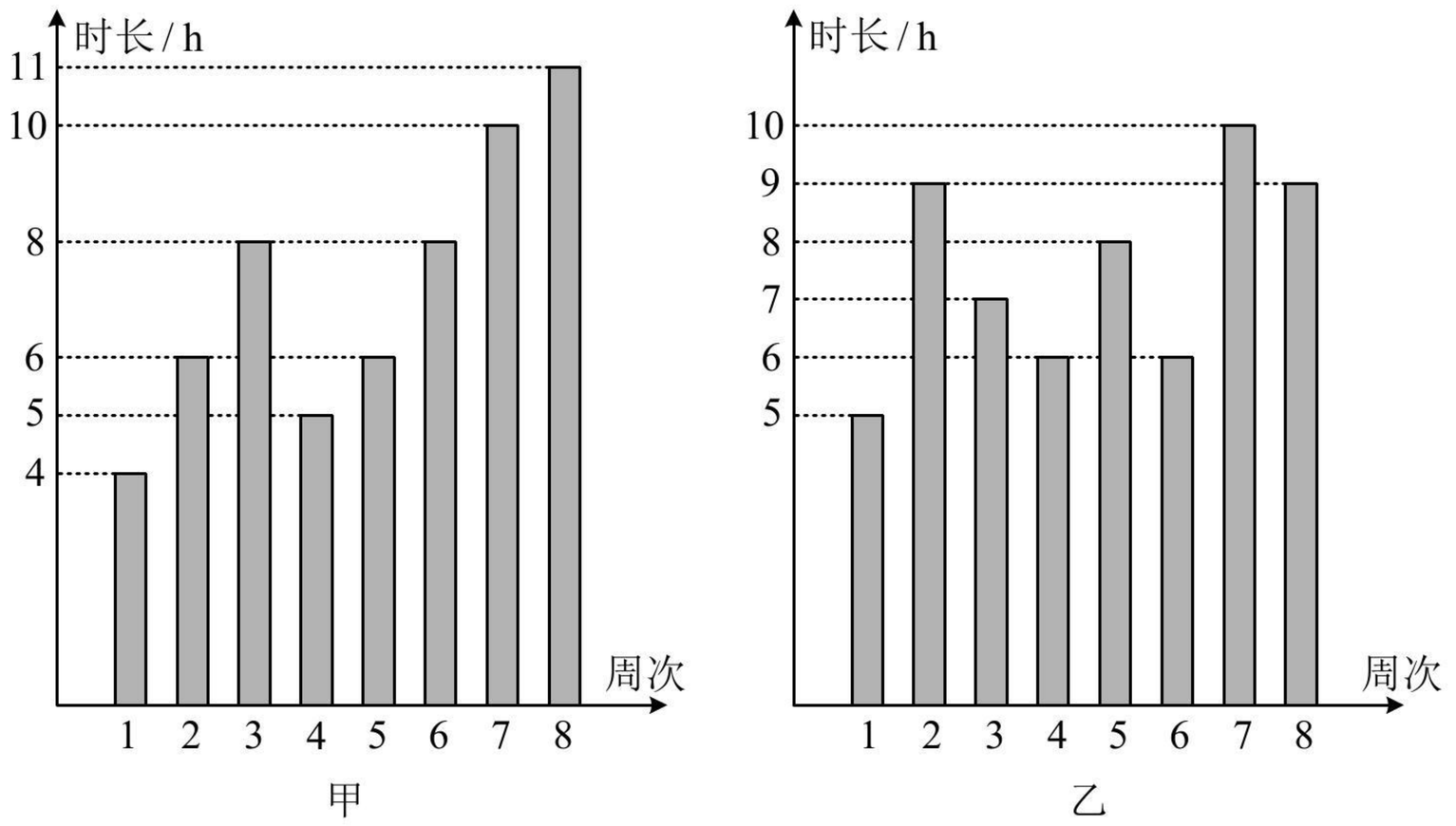
答案：0.03；86



【总结】①给出 n 个数据，让求第 p 百分位数，先将数据按从小到大的顺序排列，再计算 $i=n\times p\%$，若 i 为整数，则取第 i 个和第 $i+1$ 个数据的平均值作为第 p 百分位数，若 i 不是整数，而比 i 大的最小整数为 j，则取第 j 个数据为第 p 百分位数；②由频率分布直方图估计第 p 百分位数，只需在频率分布直方图中找到左侧小矩形面积和为 $p\%$ 的位置即可.

类型III：平均数、中位数、众数的计算

【例 3】甲、乙两位同学本学期前 8 周的各周课外阅读时长的条形统计图如图所示.



则下列结论正确的是（　　）

（A）甲同学周课外阅读时长的样本众数为 8

（B）甲同学周课外阅读时长的样本中位数是 5.5

（C）乙同学周课外阅读时长的样本平均数是 7.5

（D）乙同学周课外阅读时长大于 8 的概率的估计值大于 0.4

解析：A 项，由图可知甲的 8 个数据按从小到大的顺序依次为 4，5，6，6，8，8，10，11，其中 6 和 8 出现的次数都最多，所以众数为 6 和 8，故 A 项错误；

B 项，甲同学周课外阅读时长的样本中位数是 $\frac{6+8}{2}=7$，故 B 项错误；

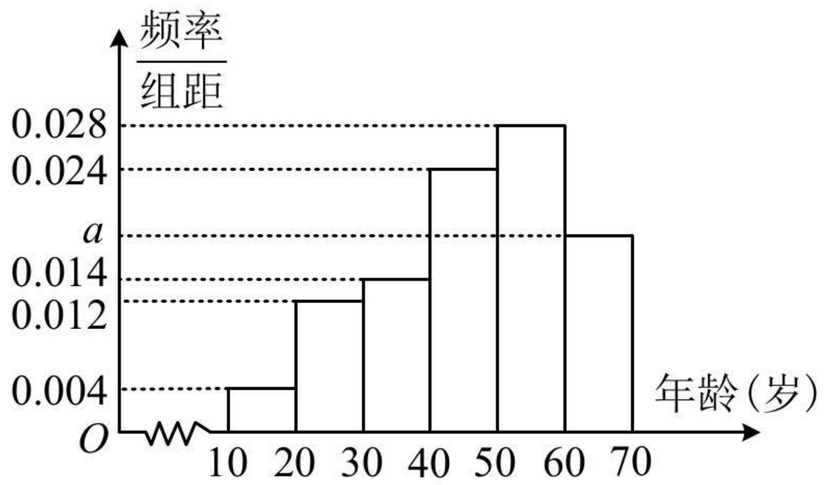
C 项，由图可知乙的 8 个数据按从小到大的顺序依次为 5，6，6，7，8，9，9，10，所以乙同学周课外阅读时长的样本平均数为 $\frac{5+6+6+7+8+9+9+10}{8}=7.5$，故 C 项正确；

D 项，乙有 3 个大于 8 的数据，所以乙同学周课外阅读时长大于 8 的概率的估计值为 $\frac{3}{8}<0.4$，故 D 项错误.

答案：C

【反思】一组数据的众数是出现次数最多的数据，若有多个数据出现次数一样多，且都比其它数据多，那它们都是众数.

【例 4】非物质文化遗产（简称“非遗”）是优秀传统文化的重要组成部分，是一个国家和民族历史文化成就的重要标志. 随着短视频这一新兴媒介形态的兴起，非遗传播获得广阔的平台，非遗文化迎来了发展的春天. 为研究非遗短视频受众的年龄结构，现从各短视频平台随机调查了 1000 名非遗短视频的粉丝，记录他们的年龄，将数据分成 6 组：$[10,20)$，$[20,30)$，…，$[50,60)$，$[60,70]$，并整理得到如下的频率分布直方图：



（1）求 a 的值；

（2）在频率分布直方图中，用每一个小矩形底边中点的横坐标作为该组粉丝年龄的平均数，估计非遗短视频粉丝年龄的平均数 m，若中位数的估计值为 n，比较 m 与 n 的大小关系.

解：（1）由图可知，$10\times0.004+10\times0.012+10\times0.014+10\times0.024+10\times0.028+10\times a=1$，解得：$a=0.018$.

（2）由图可知，从左至右 6 组的频率依次为 0.04，0.12，0.14，0.24，0.28，0.18，

所以频数依次为 40，120，140，240，280，180，

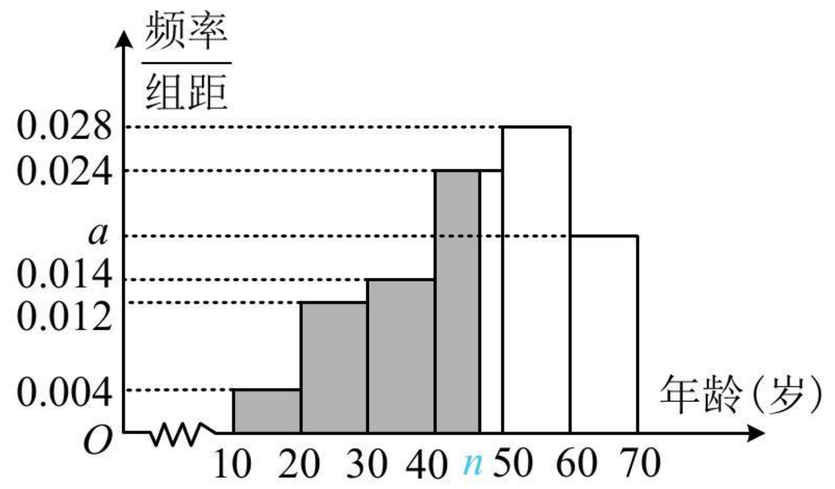
故 $m=\dfrac{15\times40+25\times120+35\times140+45\times240+55\times280+65\times180}{1000}=46.4$，

（再求中位数 n，只需找到使左侧频率和为 0.5 的位置即可）

前 3 组的频率和 $0.04+0.12+0.14=0.3<0.5$，前 4 组的频率和 $0.04+0.12+0.14+0.24=0.54>0.5$，

所以中位数 n 在 $[40,50)$ 这一组，（如图，阴影部分面积应为 0.5，前三组的面积已求出，第四组的左边阴影部分底边是 $n-40$，高为 0.024，其面积能用 n 表示，故可由此建立方程求 n）

从而 $0.3+(n-40)\times0.024=0.5$，解得：$n=\dfrac{145}{3}\approx48.3$，故 $m<n$.



【反思】由频率分布直方图估计样本数据的中位数，就是在频率分布直方图的横轴上找一个数，使不超过该数的频率和为 0.5，求解时应先判断中位数位于哪一组，再列方程计算.

类型Ⅳ：方差、标准差的计算与分析

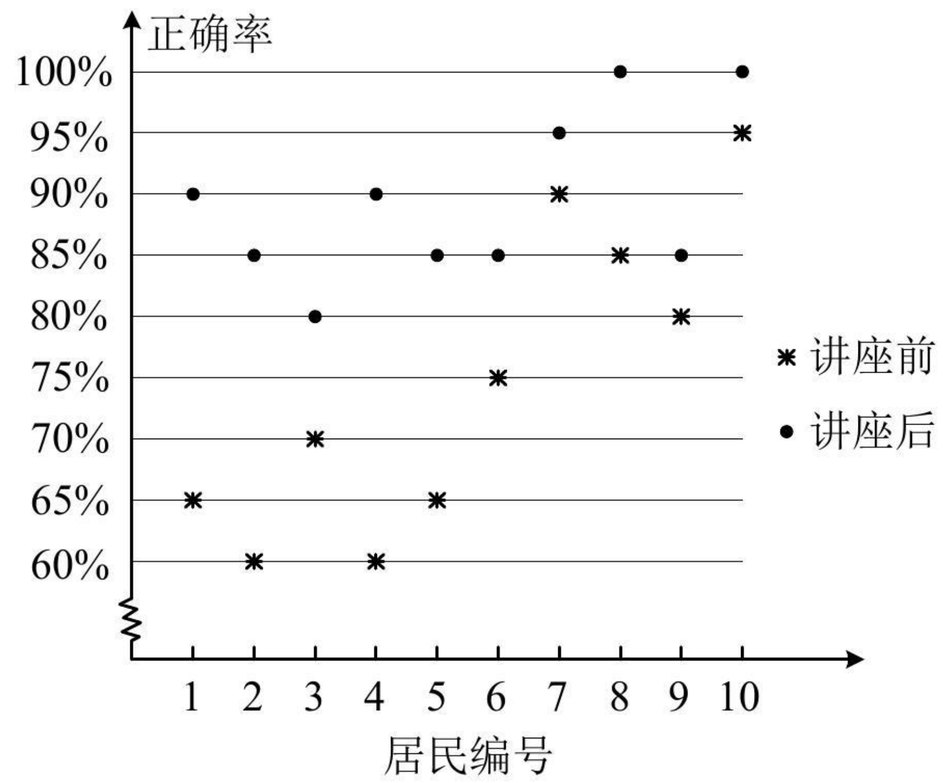
【例 5】（2022·全国甲卷）某社区通过公益讲座来普及社区居民的垃圾分类知识，随机抽取 10 位社区居民，让他们在讲座前和讲座后各回答一份垃圾分类知识问卷，这 10 位社区居民在讲座前和讲座后问卷答题的正确率如下图，则（　　）

（A）讲座前问卷答题的正确率的中位数小于 70%

（B）讲座后问卷答题的正确率的平均数大于 85%

（C）讲座前问卷答题的正确率的标准差小于讲座后正确率的标准差

（D）讲座后问卷答题的正确率的极差大于讲座前正确率的极差



解析：A 项，由图可知讲座前 10 位居民问卷答题的正确率按从小到大排列，第 5、6 位分别 70%，75%，所以中位数为 72.5%，故 A 项错误；

B 项，由图可知讲座后问卷答题的正确率的平均数为 $\frac{0.8+0.85\times4+0.9\times2+0.95+1\times2}{10}>0.85$，故 B 项正确；

C 项，计算标准差较复杂，故直接看图，观察波动情况，结合标准差的统计意义来分析，

由图可知讲座前问卷答题的正确率的数据比讲座后更分散，波动更大，所以讲座前标准差更大，故 C 项错误；

D 项，由图可知讲座后问卷答题的正确率的极差为 $100\%-80\%=20\%$，讲座前为 $95\%-60\%=35\%$，故 D 项错误.

答案：B

【变式】已知一组数据为：4，1，2，5，5，3，3，2，3，2，则这组数据的方差为_____；标准差为_____.

解析：要算方差，先算平均数，由题意，这组数据的平均数 $\overline{x}=\frac{4+1+2+5+5+3+3+2+3+2}{10}=3$，

由于 10 个数据中重复的数据较多，于是算方差可用内容提要中的公式 $s^2=\frac{1}{n}\sum_{i=1}^{k}f_i(y_i-\overline{x})^2$，

所以方差 $s^2=\frac{1}{10}[(1-3)^2+3\times(2-3)^2+3\times(3-3)^2+(4-3)^2+2\times(5-3)^2]=\frac{8}{5}$，标准差 $s=\frac{2\sqrt{10}}{5}$.

答案：$\frac{8}{5}$；$\frac{2\sqrt{10}}{5}$

【例 6】（多选）第一组数据：$x_1,x_2,\cdots,x_n$，由这组数据得到第二组数据：$y_1,y_2,\cdots,y_n$，其中 $y_i=ax_i+b(i=1,2,\cdots,n)$ 且 a，b 为正数，则下列命题正确的是（　　）

（A）当 $a=1$ 时，两组数据的平均数不相同

（B）第二组数据的极差是第一组的 a 倍

（C）第二组数据的方差是第一组的 a 倍

（D）第二组数据的标准差是第一组的 a 倍

解析：A 项，当 $a=1$ 时，设第一组数据的平均数为 $\overline{x}$，则第二组数据的平均数 $\overline{y}=\overline{x}+b\neq\overline{x}$，故 A 项正确；

B 项，不妨设 $x_1\le x_2\le x_3\le\cdots\le x_n$，则第一组数据的极差为 x_n-x_1，且 $y_1\le y_2\le y_3\le\cdots\le y_n$，所以第二组数据的极差为 $y_n-y_1=(ax_n+b)-(ax_1+b)=a(x_n-x_1)$，故 B 项正确；

C 项和 D 项，设第一组数据的标准差为 s，由方差和标准差的性质（内容提要中有证明过程）可知第二组数据的标准差为 as，方差为 a^2s^2，故 C 项错误，D 项正确.

答案：ABD

【反思】若数据 $x_i(i=1,2,\cdots,n)$ 的标准差为 s，则数据 $ax_i+b(i=1,2,\cdots,n)$ 的标准差为 $|a|s$，方差为 a^2s^2.

【变式】已知一组数据 x_1，x_2，x_3 的平均数为 4，方差为 2，则由 $3x_1-1$，$3x_2-1$，$3x_3-1$ 和 11 这四个数据组成的新数据组的方差是_____.

解析：由于数据中增加了个 11，所以无法像例 6 那样用结论计算，故直接代公式计算新老方差，并进行对比. 此处为了更方便地观察两者差异，我们选择公式 $s^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2$，

由题意，原数据组的平均数 $\overline{x}=\frac{x_1+x_2+x_3}{3}=4$，所以 $x_1+x_2+x_3=12$，

新数据组的平均数为 $\frac{(3x_1-1)+(3x_2-1)+(3x_3-1)+11}{4}=\frac{3(x_1+x_2+x_3)+8}{4}=11$，

接下来用 $s^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2$ 算方差，原数据组的方差 $s^2=\frac{1}{3}(x_1^2+x_2^2+x_3^2)-4^2=2$，所以 $x_1^2+x_2^2+x_3^2=54$，

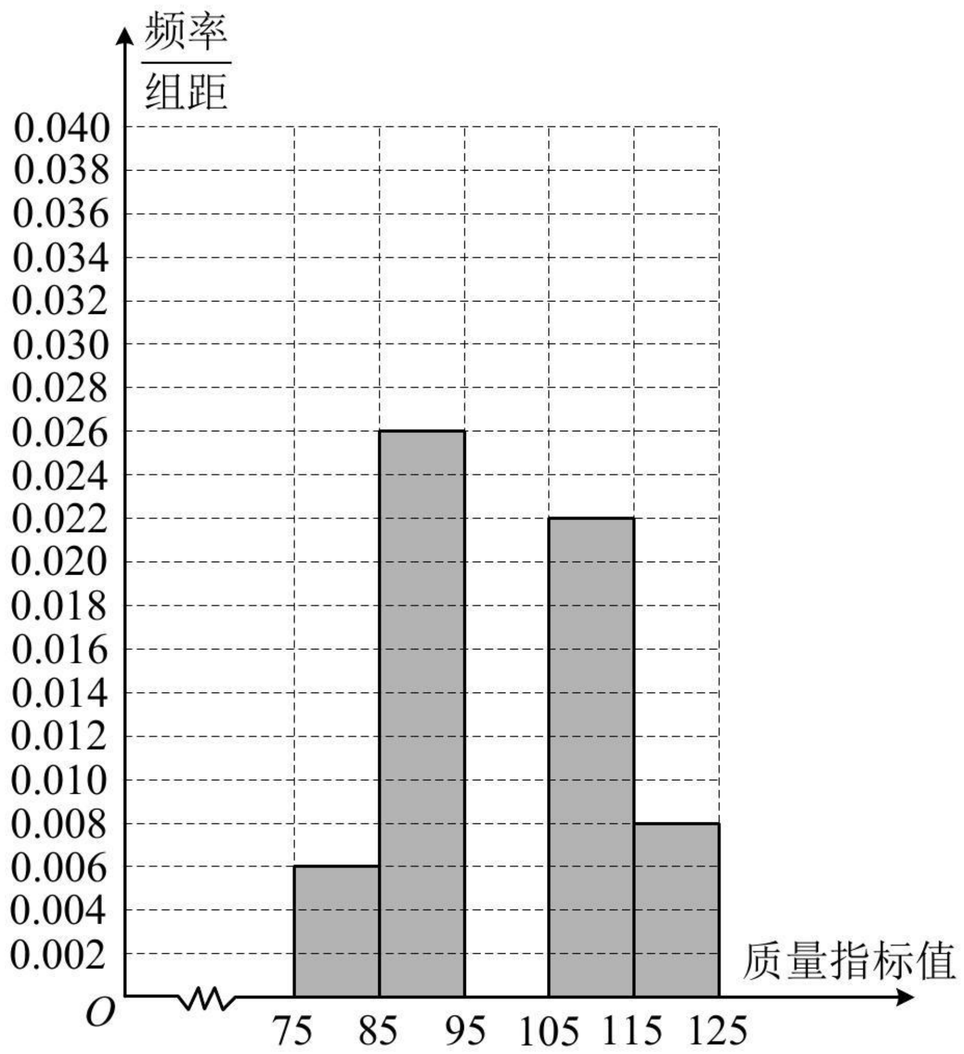
新数据组的方差为 $\frac{1}{4}[(3x_1-1)^2+(3x_2-1)^2+(3x_3-1)^2+11^2]-11^2$

$=\frac{1}{4}[9(x_1^2+x_2^2+x_3^2)-6(x_1+x_2+x_3)+124]-121=\frac{1}{4}(9\times54-6\times12+124)-121=\frac{27}{2}$.

答案：$\frac{27}{2}$

【反思】对于添项或减项的新数据组方差计算问题，常用 $s^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2$ 来对比新旧方差更方便. 需注意，用此公式的前提是 $\sum_{i=1}^{n}x_i^2$ 已知或好求，否则只能采用基本公式 $s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$.

【例 7】从某企业生产的某种产品中抽取 100 件，测量这些产品的一项质量指标值，由测量数据得到频率分布直方图如图所示.



（1）补全频率分布直方图；

（2）若同一组数据用该组区间的中点值作代表，试估计这种产品质量指标值的平均数 $\bar{x}$ 和方差 s^2.

解：（1）由图可知数据落在 $[95,105)$ 的频率为 $1-10\times(0.006+0.026+0.022+0.008)=0.38$，

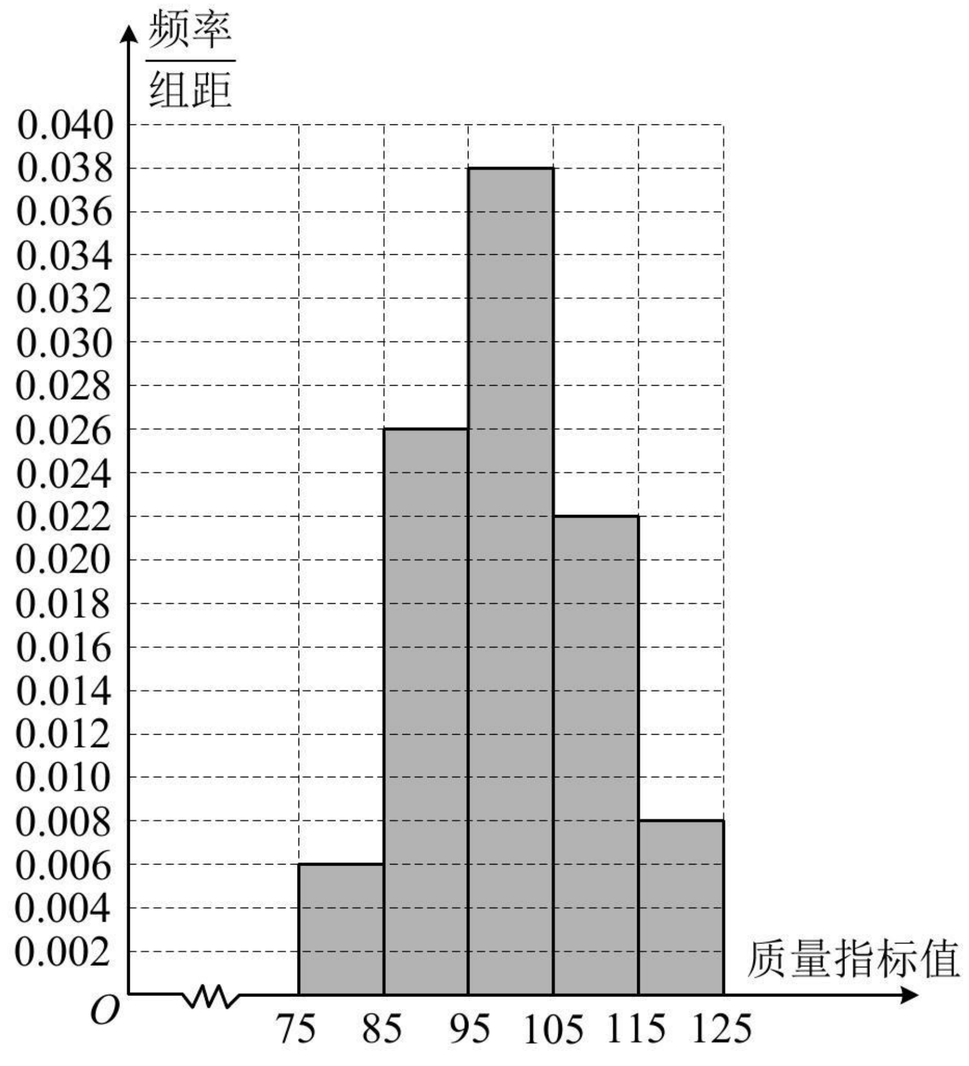
所以 $[95,105)$ 这一组的高为 0.038，故补全后的频率分布直方图如图.

（2）（由频率分布直方图估计样本平均数，用区间中点乘以对应的频率，再求和即可）

由题意，$\bar{x}=80\times0.06+90\times0.26+100\times0.38+110\times0.22+120\times0.08=100$，

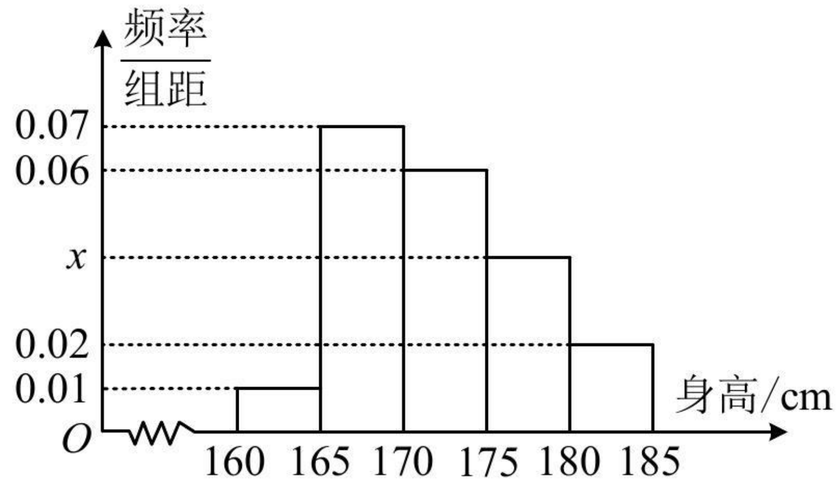
（由频率分布直方图估计样本方差，可代公式 $s^2=\sum_{i=1}^{n}(x_i-\bar{x})^2 f_i$，其中 x_i 为各组区间中点，f_i 为对应的频率）

$$s^2=(80-100)^2\times0.06+(90-100)^2\times0.26+(110-100)^2\times0.22+(120-100)^2\times0.08=104.$$



【反思】在频率分布直方图中，设第 i 组的区间中点为 x_i，频率为 f_i，其中 $i=1,2,\cdots,n$，若每组数据用区间中点值作代表，则样本平均数 $\overline{x}=\sum_{i=1}^{n}x_i f_i$，样本方差 $s^2=\sum_{i=1}^{n}(x_i-\overline{x})^2 f_i$.

【变式】为了调查某中学高一年级学生的身高情况，在高一年级随机抽取 100 名学生作为样本，把他们的身高（单位：cm）按照区间 $[160,165)$，$[165,170)$，$[170,175)$，$[175,180)$，$[180,185]$ 分组，得到样本身高的频率分布直方图如图所示.



（1）求频率分布直方图中 x 的值以及样本中身高不低于 175cm 的学生人数；

（2）统计过程中，小明与小张两位同学因事缺席，测得其余 98 名同学的平均身高为 172cm，方差为 29，之后补测得到小明与小张的身高分别为 171cm 与 173cm，试根据上述数据求样本的方差.

解：（1）由图可知，$5(0.01+0.07+0.06+x+0.02)=1$，解得：$x=0.04$，

样本中身高不低于 175cm 的学生人数为 $100\times5(x+0.02)=500x+10=30$.

（2）（此小问重新给出了有关数据，故不再使用频率分布直方图来算，下面先计算样本平均数）

设除小明与小张外的 98 名同学的身高分别为 $x_1,x_2,\cdots,x_{98}$，由题意，$\dfrac{x_1+x_2+\cdots+x_{98}}{98}=172$，

所以 $x_1+x_2+\cdots+x_{98}=98\times172$，故样本平均数 $\overline{x}=\dfrac{x_1+x_2+\cdots+x_{98}+171+173}{100}=\dfrac{98\times172+2\times172}{100}=172$，

（再算方差，可把 98 名学生的方差和 100 名学生的方差的算式都写出来，进行对比）

又 $\dfrac{1}{98}[(x_1-172)^2+(x_2-172)^2+\cdots+(x_{98}-172)^2]=29$，所以

$(x_1-172)^2+(x_2-172)^2+\cdots+(x_{98}-172)^2=29\times98=2842$，

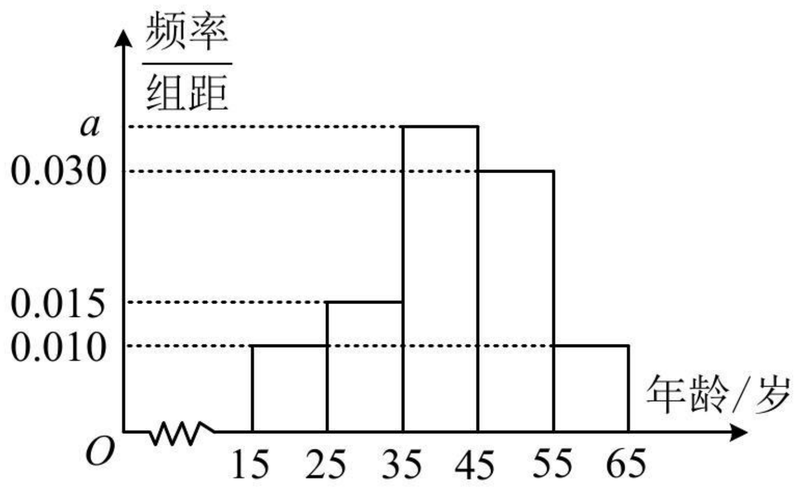
故样本方差

$s^2=\dfrac{1}{100}[(x_1-172)^2+(x_2-172)^2+\cdots+(x_{98}-172)^2+(171-172)^2+(173-172)^2]=\dfrac{1}{100}(2842+2)=28.44$.

强化训练

1.（2023·北京模拟·★）某直播间从参与购物的人群中随机选出 200 人，并将这 200 人按年龄分组，得到的频率分布直方图如图所示，则在这 200 人中，年龄在 $[25,35)$ 的人数 n，以及图中 a 的值是（　　）

（A）$n=35$，$a=0.032$　　（B）$n=35$，$a=0.32$　　（C）$n=30$，$a=0.035$　　（D）$n=30$，$a=0.35$



2.（2023·湖南长郡中学模拟·★★）已知甲、乙两组按从小到大顺序排列的数据：

甲组：27，28，37，m，40，50；乙组：24，n，34，43，48，52.

若这两组数据的第 30 百分位数，第 50 百分位数分别对应相等，则 $\dfrac{n}{m}=$_____.

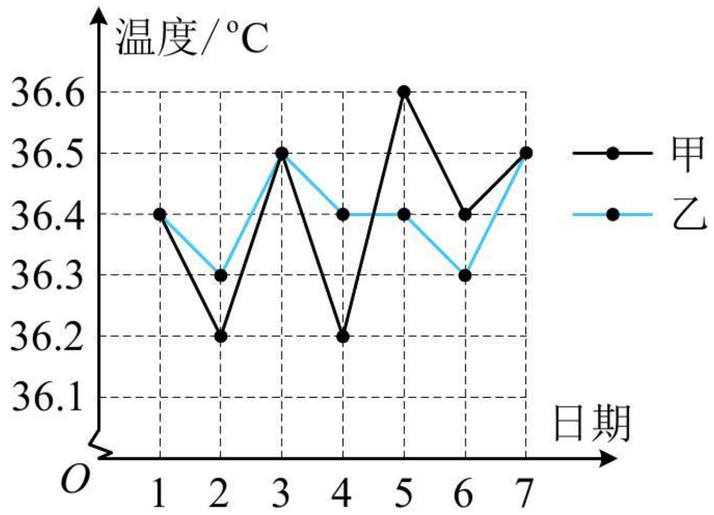
3.（2023·安徽模拟·★★）（多选）某校为做好疫情防控，每天早中晚都要对学生进行体温检测，某班级体温检测员对一周内甲、乙两名同学的体温进行了统计，其结果如图所示，则（　　）

（A）甲同学体温的极差为 $0.4^{\circ}\mathrm{C}$

（B）甲同学体温的第 60 百分位数为 $36.4^{\circ}\mathrm{C}$

（C）乙同学体温的众数为 $36.4^{\circ}\mathrm{C}$，中位数与平均数相等

（D）乙同学体温数据的方差比甲同学体温数据的方差小



4.（2023·新高考Ⅰ卷·★★）（多选）有一组样本数据 $x_1,x_2,\cdots,x_6$，其中 x_1 是最小值，x_6 是最大值，则（　　）

（A）x_2,x_3,x_4,x_5 的平均数等于 $x_1,x_2,\cdots,x_6$ 的平均数

（B）x_2,x_3,x_4,x_5 的中位数等于 $x_1,x_2,\cdots,x_6$ 的中位数

（C）x_2,x_3,x_4,x_5 的标准差不小于 $x_1,x_2,\cdots,x_6$ 的标准差

（D）x_2,x_3,x_4,x_5 的极差不大于 $x_1,x_2,\cdots,x_6$ 的极差

5.（2021·新高考Ⅰ卷·★★）（多选）有一组样本数据 $x_1,x_2,\cdots,x_n$，由这组数据得到样本数据 $y_1,y_2,\cdots,y_n$，其中 $y_i=x_i+c(i=1,2,\cdots,n)$，$c$ 为非零常数，则（　　）

（A）两组样本数据的样本平均数相同

（B）两组样本数据的样本中位数相同

（C）两组样本数据的样本标准差相同

（D）两组样本数据的样本极差相同

6.（2023·广东模拟·★★★）（多选）有一组样本数据 $x_1,x_2,\cdots,x_n$，其样本平均数为 $\overline{x}$，现加入一个新数据 $x_{n+1}(x_{n+1}<\overline{x})$，组成新的样本数据 $x_1,x_2,\cdots,x_n,x_{n+1}$，与原样本数据相比，新的样本数据可能（　　）

（A）平均数不变　　（B）众数不变　　（C）极差变小　　（D）第 20 百分位数变大

7．（2020・新课标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）$p_1 = p_4 = 0.3$，$p_2 = p_3 = 0.2$

8．（2023・河南模拟・★★★）为了让学生了解环保知识，增强环保意识，某班举行了一次环保知识有奖竞赛活动，有 20 名学生参加活动，已知这 20 名学生得分的平均数为 m，方差为 n，若将 m 当成一个学生的分数与原来的 20 名学生的分数一起，算出这 21 个分数的平均数为 m'，方差为 n'，则（　　）

（A）$20m = 21m'$，$21n = 20n'$　　（B）$m = m'$，$20n = 21n'$

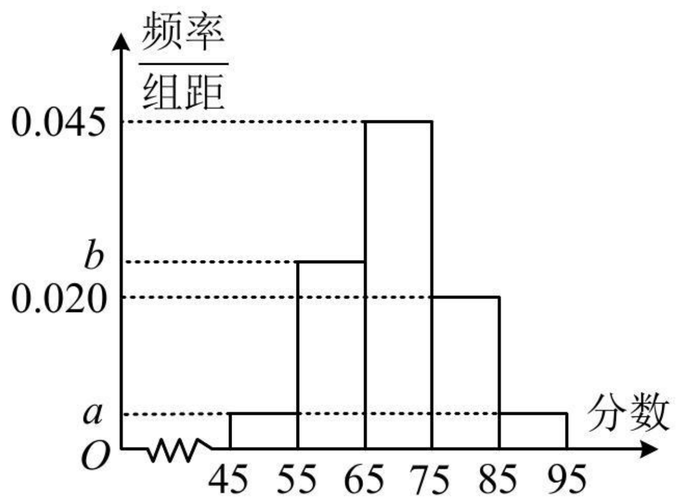
（C）$20m = 21m'$，$20n = 21n'$　　（D）$m = m'$，$21n = 20n'$

9．（2023・辽宁模拟・★★★）某班为了了解学生每月购买零食的支出情况，利用分层随机抽样抽取了一个 9 人的样本，统计如下：

	学生数	平均支出（元）	支出平方的累加值	方差
女生	4	$\overline{x} = 115$	$\sum_{i=1}^{4} x_i^2 = 53800$	225
男生	5	$\overline{y} = 106$	$\sum_{i=1}^{5} y_i^2 = 57700$	304

则样本的 9 人每月购买零食支出的平均数为_____元，方差为_____.（精确到小数点后一位）

10.（2022·上海模拟·★★）某高校承办了奥运会的志愿者选拔面试工作，现随机抽取了 100 名候选者的面试成绩并分成五组：[45,55)，[55,65)，[65,75)，[75,85)，[85,95]，绘制成如图所示的频率分布直方图，已知第三、四、五组的频率之和为 0.7，第一组和第五组的频率相同.



（1）求图中 a，b 的值；

（2）估计这 100 名候选者面试成绩的平均数、方差和第 60 百分位数（精确到 0.1）.

参考数据：$69.5^2=4830.25$.

11.（2023·全国乙卷·★★）某厂为比较甲乙两种工艺对橡胶产品伸缩率的处理效应，进行 10 次配对试验，每次配对试验，选用材质相同的两个橡胶产品，随机地选其中一个用甲工艺处理，另一个用乙工艺处理，测量处理后的橡胶产品的伸缩率，甲、乙两种工艺处理后的橡胶产品的伸缩率分别记为 x_i，$y_i(i=1,2,\cdots,10)$，试验结果如下：

试验序号 i	1	2	3	4	5	6	7	8	9	10
伸缩率 x_i	545	533	551	522	575	544	541	568	596	548
伸缩率 y_i	536	527	543	530	560	533	522	550	576	536

记 $z_i=x_i-y_i(i=1,2,\cdots,10)$，记 $z_1,z_2,\cdots,z_{10}$ 的样本平均数为 $\overline{z}$，样本方差为 s^2.

（1）求 $\overline{z}$，s^2，

（2）判断甲工艺处理后的橡胶产品的伸缩率较乙工艺处理后的橡胶产品的伸缩率是否有显著提高.（如果 $\overline{z}\geqslant 2\sqrt{\dfrac{s^2}{10}}$，则认为甲工艺处理后的橡胶产品的伸缩率较乙工艺处理后的橡胶产品的伸缩率有显著提高，否则不认为有显著提高）

12.（2022·全国模拟·★★★）已知 A，B 两家公司的员工月均工资（单位：万元）的情况分别如图 1，图 2 所示：

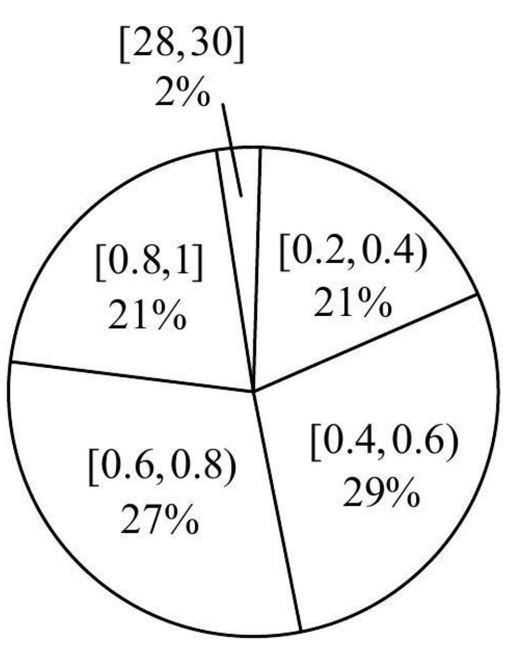


图1

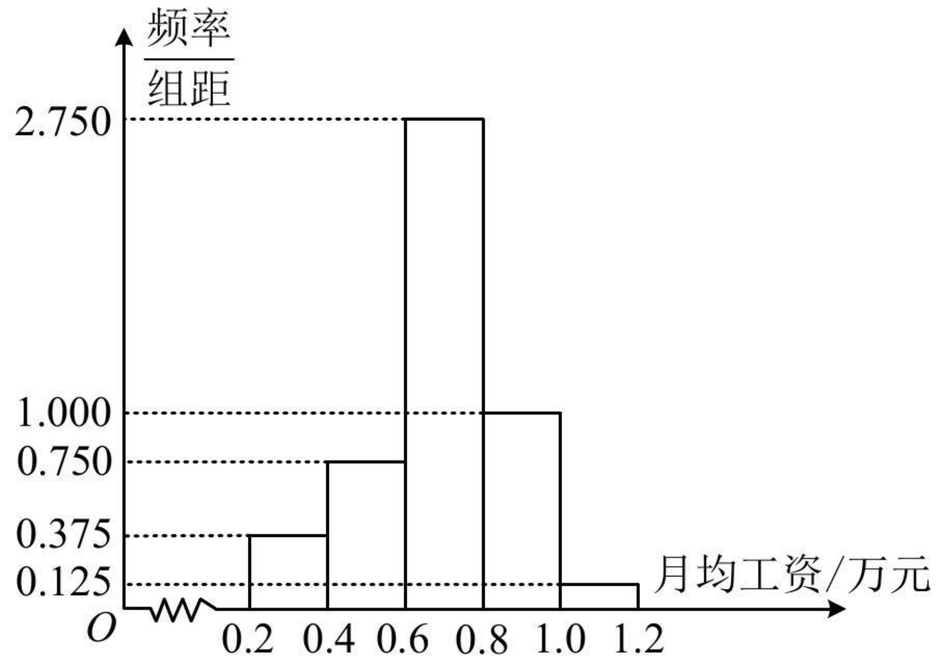
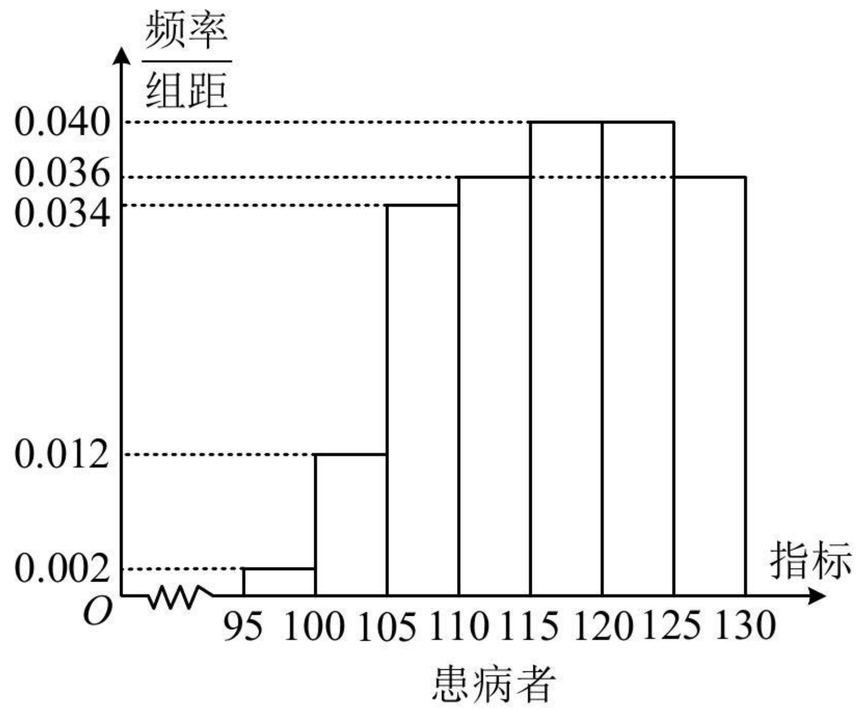


图2

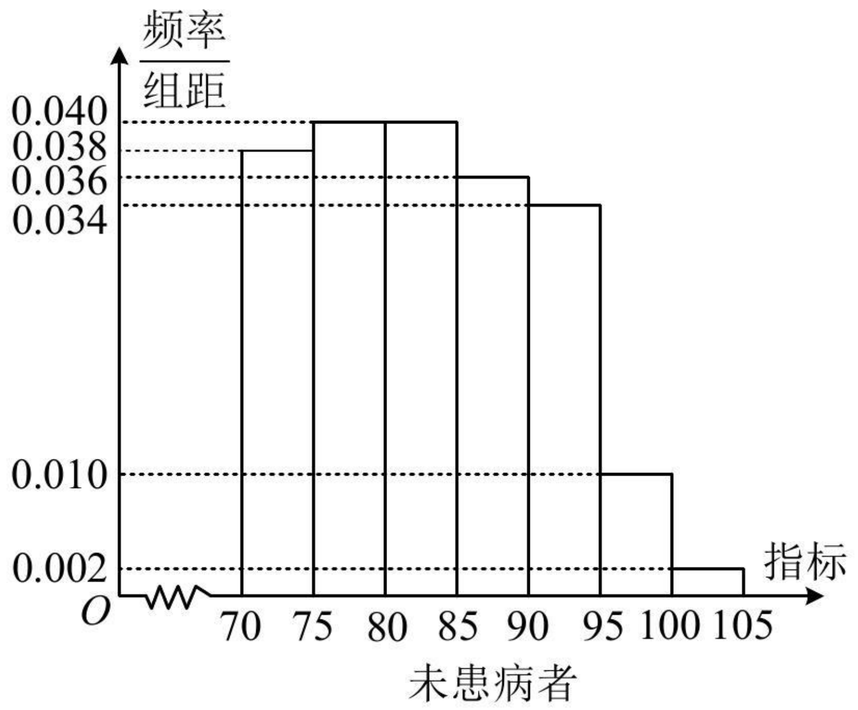
（1）以每组数据的区间中点值为代表，根据图 1 估计 A 公司员工月均工资的平均数、中位数，你认为哪个数据更能反映该公司普通员工的工资水平？请说明理由；

（2）小明拟到 A，B 两家公司中的一家应聘，以公司普通员工的工资水平作为决策依据，他应该选哪家公司？

13.（2023·新高考Ⅱ卷·★★★）某研究小组经过研究发现某种疾病的患病者与未患病者的某项医学指标有明显差异，经过大量调查，得到如下的患病者和未患病者该项指标的频率分布直方图：



患病者



未患病者

利用该指标制定一个检测标准，需要确定临界值 c，将该指标大于 c 的人判定为阳性，小于或等于 c 的人判定为阴性. 此检测标准的漏诊率是将患病者判定为阴性的概率，记为 $p(c)$；误诊率是将未患病者判定为阳性的概率，记为 $q(c)$. 假设数据在组内均匀分布. 以事件发生的频率作为相应事件发生的概率.

（1）当漏诊率 $p(c)=0.5\%$ 时，求临界值 c 和误诊率 $q(c)$；

（2）设函数 $f(c)=p(c)+q(c)$. 当 $c\in[95,105]$ 时，求 $f(c)$ 的解析式，并求 $f(c)$ 在区间 $[95,105]$ 的最小值.