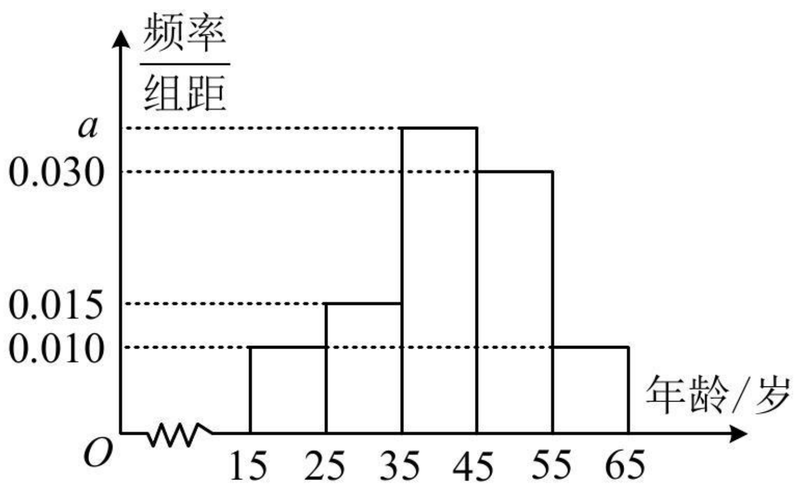
第 2 节 用样本估计总体（★★★）

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）$n=30$，$a=0.35$



答案：C

解析：可用样本量乘以 $[25,35)$ 这一组的频率来求 n，由题意，年龄在 $[25,35)$ 的人数 $n=200\times10\times0.015=30$；

再求 a，可用面积和为 1 来建立方程，$10\times0.01+10\times0.015+10\times a+10\times0.03+10\times0.01=1$，解得：$a=0.035$.

2.（2023·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 $\frac{n}{m}=$_____.

答案：$\frac{7}{10}$

解析：先求出两组数据的第 30、50 百分位数，

因为 $6\times30\%=1.8$，所以甲、乙两组数据的第 30 百分位数分别为 28，n，由题意，$n=28$；

因为 $6\times50\%=3$，所以甲、乙两组数据的第 50 百分位数分别为 $\frac{37+m}{2}$，$\frac{34+43}{2}$，

由题意，$\frac{37+m}{2}=\frac{34+43}{2}$，解得：$m=40$，所以 $\frac{n}{m}=\frac{28}{40}=\frac{7}{10}$.

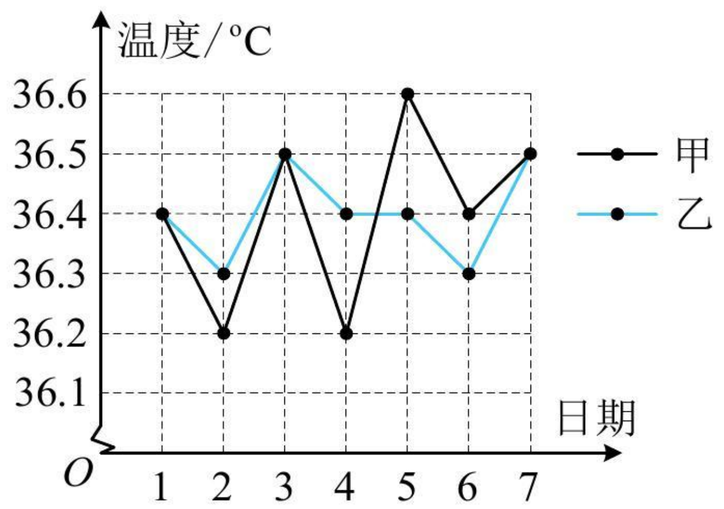
3.（2023·安徽模拟·★★）（多选）某校为做好疫情防控，每天早中晚都要对学生进行体温检测，某班级体温检测员对一周内甲、乙两名同学的体温进行了统计，其结果如图所示，则（　　）

（A）甲同学体温的极差为 $0.4^{\circ}C$

（B）甲同学体温的第 60 百分位数为 $36.4^{\circ}C$

（C）乙同学体温的众数为 $36.4^{\circ}C$，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



答案：ACD

解析：A 项，由图可知甲同学体温的数据从小到大依次为 36.2，36.2，36.4，36.4，36.5，36.5，36.6，所以甲同学体温的极差为$36.6-36.2=0.4^{\circ}\mathrm{C}$，故 A 项正确；

B 项，$i=7\times60\%=4.2$，所以甲同学体温的第 60 百分位数为$36.5^{\circ}\mathrm{C}$，故 B 项错误；

C 项，乙同学的体温数据从小到大依次为 36.3，36.3，36.4，36.4，36.4，36.5，36.5，所以众数为 36.4，中位数为 36.4，平均数为$\frac{36.3\times2+36.4\times3+36.5\times2}{7}=36.4$，故 C 项正确；

D 项，从折线图来看，乙同学体温的波动更小，其体温更稳定，所以方差也更小，故 D 项正确.

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

答案：BD

解析：A 项，x_1和x_6偏离平均数的程度不一定相同，所以去掉它们后，平均数可能发生变化，故能想象 A 项错误，我们举个例子，

不妨设这组数据为 0，2，3，4，5，6，

则原平均数$\overline{x}=\frac{0+2+3+4+5+6}{6}=\frac{10}{3}$，

去掉 0 和 6 之后的平均数$\overline{x}'=\frac{2+3+4+5}{4}=\frac{7}{2}\neq\overline{x}$，

故 A 项错误；

B 项，不妨假设$x_1\leqslant x_2\leqslant\cdots\leqslant x_6$，则$x_2,x_3,x_4,x_5$和$x_1,x_2,\cdots,x_6$的中位数都是$\frac{x_3+x_4}{2}$，故 B 项正确；

C 项，x_1和x_6偏离平均数较大，去掉它们后，标准差可能减小，故通过直观想象能得出 C 项错误，举个例子，不妨设这组数据为 1，2，3，5，6，7，

则$\overline{x}=\frac{1+2+3+5+6+7}{6}=4$，$s^2=\frac{1}{6}[(1-4)^2+(2-4)^2+$

$(3-4)^2+(5-4)^2+(6-4)^2+(7-4)^2]=\frac{14}{3}$，

去掉 1 和 7 后，$\overline{x}'=\dfrac{2+3+5+6}{4}=4$，

$s'^2=\dfrac{1}{4}[(2-4)^2+(3-4)^2+(5-4)^2+(6-4)^2]=\dfrac{5}{2}$，

所以 $s'^2<s^2$，从而 $s'<s$，故 C 项错误；

D 项，沿用 B 项的假设，则 x_2,x_3,x_4,x_5 的极差为 x_5-x_2，$x_1,x_2,\cdots,x_6$ 的极差为 x_6-x_1，

要比较两个极差的大小，可再将它们作差判断正负，

因为 $(x_6-x_1)-(x_5-x_2)=(x_6-x_5)+(x_2-x_1)\geqslant 0$，所以 $x_5-x_2\leqslant x_6-x_1$，故 D 项正确.

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

答案：CD

解析：A 项，设 $x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_n$ 的平均数分别为 $\overline{x}$ 和 $\overline{y}$，因为 $c\neq 0$，所以 $\overline{y}=\overline{x}+c\neq\overline{x}$，故 A 项错误；

分析中位数、极差要用数据的大小关系，先做个假设，不妨设 $x_1\leqslant x_2\leqslant x_3\leqslant\cdots\leqslant x_n$，则 $y_1\leqslant y_2\leqslant y_3\leqslant\cdots\leqslant y_n$，

B 项，直观感觉，在一组数据的每个数上加 c，中位数必定也加 c，若要严格论证，可对 n 分奇偶讨论，

当 n 为奇数时，数据 $x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_n$ 的中位数分别为 $x_{\frac{n+1}{2}}$ 和 $y_{\frac{n+1}{2}}$，且 $y_{\frac{n+1}{2}}=x_{\frac{n+1}{2}}+c\neq x_{\frac{n+1}{2}}$；

当 n 为偶数时，数据 $x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_n$ 的中位数分别为 $\dfrac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}$ 和 $\dfrac{y_{\frac{n}{2}}+y_{\frac{n}{2}+1}}{2}$，

且 $\dfrac{y_{\frac{n}{2}}+y_{\frac{n}{2}+1}}{2}=\dfrac{x_{\frac{n}{2}}+c+x_{\frac{n}{2}+1}+c}{2}=\dfrac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}+c\neq\dfrac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}$；所以两组数据的中位数必定不同，故 B 项错误；

C 项，由内容提要第 6 点③的公式 $s_y=as$，标准差只与 x_i 的系数有关，本题 x_i 的系数为 1，所以标准差不变，故 C 项正确；

D 项，数据 $x_1,x_2,\cdots,x_n$ 的极差为 x_n-x_1，数据 $y_1,y_2,\cdots,y_n$ 的极差为 $y_n-y_1=(x_n+c)-(x_1+c)=x_n-x_1$，故 D 项正确.

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

答案：BD

解析：A 项，直观感觉可知加入小于平均值的数据，会拉低平均数，下面给出严格论证，

因为 $x_{n+1}<\overline{x}$，所以新样本数据的平均数 $\overline{x}'=\dfrac{x_1+x_2+\cdots+x_n+x_{n+1}}{n+1}=\dfrac{n\overline{x}+x_{n+1}}{n+1}<\dfrac{n\overline{x}+\overline{x}}{n+1}=\overline{x}$，故 A 项错误；

B 项，无论新加入一个什么数据，出现次数最多的数都可能不变，所以众数可能不变，故 B 项正确；

C 项，由于 $x_{n+1}<\overline{x}$，所以加入 x_{n+1} 后，样本中最大的数据必定不变，当 x_{n+1} 不小于原来的最小数据时，极差不变，当 x_{n+1} 小于原最小数据时，极差会变大，从而极差不可能变小，故 C 项错误；

D 项，直观感觉可知只要加入的 x_{n+1} 比原来的第 20 百分位数大，那么加入后由于数据个数变多了，第 20 百分位数可能后移，从而变大，下面举个例子，

假设原样本数据是 1，2，3，…，100，共 100 个，因为 $i=100\times20\%=20$，所以该组数据的第 20 百分位数为 $\frac{20+21}{2}=20.5$，现在新加入数据 $x_{101}=21$，则此时 $i=101\times20\%=20.2$，从而新样本数据的第 20 百分位数为第 21 个数据，也即 21，比原来大，故 D 项正确.

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 $\sum\limits_{i=1}^{4}p_i=1$，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）$p_1=p_4=0.3$，$p_2=p_3=0.2$

答案：B

解析：若计算标准差再比较，则较为繁琐，可从标准差的统计意义来分析，标准差可刻画数据的波动程度，波动程度越大，标准差越大，

由所给数据的等间距性和对应频率的对称性可知四个选项求得的样本平均数都是 2.5，

远离平均数的数据越多，则数据波动程度越大，标准差也就越大，

而 1、2、3、4 四个数据中 1 和 4 偏离样本平均数较大，它们出现的频率越高，则标准差越大，故选 B.

【反思】由标准差的计算公式 $s=\sqrt{(x_1-\overline{x})^2f_1+(x_2-\overline{x})^2f_2+\cdots+(x_n-\overline{x})^2f_n}$（其中 $f_i(i=1,2,\cdots,n)$ 表示数据 x_i 出现的频率）知，远离平均数的数据越多，标准差越大.

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）$m=m'$，$21n=20n'$

答案：B

解析：直接观察不易得出结果，可代平均数、方差的计算公式来看，本题可以代基本公式 $s^2=\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\overline{x})^2$ 来对比，但用 $s^2=\frac{1}{n}\sum\limits_{i=1}^{n}x_i^2-\overline{x}^2$ 更方便，

记原来的 20 名学生的得分分别为 $x_1,x_2,\cdots,x_{20}$，

则 $m=\frac{x_1+x_2+\cdots+x_{20}}{20}$，所以 $x_1+x_2+\cdots+x_{20}=20m$，

故 $m'=\frac{x_1+x_2+\cdots+x_{20}+m}{21}=\frac{20m+m}{21}=m$，

$$n=\frac{1}{20}(x_1^2+x_2^2+\cdots+x_{20}^2)-m^2 \quad ①,$$

$$n'=\frac{1}{21}(x_1^2+x_2^2+\cdots+x_{20}^2+m^2)-m'^2$$

$$=\frac{1}{21}(x_1^2+x_2^2+\cdots+x_{20}^2+m^2)-m^2 \quad ②,$$

由①可得 $x_1^2+x_2^2+\cdots+x_{20}^2=20(n+m^2)$，

代入②得 $n'=\frac{1}{21}[20(n+m^2)+m^2]-m^2=\frac{20}{21}n$，即 $20n=21n'$.

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）

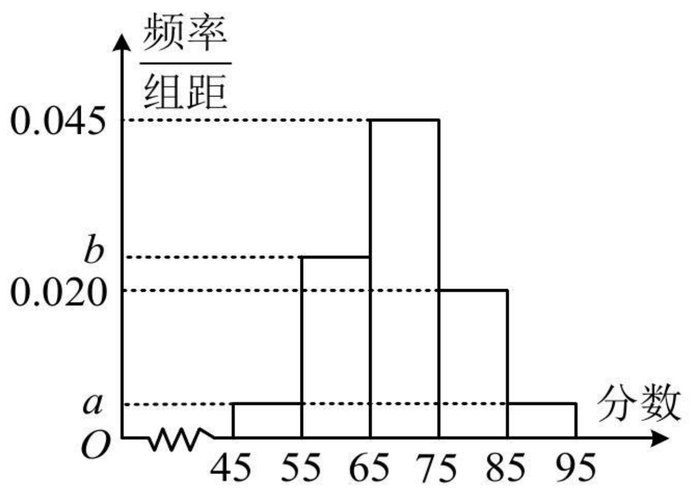
答案：110，288.9

解析：由表中数据可知样本的 9 人每月购买零食支出的平均数为 $\frac{4\overline{x}+5\overline{y}}{9}=\frac{4\times115+5\times106}{9}=110$ 元；

表中数据给的是 $\sum_{i=1}^{4}x_i^2$ 和 $\sum_{i=1}^{5}y_i^2$，故算方差用内容提要第 6 点①的公式 $s^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2$，

所以方差 $s^2=\frac{1}{9}(\sum_{i=1}^{4}x_i^2+\sum_{i=1}^{5}y_i^2)-110^2=\frac{1}{9}\times(53800+57700)-12100\approx288.9$.

10.（2022·上海模拟改·★★）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：$69.5^2=4830.25$.

解：（1）（由面积和为 1，第三、四、五组的频率之和为 0.7 可分别建立一个方程，求出 a 和 b）

由题意，$\begin{cases}10\times a+10\times b+10\times0.045+10\times0.02+10\times a=1\\10\times0.045+10\times0.02+10\times a=0.7\end{cases}$，解得：$a=0.005$，$b=0.025$.

（2）设 100 名候选者面试成绩的平均数为 $\overline{x}$，方差为 s^2，

则 $\overline{x}=50\times0.05+60\times0.25+70\times0.45+80\times0.2+90\times0.05=69.5$，

$s^2=50^2\times0.05+60^2\times0.25+70^2\times0.45+80^2\times0.2+90^2\times0.05-69.5^2=84.75$；

（再估计第 60 百分位数，只需在频率分布直方图中找到从左至右频率和为 0.6 的位置即可）

由图可知前两组的频率和为 $0.05+0.25=0.3<0.6$，前三组的频率和为 $0.05+0.25+0.45=0.75>0.6$，

所以第 60 百分位数应在 $[65,75)$ 这一组，设为 y，则 $0.3+(y-65)\times0.045=0.6$ ①，

解得：$y\approx71.7$，故可估计这 100 名候选者面试成绩的平均数为 69.5，第 60 百分位数为 71.7.

【反思】若不清楚解答过程中的式①怎么来的，可参考本节例 2 的变式 2.

11.（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：

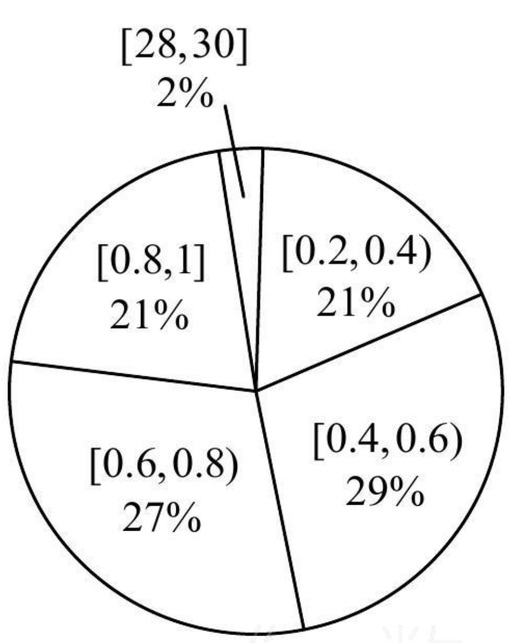


图1

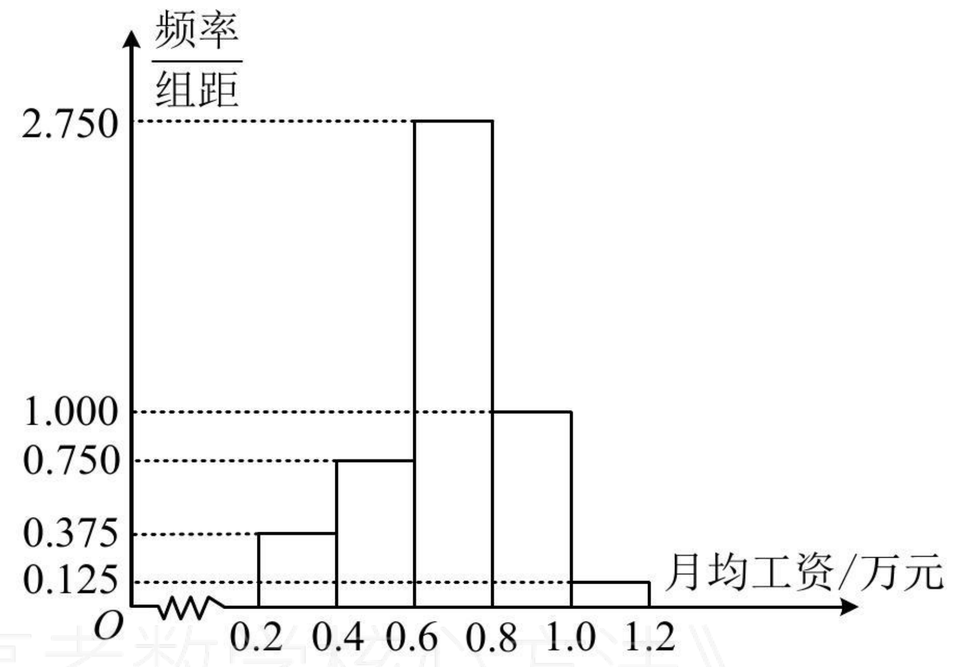


图2

（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？

解：（1）设 A 公司共有 a 名员工，该公司员工月均工资的平均数为 $\overline{x}_A$，中位数为 x_A，

则 $\overline{x}_A=\dfrac{0.21a\times0.3+0.29a\times0.5+0.27a\times0.7+0.21a\times0.9+0.02a\times29}{a}=1.166$ 万元，

因为 $[0.2,0.4)$ 与 $[0.4,0.6)$ 的频率和为 $0.21+0.29=0.5$，所以 A 公司员工月均工资的中位数 x_A 约为 0.6 万元，

（要分析平均数和中位数谁更有参考性，应先找到平均数明显大于中位数的原因，再加以阐释）

因为平均数会受少数极端数据的影响，该公司有2%的少数员工的月均工资大幅高于普通员工，从而拉高了员工的月均工资平均数，在这种情况下，平均数不能很好地反映普通员工的工资水平，而中位数不受少数极端数据的影响，可以较好地反映普通员工的工资水平，所以中位数更能反映该公司普通员工的工资水平.

（2）（要做决策，先计算 B 公司员工月均工资的平均数和中位数，与前面的 x_A 比较）

设 B 公司员工的月均工资平均数的估计值为 $\overline{x}_B$，中位数的估计值为 x_B，

则由图可知 $\overline{x}_B=0.2\times(0.375\times0.3+0.75\times0.5+2.75\times0.7+1\times0.9+0.125\times1.1)=0.69$，

图 2 中前两组的频率和 $0.2\times(0.375+0.75)=0.225<0.5$，前三组的频率和 $0.2\times(0.375+0.75+2.75)=0.775>0.5$，

所以中位数 x_B 在 $[0.6,0.8)$ 内，且 $0.225+(x_B-0.6)\times 2.75=0.5$，解得：$x_B=0.7$，

（求出了 B 公司员工月均工资的平均数和中位数，应先结合图形阐释他们是否有代表性，再与 A 公司比较）

从图 2 可以看出，B 公司员工月均工资数据比较集中，且没有出现极端值，月均工资的平均数和中位数都能很好地反映该公司普通员工的工资水平，B 公司员工月均工资的平均数和中位数均大于 A 公司员工月均工资的中位数，故以公司普通员工的工资水平作为决策依据，他应该选 B 公司.

12.（2021·全国乙卷·★★★）某厂研制了一种生产高精产品的设备，为检验新设备生产产品的某项指标有无提高，用一台旧设备和一台新设备各生产了 10 件产品，得到各件产品该项指标数据如下：

旧设备	9.8	10.3	10.0	10.2	9.9	9.8	10.0	10.1	10.2	9.7
新设备	10.1	10.4	10.1	10.0	10.1	10.3	10.6	10.5	10.4	10.5

旧设备和新设备生产产品的该项指标的样本平均数分别记为 $\overline{x}$ 和 $\overline{y}$，样本方差分别记为 s_1^2 和 s_2^2.

（1）求 $\overline{x}$，$\overline{y}$，s_1^2，s_2^2；

（2）判断新设备生产产品的该项指标的均值较旧设备是否有显著提高（如果 $\overline{y}-\overline{x}\geq 2\sqrt{\dfrac{s_1^2+s_2^2}{10}}$，则认为新设备生产产品的该项指标的均值较旧设备有显著提高，否则不认为有显著提高）.

解：（1）（计算平均数前，最好先将数据排序，梳理出重复数据，例如，旧设备的数据可整理为 9.7，2 个 9.8，9.9，2 个 10，10.1，2 个 10.2，10.3，这样计算时就不会遗漏或看错）

由表中数据可得，$\overline{x}=\dfrac{1}{10}(9.7+9.8\times 2+9.9+10\times 2+10.1+10.2\times 2+10.3)=10$，

$\overline{y}=\dfrac{1}{10}(10+10.1\times 3+10.3+10.4\times 2+10.5\times 2+10.6)=10.3$，

$s_1^2=\dfrac{1}{10}[(9.7-10)^2+2\times(9.8-10)^2+(9.9-10)^2+2\times(10-10)^2+(10.1-10)^2+2\times(10.2-10)^2+(10.3-10)^2]=0.036$，

$s_2^2=\dfrac{1}{10}[(10-10.3)^2+3\times(10.1-10.3)^2+(10.3-10.3)^2+2\times(10.4-10.3)^2+2\times(10.5-10.3)^2+(10.6-10.3)^2]=0.04$.

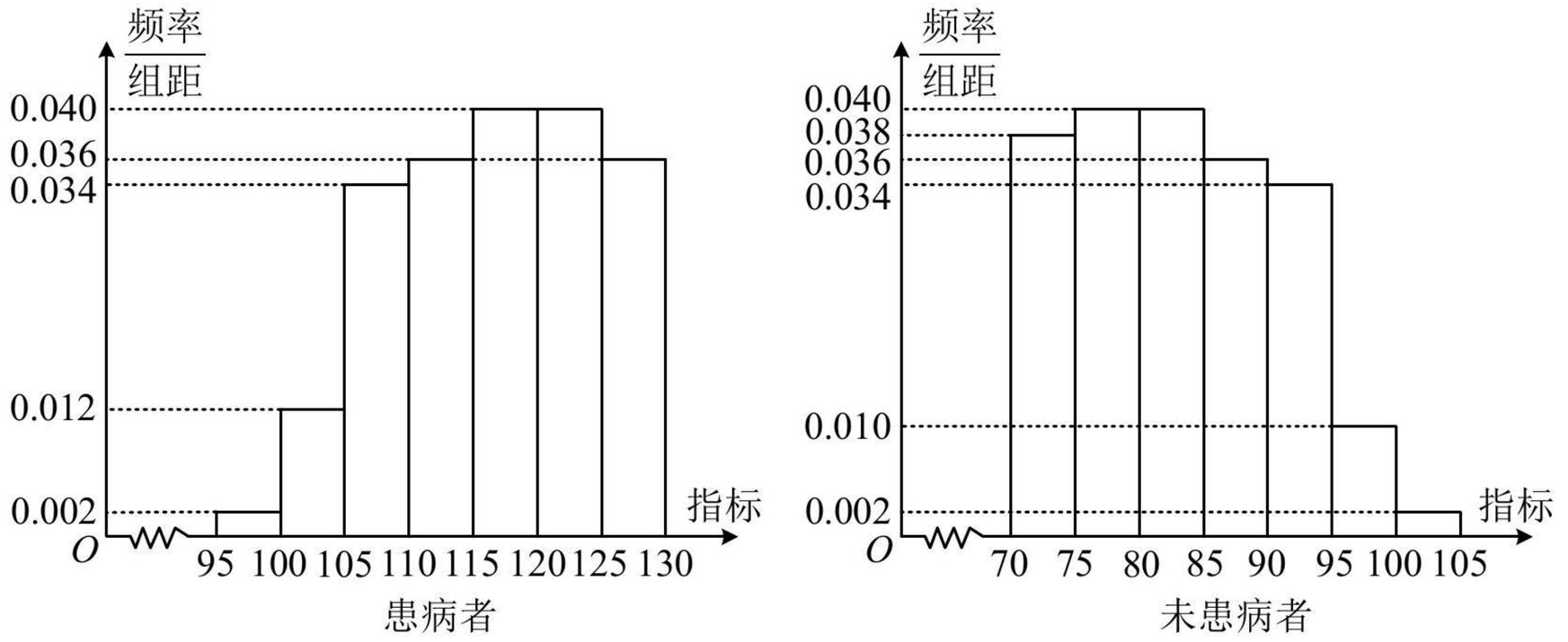
（2）（读完题可发现，只需结合（1）的结果计算 $\overline{y}-\overline{x}$ 和 $2\sqrt{\dfrac{s_1^2+s_2^2}{10}}$，并加以比较，即可解决问题）

由（1）知 $\overline{y}-\overline{x}=0.3$，$2\sqrt{\dfrac{s_1^2+s_2^2}{10}}=2\sqrt{\dfrac{0.036+0.04}{10}}=2\sqrt{0.0076}$，（要比较大小，简单估算一下即可）

因为 $2\sqrt{0.0076}<2\sqrt{0.01}=2\times 0.1=0.2<0.3$，所以 $\overline{y}-\overline{x}>2\sqrt{\dfrac{s_1^2+s_2^2}{10}}$，

故新设备生产产品的该项指标的均值较旧设备有显著提高.

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：



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.

解：（1）（给的是漏诊率，故先看患病者的图，漏诊率为 0.5%即小于或等于 c 的频率为 0.5%，可由此求 c）

由患病者的图可知，$[95,100)$ 这组的频率为 $5\times0.002=0.01>0.005$，所以 c 在 $[95,100)$ 内，

且 $(c-95)\times0.002=0.005$，解得：$c=97.5$；

（要求 $q(c)$，再来看未患病者的图，$q(c)$ 是误诊率，也即未患病者判定为阳性（指标大于 c）的概率）

由未患病者的图可知指标大于 97.5 的概率为 $(100-97.5)\times0.01+5\times0.002=0.035$，所以 $q(c)=3.5\%$.

（2）（$[95,105]$ 包含两个分组，故应分类讨论）当 $95\le c<100$ 时，$p(c)=(c-95)\times0.002$，

$q(c)=(100-c)\times0.01+5\times0.002$，所以 $f(c)=p(c)+q(c)=-0.008c+0.82$，

故 $f(c)>-0.008\times100+0.82=0.02$ ①；

当 $100\le c\le105$ 时，$p(c)=5\times0.002+(c-100)\times0.012$，$q(c)=(105-c)\times0.002$，

所以 $f(c)=p(c)+q(c)=0.01c-0.98$，故 $f(c)\ge f(100)=0.01\times100-0.98=0.02$ ②；

所以 $f(c)=\begin{cases}-0.008c+0.82,95\le c<100\\0.01c-0.98,100\le c\le105\end{cases}$，且由①②可得 $f(c)_{\min}=0.02$.

14.（2022·全国模拟·★★★★）随着高校强基计划招生的持续开展，我市高中生抓起了参与数学兴趣小组的热潮. 为调查我市高中生对数学学习的喜好程度，从甲、乙两所高中各随机抽取了 40 名学生，记录他们在一周内平均每天学习数学的时间，并将其分成了 6 个区间：$(0,10]$，$(10,20]$，$(20,30]$，$(30,40]$，$(40,50]$，$(50,60]$，整理得到如下的频率分布直方图.

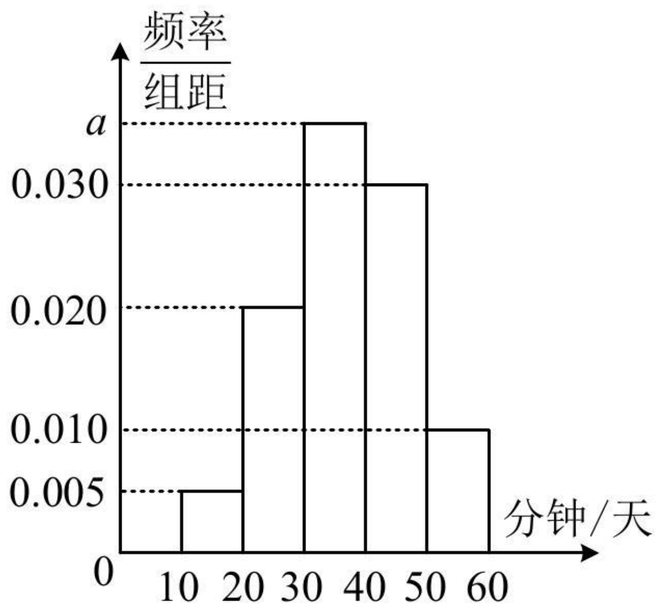


图1:甲高中

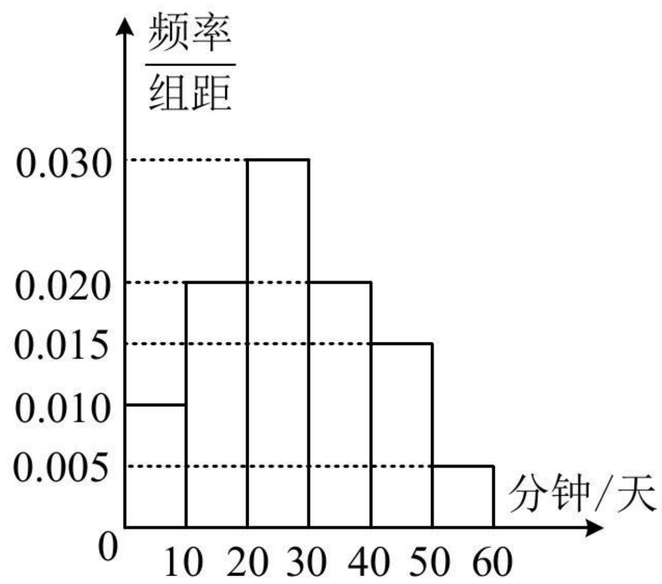


图2:乙高中

（1）求图 1 中 a 的值，并估计甲高中学生一周内平均每天学习数学时间的众数；

（2）设甲、乙高中学生一周内平均每天学习数学时间的样本方差分别为 $s_{甲}^2$ 和 $s_{乙}^2$，比较它们的大小；（无需计算，说明理由即可）

（3）若从甲、乙两所高中分别抽取样本量为 m，n 的两个样本，经计算得到它们的平均数和方差分别为 $\bar{x}$，s_1^2 与 $\bar{y}$，s_2^2，记总样本的平均数为 $\bar{w}$，样本方差为 s^2，证明：

（i）$\bar{w}=\dfrac{m}{m+n}\bar{x}+\dfrac{n}{m+n}\bar{y}$；（ii）$s^2=\dfrac{1}{m+n}\{m[s_1^2+(\bar{x}-\bar{w})^2]+n[s_2^2+(\bar{y}-\bar{w})^2]\}$.

解：（1）由图 1 可知，$10\times(0.005+0.02+a+0.03+0.01)=1$，解得：$a=0.035$，

（由频率分布直方图估计众数，直接取最高的小矩形区间中点即可）

可估计甲高中学生一周内平均每天学习数学时间的众数为 35 分钟.

（2）比较图 1 和图 2 可知，甲的数据更集中，乙的数据更分散，故 $s_{甲}^2<s_{乙}^2$.

（3）（i）（可按 $\bar{w}=\dfrac{\text{两所高中抽取的数据总和}}{\text{总样本量}}$ 来计算总样本平均数）由题意，$\bar{w}=\dfrac{m\bar{x}+n\bar{y}}{m+n}=\dfrac{m}{m+n}\bar{x}+\dfrac{n}{m+n}\bar{y}$.

（ii）**证法 1：**记甲高中抽取的样本数据为 $x_1,x_2,\cdots,x_m$，乙高中抽取的样本数据为 $y_1,y_2,\cdots,y_n$，

（要证问题的结论，先把 s_1^2，s_2^2，s^2 的计算公式写出来，再加以对比，方差公式可用 $s^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i^2-\bar{x}^2$）

则 $s_1^2=\dfrac{1}{m}(x_1^2+x_2^2+\cdots+x_m^2)-\bar{x}^2$ ①，$s_2^2=\dfrac{1}{n}(y_1^2+y_2^2+\cdots+y_n^2)-\bar{y}^2$ ②，

$s^2=\dfrac{1}{m+n}(x_1^2+x_2^2+\cdots+x_m^2+y_1^2+y_2^2+\cdots+y_n^2)-\bar{w}^2$ ③，

（要证的式子中不含 $x_1,x_2,\cdots,x_m$ 和 $y_1,y_2,\cdots,y_n$，应消去它们，可由①②把它们的平方和解出来，代入③）

由①②可得 $\begin{cases}x_1^2+x_2^2+\cdots+x_m^2=m(s_1^2+\bar{x}^2)\\y_1^2+y_2^2+\cdots+y_n^2=n(s_2^2+\bar{y}^2)\end{cases}$，代入③可得 $s^2=\dfrac{1}{m+n}[m(s_1^2+\bar{x}^2)+n(s_2^2+\bar{y}^2)]-\bar{w}^2$ ④，

（把上式右侧与我们要证明的式子的右侧比较，发现有诸多部分是相同的，可作差将它们抵消再看）

所以 $\dfrac{1}{m+n}\{m[s_1^2+(\bar{x}-\bar{w})^2]+n[s_2^2+(\bar{y}-\bar{w})^2]\}-\{\dfrac{1}{m+n}[m(s_1^2+\bar{x}^2)+n(s_2^2+\bar{y}^2)]-\bar{w}^2\}$

$=\dfrac{1}{m+n}[m(\bar{x}-\bar{w})^2+n(\bar{y}-\bar{w})^2-m\bar{x}^2-n\bar{y}^2+(m+n)\bar{w}^2]$

$$=\frac{1}{m+n}[m\bar{x}^2-2m\bar{x}\bar{w}+m\bar{w}^2+n\bar{y}^2-2n\bar{y}\bar{w}+n\bar{w}^2-m\bar{x}^2-n\bar{y}^2+(m+n)\bar{w}^2]=\frac{2}{m+n}[(m+n)\bar{w}^2-(m\bar{x}+n\bar{y})\bar{w}] \quad ⑤,$$

（观察发现只要把外面的 $\frac{1}{m+n}$ 乘进去，可结合（i）的结果将 $(m\bar{x}+n\bar{y})$ 化为 $\bar{w}$ ）

由（i）知 $\bar{w}=\frac{m}{m+n}\bar{x}+\frac{n}{m+n}\bar{y}$，所以式⑤中

$$\frac{2}{m+n}[(m+n)\bar{w}^2-(m\bar{x}+n\bar{y})\bar{w}]=2[\bar{w}^2-(\frac{m}{m+n}\bar{x}+\frac{n}{m+n}\bar{y})\bar{w}]=2(\bar{w}^2-\bar{w}\cdot\bar{w})=0,$$

从而 $\frac{1}{m+n}\{m[s_1^2+(\bar{x}-\bar{w})^2]+n[s_2^2+(\bar{y}-\bar{w})^2]\}-\{\frac{1}{m+n}[m(s_1^2+\bar{x}^2)+n(s_2^2+\bar{y}^2)]-\bar{w}^2\}=0$，

故 $\frac{1}{m+n}\{m[s_1^2+(\bar{x}-\bar{w})^2]+n[s_2^2+(\bar{y}-\bar{w})^2]\}=\frac{1}{m+n}[m(s_1^2+\bar{x}^2)+n(s_2^2+\bar{y}^2)]-\bar{w}^2$，

代入④得 $s^2=\frac{1}{m+n}\{m[s_1^2+(\bar{x}-\bar{w})^2]+n[s_2^2+(\bar{y}-\bar{w})^2]\}$.

证法 2：记甲高中抽取的样本数据为 $x_1,x_2,\cdots,x_m$，乙高中抽取的样本数据为 $y_1,y_2,\cdots,y_n$，

（先把 s^2，s_1^2，s_2^2 的计算公式写出来，再观察它和要证明的式子的结构特征，方差公式可用 $s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ ）

由题意，$s^2=\frac{1}{m+n}[\sum_{i=1}^{m}(x_i-\bar{w})^2+\sum_{i=1}^{n}(y_i-\bar{w})^2]$，$s_1^2=\frac{1}{m}\sum_{i=1}^{m}(x_i-\bar{x})^2$ ①，$s_2^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$ ②，

（要证明的式子中有 $(\bar{x}-\bar{w})^2$，$(\bar{y}-\bar{w})^2$ 这种结构，故考虑在 s^2 的式子中将其凑出来）

所以 $s^2=\frac{1}{m+n}[\sum_{i=1}^{m}(x_i-\bar{w})^2+\sum_{i=1}^{n}(y_i-\bar{w})^2]=\frac{1}{m+n}[\sum_{i=1}^{m}(x_i-\bar{x}+\bar{x}-\bar{w})^2+\sum_{i=1}^{n}(y_i-\bar{y}+\bar{y}-\bar{w})^2]$

$$=\frac{1}{m+n}[\sum_{i=1}^{m}(x_i-\bar{x})^2+\sum_{i=1}^{m}(\bar{x}-\bar{w})^2+\sum_{i=1}^{m}2(x_i-\bar{x})(\bar{x}-\bar{w})+\sum_{i=1}^{n}(y_i-\bar{y})^2+\sum_{i=1}^{n}(\bar{y}-\bar{w})^2+\sum_{i=1}^{n}2(y_i-\bar{y})(\bar{y}-\bar{w})],$$

（此时我们发现可结合式①②，先化掉 $\sum_{i=1}^{m}(x_i-\bar{x})^2$ 和 $\sum_{i=1}^{n}(y_i-\bar{y})^2$ 这两部分，向要证明的结构靠拢）

结合①②可得 $s^2=\frac{1}{m+n}[ms_1^2+m(\bar{x}-\bar{w})^2+\sum_{i=1}^{m}2(x_i-\bar{x})(\bar{x}-\bar{w})+ns_2^2+n(\bar{y}-\bar{w})^2+\sum_{i=1}^{n}2(y_i-\bar{y})(\bar{y}-\bar{w})]$ ③，

（把式③与要证的式子对比，发现接下来只需证 $\sum_{i=1}^{m}2(x_i-\bar{x})(\bar{x}-\bar{w})+\sum_{i=1}^{n}2(y_i-\bar{y})(\bar{y}-\bar{w})=0$ 即可）

$$\sum_{i=1}^{m}2(x_i-\bar{x})(\bar{x}-\bar{w})=2(\bar{x}-\bar{w})\sum_{i=1}^{m}(x_i-\bar{x})=2(\bar{x}-\bar{w})(x_1-\bar{x}+x_2-\bar{x}+\cdots+x_m-\bar{x})$$

$=2(\bar{x}-\bar{w})[(x_1+x_2+\cdots+x_m)-m\bar{x}]=2(\bar{x}-\bar{w})(m\bar{x}-m\bar{x})=0$，同理，$\sum_{i=1}^{n}2(y_i-\bar{y})(\bar{y}-\bar{w})=0$，

代入③可得 $s^2=\frac{1}{m+n}[ms_1^2+m(\bar{x}-\bar{w})^2+ns_2^2+n(\bar{y}-\bar{w})^2]=\frac{1}{m+n}\{m[s_1^2+(\bar{x}-\bar{w})^2]+n[s_2^2+(\bar{y}-\bar{w})^2]\}$.

【反思】第（3）问来自必修二第 216 页推广探索中的一道题，对符号运算的能力要求颇高，新教材上类似的习题不少，启示了我们在学习概率统计有关公式时，不仅要做到识记、应用，还应强化公式的推导.