分层计算总体的方差

若一个总体划分为两层，通过按样本量比例分配分层随机抽样，各层抽取的样本量、样本平均数和样本方差分别为：m，$\bar{x}$，S_1^2；n，$\bar{y}$，S_2^2．记总的样本平均数为$\bar{w}$，样本方差为S^2，

①$\bar{w}=\dfrac{m\bar{x}+n\bar{y}}{m+n}=\dfrac{m}{m+n}\bar{x}+\dfrac{n}{m+n}\bar{y}$；

②$S^2=\dfrac{1}{m+n}\left[\sum\limits_{i=1}^{m}(x_i-\bar{w})^2+\sum\limits_{j=1}^{n}(y_j-\bar{w})^2\right]=\dfrac{1}{m+n}\left[\sum\limits_{i=1}^{m}(x_i-\bar{x}+\bar{x}-\bar{w})^2+\sum\limits_{j=1}^{n}(y_j-\bar{y}+\bar{y}-\bar{w})^2\right]$

$=\dfrac{1}{m+n}\left[\sum\limits_{m}^{i=1}(x_i-\bar{x})^2+2\sum\limits_{m}^{i=1}(x_i-\bar{x})(\bar{x}-\bar{w})+m(\bar{x}-\bar{w})^2+\right.$

$\left.\sum\limits_{n}^{j=1}(y_j-\bar{y})^2+2\sum\limits_{n}^{j=1}(y_j-\bar{y})(\bar{y}-\bar{w})+n(\bar{y}-\bar{w})^2\right]$

又$\sum\limits_{i=1}^{m}(x_i-\bar{x})(\bar{x}-\bar{w})=\sum\limits_{i=1}^{m}x_i(\bar{x}-\bar{w})-m\bar{x}(\bar{x}-\bar{w})=m\bar{x}(\bar{x}-\bar{w})-m\bar{x}(\bar{x}-\bar{w})=0$

同理$\sum\limits_{j=1}^{n}(y_j-\bar{y})(\bar{y}-\bar{w})=0$，

$\therefore S^2=\dfrac{1}{m+n}\left[\sum\limits_{i=1}^{m}(x_i-\bar{x})^2+m(\bar{x}-\bar{w})^2+\sum\limits_{j=1}^{n}(y_j-\bar{y})^2+n(\bar{y}-\bar{w})^2\right]$

$=\dfrac{1}{m+n}\left[mS_1^2+m(\bar{x}-\bar{w})^2+nS_2^2+n(\bar{y}-\bar{w})^2\right]$

$=\dfrac{1}{m+n}\left\{m\left[S_1^2+(\bar{x}-\bar{w})^2\right]+n\left[S_2^2+\left(\bar{y}-\bar{w}\right)^2\right]\right\}$.

推广：

1．已知采用分层抽样得到的样本数据由两部分组成，第一部分样本数据$x_i\left(i=1,2,\cdots,m\right)$的平均数为$\bar{x}$，方差为$s_x^2$；第二部分样本数据$y_i\left(i=1,2,\cdots,n\right)$的平均数为$\bar{y}$，方差为$s_y^2$，设$\bar{x}\le\bar{y},s_x^2\le s_y^2$，则以下命题正确的是（　　）

A．设总样本的平均数为$\bar{z}$，则$\bar{x}\le\bar{z}\le\bar{y}$

B．设总样本的平均数为$\bar{z}$，则$\bar{z}^2\ge\bar{x}\cdot\bar{y}$

C．设总样本的方差为s^2，则$s_x^2\le s^2\le s_y^2$

D．若$m=n,\bar{x}=\bar{y}$，则$s^2=\dfrac{s_x^2+s_y^2}{2}$

1．AD

【分析】对于 A 选项，因为$\bar{x}\le\bar{y}$，由$\bar{z}=\dfrac{m}{m+n}\bar{x}+\dfrac{n}{m+n}\bar{y}$放缩可得$\bar{x}\le\bar{z}\le\bar{y}$；

对于 B 选项，举例说明 B 不正确；

对于 C 选项，举例说明 C 不正确；

对于 D 选项，若$m=n,\bar{x}=\bar{y}$，代入总体方差计算公式，可得$s^2=\dfrac{s_x^2+s_y^2}{2}$.

【详解】对于 A 选项，因为$\bar{x}\le\bar{y}$，所以$\bar{z}=\dfrac{m}{m+n}\bar{x}+\dfrac{n}{m+n}\bar{y}\le\dfrac{m}{m+n}\bar{y}+\dfrac{n}{m+n}\bar{y}=\bar{y}$

$\bar{z}=\dfrac{m}{m+n}\bar{x}+\dfrac{n}{m+n}\bar{y}\ge\dfrac{m}{m+n}\bar{x}+\dfrac{n}{m+n}\bar{x}=\bar{x}$，即$\bar{x}\le\bar{z}\le\bar{y}$，A 正确；

对于 B 选项，取第一部分数据为$1,1,1,1,1$，则$\bar{x}=1$，$s_x^2=0$，取第二部分数据为$-3,9$，则$\bar{y}=3$，

$s_y^2=36$，则$\bar{z}^2=(\dfrac{5}{7}\times1+\dfrac{2}{7}\times3)^2=\dfrac{121}{49}<3=\bar{x}\cdot\bar{y}$，B 不正确；

对于 C 选项，取第一部分数据为$-2,-1,0,1,2$，则$\bar{x}=0$，$s_x^2=2$，

取第二部分数据为$1,2,3,4,5$，则$\bar{y}=3$，$s_y^2=2$，则$\bar{z}=\dfrac{m}{m+n}\bar{x}+\dfrac{n}{m+n}\bar{y}=\dfrac{5}{10}\times0+\dfrac{5}{10}\times3=\dfrac{3}{2}$，

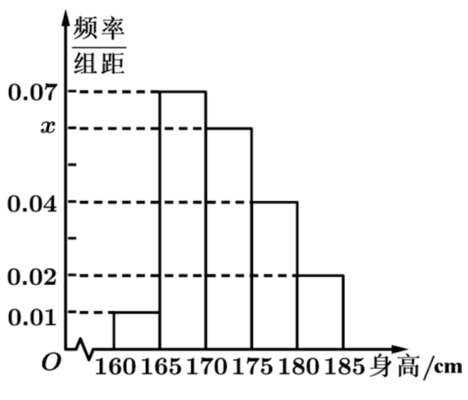
$s^2=\dfrac{m}{m+n}\left[s_x^2+(\bar{x}-\bar{z})^2\right]+\dfrac{n}{m+n}\left[s_y^2+(\bar{y}-\bar{z})^2\right]=\dfrac{5}{10}(2+\dfrac{9}{4})+\dfrac{5}{10}(2+\dfrac{9}{4})=\dfrac{17}{4}>2=s_y^2$，C 不正确；

对于 D 选项，若$m=n,\bar{x}=\bar{y}$，则$\bar{z}=\bar{x}=\bar{y}$

$s^2=\dfrac{m}{m+n}\left[s_x^2+(\bar{x}-\bar{z})^2\right]+\dfrac{n}{m+n}\left[s_y^2+(\bar{y}-\bar{z})^2\right]=\dfrac{s_x^2+s_y^2}{2}$，D 正确.

故选：AD.

2．随机抽取 100 名学生，测得他们的身高（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的值及身高在$170cm$及以上的学生人数；

（2）估计该校 100 名生学身高的 75％分位数.

【详解】(1) 由频率分布直方图可知$5\times(0.01+0.07+x+0.04+0.02+0.01)=1$，解得$x=0.06$，

身高在$170cm$及以上的学生人数$100\times5\times\left(0.06+0.04+0.02\right)=60$（人）.

（2）$\left[180,185\right]$的人数占比为$5\times0.02=10\%$，

$\left[175,180\right]$的人数占比为$5\times0.04=20\%$，

所以该校 100 名生学身高的 75％分位数落在$\left[175,180\right]$，

设该校 100 名生学身高的 75％分位数为x，

则$0.04\left(180-x\right)+0.1=25\%$，解得$x=176.25$，

故该校 100 名生学身高的 75％分位数为176.25.

3.非线性经验回归

当经验回归方程并非形如$y=bx+a$（$a,b\in R$）时，称之为非线性经验回归方程，当两个变量不呈线性相关关系时，依据样本点的分布选择合适的曲线方程来模拟，常见的非线性经验回归方程的转换方式总结如下：

曲线方程	变换公式	变换后的线性关系式
$y=ax^b$	$c=\ln a,v=\ln x,u=\ln y$	$u=c+bv$
$y=ae^{bx}$	$c=\ln a,u=\ln y$	$u=c+bx$
$y=ae^{\frac{b}{x}}$	$c=\ln a,v=\frac{1}{x},u=\ln y$	$u=c+bv$
$y=a+b\ln x$	$v=\ln x$	$y=a+bv$
$y=a+b\sqrt{x}$	$v=\sqrt{x},u=y$	$u=a+bv$