分层抽样中样本方差的求解探索

谢新华（福建省莆田第二中学　351131）

【摘　要】　高中数学 2019 新人教版必修第二册统计部分，新增了分层随机抽样中方差的计算问题，求解时需要对方差的公式进行合理变形转化，这对于高一学生来说是难点.本文先以分两层抽样的情况为例，分析总体样本方差求解的策略，在此基础上进一步推广到一般的分层随机抽样，归纳总体样本方差的计算方法，与读者分享.

【关键词】　分层；抽样；方差；变形

我们首先探究分两层随机抽样的情况，假设第一层有 m 个数，分别为 $x_1,x_2,\cdots,x_m$，平均数为 $\bar{x}$，方差为 s_x^2；第二层有 n 个数，分别为 $y_1,y_2,\cdots,y_n$，平均数为 $\bar{y}$，方差为 s_y^2.

则　$\bar{x}=\dfrac{1}{m}\sum\limits_{i=1}^{m}x_i,s_x^2=\dfrac{1}{m}\sum\limits_{i=1}^{m}(x_i-\bar{x})^2$，

$\bar{y}=\dfrac{1}{n}\sum\limits_{i=1}^{n}y_i,s_y^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}(y_i-\bar{y})^2$.

若记合在一起的样本平均数为 $\bar{a}$，方差为 s^2，

则　$\bar{a}=\dfrac{m\bar{x}+n\bar{y}}{m+n}$，

$$s^2=\frac{1}{m+n}\left[\sum_{i=1}^{m}(x_i-\bar{a})^2+\sum_{i=1}^{n}(y_i-\bar{a})^2\right],$$

因为　$(x_i-\bar{a})^2=[(x_i-\bar{x})+(\bar{x}-\bar{a})]^2$

$=(x_i-\bar{x})^2+(\bar{x}-\bar{a})^2+2(\bar{x}-\bar{a})(x_i-\bar{x})$，

$$\sum_{i=1}^{m}2(\bar{x}-\bar{a})(x_i-\bar{x})$$

$$=2(\bar{x}-\bar{a})\cdot\sum_{i=1}^{m}(x_i-\bar{x})=0,$$

所以　$\sum\limits_{i=1}^{m}(x_i-\bar{a})^2$

$=\sum\limits_{i=1}^{m}(x_i-\bar{x})^2+\sum\limits_{i=1}^{m}(\bar{x}-\bar{a})^2$

$=ms_x^2+m(\bar{x}-\bar{a})^2$，

同理 $\sum\limits_{i=1}^{n}(y_i-\bar{a})^2=\sum\limits_{i=1}^{n}(y_i-\bar{x})^2+\sum\limits_{i=1}^{n}(\bar{y}-\bar{a})^2$

$=ns_y^2+n(\bar{y}-\bar{a})^2$，

所以总体样本的方差为

$$s^2=\frac{1}{m+n}[ms_x^2+m(\bar{x}-\bar{a})^2+ns_y^2+n(\bar{y}-\bar{a})^2]$$

$$=\frac{1}{m+n}\left[ms_x^2+ns_y^2+\frac{mn}{m+n}(\bar{x}-\bar{y})^2\right],$$

另一方面 $s_x^2=\dfrac{1}{m}\sum\limits_{i=1}^{m}(x_i-\bar{x})^2=\dfrac{1}{m}\sum\limits_{i=1}^{m}x_i^2-\bar{x}^2$，

$s_y^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}(y_i-\bar{y})^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}y_i^2-\bar{y}^2$，

所以总体样本的方差为

$$s^2=\frac{1}{m+n}\left(\sum_{i=1}^{m}x_i^2+\sum_{i=1}^{n}y_i^2\right)-\bar{a}^2$$

$$=\frac{1}{m+n}(ms_x^2+m\bar{x}^2+ns_y^2+n\bar{y}^2)-\bar{a}^2.$$

例 1　在了解全校学生每年平均阅读了多少本文学经典名著时，甲同学抽取了一个容量为 10 的样本，并算得样本的平均数为 5，方差为 9；乙同学抽取了一个容量为 8 的样本，并算得样本的平均数为 6，方差为 16.已知甲、乙两同学抽取的样本合在一起组成一个容量为 18 的样本，求合在一起后的样本均值与样本方差.(精确到 0.1)

解法 1　设抽到甲的一组样本数为 $x_1,x_2,\cdots,x_{10}$，抽到乙的一组样本数为 $y_1,y_2,\cdots,y_8$，

由题意知　$\bar{x}=\dfrac{x_1+x_2+\cdots+x_{10}}{10}=5$，

$\bar{y}=\dfrac{y_1+y_2+\cdots+y_8}{8}=6$，

所以　s_1^2

$=\dfrac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_{10}-\bar{x})^2}{10}$

$=9$，

$s_2^2=\dfrac{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_8-\bar{y})^2}{8}$

$=16$，

两个样本合在一起后的样本均值为

$\bar{z}=\dfrac{10\bar{x}+8\bar{y}}{18}=\dfrac{10\times5+8\times6}{10+8}\approx5.4$，

总体样本方差为

$$s^2=\frac{1}{18}[10s_1^2+10(\bar{x}-\bar{z})^2+8s_2^2+8(\bar{y}-\bar{z})^2]$$

$$=\frac{1}{18}\left[10s_1^2+10\left(\bar{x}-\frac{10\bar{x}+8\bar{y}}{18}\right)^2+8s_2^2\right]+$$

$$\frac{1}{18}\times 8\left(\bar{y}-\frac{10\bar{x}+8\bar{y}}{18}\right)^2$$

$$=\frac{1}{18}\left[10s_1^2+8s_2^2+\frac{10\times 8}{18}(\bar{x}-\bar{y})^2\right]$$

$$=\frac{1}{10+8}\times$$

$$\left[(10\times 9+8\times 16)+\frac{10\times 8}{18}(5-6)^2\right]$$

$$=\frac{1}{18}\times\left(218+\frac{40}{9}\right)\approx 12.36.$$

解法 2　设抽到甲的一组样本数为 $x_1,x_2,\cdots,x_{10}$，抽到乙的一组样本数为 $y_1,y_2,\cdots,y_8$，

由题意知　$\bar{x}=\dfrac{x_1+x_2+\cdots+x_{10}}{10}=5$，

$$\bar{y}=\frac{y_1+y_2+\cdots+y_8}{8}=6,$$

$$s_1^2=\frac{1}{10}(x_1^2+x_2^2+\cdots+x_{10}^2)-\bar{x}^2=9,$$

$$s_2^2=\frac{1}{8}(y_1^2+y_2^2+\cdots+y_8^2)-\bar{y}^2=16,$$

两个样本合在一起后的样本均值为

$$\bar{z}=\frac{10\bar{x}+8\bar{y}}{18}=\frac{10\times 5+8\times 6}{10+8}\approx 5.4,$$

总体样本方差为

$$s^2=\frac{1}{18}(10s_1^2+8s_2^2)-\bar{z}^2$$

$$=\frac{1}{18}(10\times 9+8\times 16)-5.4^2\approx 12.36.$$

例 2　已知总体划分为 3 层，通过分层随机抽样，各层抽取的样本量、样本均值和样本方差分别为：$l,\bar{x},s_1^2$；$m,\bar{y},s_2^2$；$n,\bar{z},s_3^2$.记总体的样本均值为 $\bar{w}$，样本方差为 s^2.求证：

$$(1)\bar{w}=\frac{l}{l+m+n}\bar{x}+\frac{m}{l+m+n}\bar{y}+$$

$$\frac{n}{l+m+n}\bar{z};$$

$$(2)s^2=\frac{1}{l+m+n}\{l[s_1^2+(\bar{x}-\bar{w})^2]+$$

$$m[s_2^2+(\bar{y}-\bar{w})^2]+n[s_3^2+(\bar{z}-\bar{w})^2]\}.$$

证明　(1) 总体的样本均值为

$$\bar{w}=\frac{l\bar{x}+m\bar{y}+n\bar{z}}{l+m+n}$$

$$=\frac{l}{l+m+n}\bar{x}+\frac{m}{l+m+n}\bar{y}+\frac{n}{l+m+n}\bar{z}.$$

$$(2)s^2=\frac{1}{l+m+n}\sum_{i=1}^{l}(x_i-\bar{w})^2+$$

$$\frac{1}{l+m+n}\sum_{j=1}^{m}(y_j-\bar{w})^2+$$

$$\frac{1}{l+m+n}\sum_{k=1}^{n}(z_k-\bar{w})^2$$

$$=\frac{1}{l+m+n}\sum_{l}^{i=1}(x_i-\bar{x}+\bar{x}-\bar{w})^2+$$

$$\frac{1}{l+m+n}\sum_{m}^{j=1}(y_j-\bar{y}+\bar{y}-\bar{w})^2+$$

$$\frac{1}{l+m+n}\sum_{n}^{k=1}(z_k-\bar{z}+\bar{z}-\bar{w})^2.$$

因为　$\sum_{i=1}^{l}(x_i-\bar{x})=\sum_{i=1}^{l}x_i-l\bar{x}=0$，

所以　$\sum_{i=1}^{l}2(x_i-\bar{x})(\bar{x}-\bar{w})$

$$=2(\bar{x}-\bar{w})\sum_{i=1}^{l}(x_i-\bar{x})=0.$$

同理　$\sum_{j=1}^{m}2(y_j-\bar{y})(\bar{y}-\bar{w})=0$，

$$\sum_{k=1}^{n}2(z_k-\bar{z})(\bar{z}-\bar{w})=0.$$

所以总体的方差为

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

$$\frac{1}{l+m+n}[\sum_{m}^{j=1}(y_j-\bar{y})^2+\sum_{m}^{j=1}(\bar{y}-\bar{w})^2]+$$

$$\frac{1}{l+m+n}[\sum_{n}^{k=1}(z_k-\bar{z})^2+\sum_{n}^{k=1}(\bar{z}-\bar{w})^2]$$

$$=\frac{1}{l+m+n}\cdot l[s_1^2+(\bar{x}-\bar{w})^2]+$$

$$\frac{1}{l+m+n}\cdot m[s_2^2+(\bar{y}-\bar{w})^2]+$$

$$\frac{1}{l+m+n}\cdot n[s_3^2+(\bar{z}-\bar{w})^2]\}.$$

推广　在分层抽样时，如果总体分为 k 层，而且第 j 层抽取的样本量为 n_j，第 j 层的样本均值为 $\bar{x}_j$，样本方差为 s_j^2，$j=1,2,\cdots,k$.记 $n=\sum_{j=1}^{k}n_j$.

求证：总体的样本均值和方差分别为

$$\bar{x}=\frac{1}{n}\sum_{j=1}^{k}(n_j\bar{x}_j),s^2=\frac{1}{n}\sum_{j=1}^{k}[n_js_j^2+n_j(\bar{x}_j-\bar{x})^2].$$

证明　因为第 j 层抽取的样本均值为

（下转第 29 页）

AB 过定点$\left(\dfrac{kx_0-2y_0}{k},\dfrac{2p-ky_0}{k}\right)$；

(2)$k_{PA}k_{PB}=k(k\neq 0)$ 的充要条件是直线 AB 过定点$\left(\dfrac{kx_0-2p}{k},-y_0\right)$.

4　应用拓展

例 3　设椭圆 $C:\dfrac{x^2}{2}+y^2=1$ 的右焦点为 F，过 F 的直线 l 与 C 交于 A,B 两点，点 M 的坐标为 $(2,0)$.设 O 为坐标原点.证明：$\angle OMA=\angle OMB$.

(2018 年全国 Ⅰ 卷)

证明　将坐标原点平移到点 M，则有

$$\begin{cases}x=x'+2,\\ y=y'.\end{cases}$$

设直线 $A'B'$ 在新坐标系下的方程为

$$y'=m(x'+1),$$

即
$$\frac{y'}{m}-x'=1,$$

椭圆方程为
$$\frac{(x'+2)^2}{2}+y'^2=1,$$

$$x'^2+2y'^2+4'+2=0.$$

联立椭圆与直线方程得

$$x'^2+2y'^2+4x'\left(\frac{y'}{m}-x'\right)+2\left(\frac{y'}{m}-x'\right)^2=0,$$

上式两边同时除以 x'^2，并记 $k'=\dfrac{y'}{x'}$ 得

$$\frac{2m^2+2}{m^2}k'^2-1=0,$$

方程的两个根即为直线 MA,MB 的斜率.

所以
$$k_{MA}+k_{MB}=k_{M'A'}+k_{M'B'}=0,$$

即两直线的倾斜角互补，

故
$$\angle OMA=\angle OMB.$$

例 4　设 F 是椭圆 $E:\dfrac{x^2}{16}+\dfrac{y^2}{12}=1$ 的左焦点，过点 $P(-8,0)$ 的直线交椭圆于不同的两点 A,B，求证 $k_{FA}+k_{FB}$ 为定值.

证明　将坐标原点平移到点 F，

则有
$$\begin{cases}x=x'-2,\\ y=y'.\end{cases}$$

设直线 $A'B'$ 在新坐标系下的方程为

$$x'=my'-6,$$

即
$$\frac{my'}{6}-\frac{x'}{6}=1,$$

椭圆方程为
$$\frac{(x'-2)^2}{16}+\frac{y'^2}{12}=1,$$

$$3x'^2+4y'^2-12x'-36=0.$$

联立椭圆与直线方程得

$$3x'^2+4y'^2-12x'\left(\frac{my'}{6}-\frac{x'}{6}\right)-36\left(\frac{my'}{6}-\frac{x'}{6}\right)^2=0,$$

$$4x'^2+(4-m^2)y'^2=0,$$

上式两边同时除以 x'^2，并记 $k'=\dfrac{y'}{x'}$ 得

$$(4-m^2)k'^2+4=0,$$

方程的两个根即为直线 FA,FB 的斜率.

从而
$$k_{FA}+k_{FB}=k_{F'A'}+k_{F'B'}=0$$

为定值.

(上接第 26 页)

$$\bar{x}_j=\frac{1}{n}\sum_{j=1}^{nj}x_j,$$

所以
$$\frac{1}{n}\sum_{j=1}^{nj}x_j=n_j\bar{x}_j,$$

总体的样本均值为

$$\bar{x}=\frac{1}{n}\sum_{j=1}^{k}\left(\sum_{j=1}^{nk}x_j\right)=\frac{1}{n}\sum_{j=1}^{k}(n_j\bar{x}_j).$$

又因为第 j 层抽取的样本方差为

$$s_j^2=\frac{1}{n_j}\sum_{j=1}^{nj}(x_j-\bar{x}_j)^2,$$

所以
$$\sum_{j=1}^{nj}(x_j-\bar{x}_j)^2=n_js_j^2,$$

因为
$$\sum_{j=1}^{nj}(x_j-\bar{x})^2$$

$$=\sum_{j=1}^{nj}(x_j-\bar{x}_j+\bar{x}_j-\bar{x})^2$$

$$=\sum_{j=1}^{nj}(x_j-\bar{x}_j)^2+2(x_j-\bar{x})\sum_{j=1}^{nj}(x_j-\bar{x}_j)+n_j(\bar{x}_j-\bar{x})^2,$$

$$\sum_{j=1}^{nj}(x_j-\bar{x}_j)=\sum_{j=1}^{nj}x_j-n_j\bar{x}_j=n_j\bar{x}_j-n_j\bar{x}_j=0,$$

所以
$$\sum_{j=1}^{nj}(x_j-\bar{x})^2=n_js_j^2+n_j(\bar{x}_j-\bar{x})^2,$$

故总体的样本方差为

$$s^2=\frac{1}{n}\sum_{j=1}^{k}\left[\sum_{j=1}^{nj}(x_j-\bar{x})^2\right]$$

$$=\frac{1}{n}\sum_{j=1}^{k}x_j[n_js_j^2+n_j(\bar{x}_j-\bar{x})^2].$$