统计中一个恒等式的应用

马德锦

(江苏省兴化市周庄高级中学,225711)

一、问题的提出

在统计内容的学习中,不少同学往往只记住方差、线性回归系数和相关系数的原始公式,而不能根据题目中给出的具体数据条件对原始的公式进行适当的变形,以充分利用题目中的数据条件简化运算.

事实上,在方差、分层抽样方差、线性回归系数和相关系数等统计计算中,都涉及到 $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ 的变形公式,即恒等式 $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}$,其中 $\bar{x},\bar{y}$ 分别是数据 $x_i,y_i(i=1,2,3,\cdots,n)$ 的平均数.

本文系统研究这个恒等式,给出其证明并举例说明其应用,以期帮助同学们灵活掌握方差、线性回归系数和相关系数等统计计算,提升自身的数学抽象、数学运算、数据分析等数学核心素养.

二、恒等式的证明

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}(x_iy_i-\bar{x}y_i-\bar{y}x_i+\bar{x}\bar{y})=\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}\bar{x}y_i-\sum_{i=1}^{n}\bar{y}x_i+\sum_{i=1}^{n}\bar{x}\bar{y}=\sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i-\bar{y}\sum_{i=1}^{n}x_i+n\bar{x}\bar{y}=\sum_{i=1}^{n}x_iy_i-\bar{x}(n\bar{y})-\bar{y}(n\bar{x})+n\bar{x}\bar{y}=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}.$$ 得证.

恒等式还可进一步变形为 $\frac{1}{n}\left[n\sum_{i=1}^{n}x_iy_i-(n\bar{x})(n\bar{y})\right]=\frac{1}{n}\left(n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\right)$.

三、恒等式的应用

1. 计算方差

设数据 $x_i(i=1,2,3,\cdots,n)$ 的平均数、方差分别为 $\bar{x},S^2$,则方差 $S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2$.

证明 在上述恒等式中,令 $x_i=y_i(i=1,2,3,\cdots,n)$,则 $\bar{y}=\bar{x}$,进而 $S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2$.

评注 这个方差的变形公式可记忆为:方差等于每个数据平方的平均减去这组数据平均的平方. 其推广形式为:离散随机变量方差公式 $D(X)=E(X^2)-(EX)^2$.

例 1 某学校从在校学生中,用分层抽样的方法抽取男生 m 人,女生 n 人. 测得他们身高后,计算得到男生身高样本均值为 $\bar{y}_{男}$ cm,方差为 $s_{男}^2$ cm^2;女生身高样本均值为 $\bar{z}_{女}$ cm,方差为 $s_{女}^2$ cm^2. 求 $m+n$ 个身高数据样本均值和方差[1].

解 设男生样本为 $y_1,\cdots,y_m$,女生样本为 $z_1,\cdots,z_n$,所有数据样本均值为 $\bar{x}_{总}$,方差为 $s_{总}^2$,则 $\bar{x}_{总}=\frac{m\bar{y}_{男}+n\bar{z}_{女}}{m+n}=\frac{m}{m+n}\bar{y}_{男}+\frac{n}{m+n}\bar{z}_{女}$.

根据方差的变形公式,得 $s_{总}^2=\frac{1}{m+n}\left(\sum_{i=1}^{m}y_i^2+\sum_{j=1}^{n}z_j^2\right)-\bar{x}_{总}^2$. 又 $s_{男}^2=\frac{1}{m}\sum_{i=1}^{m}y_i^2-\bar{y}_{男}^2$,$s_{女}^2=\frac{1}{n}\sum_{j=1}^{n}z_j^2-\bar{z}_{女}^2$,所以

$$s_{总}^2=\frac{1}{m+n}[m(s_{男}^2+\bar{y}_{男}^2)+n(s_{女}^2+\bar{z}_{女}^2)]-\bar{x}_{总}^2. \quad ①$$

评注 可以证明这里得到的分层抽样方差公式① 与以前学习过的分层抽样方差公式

$s_{总}^2=\frac{m}{m+n}[s_{男}^2+(\bar{y}_{男}-\bar{x}_{总})^2]+\frac{n}{m+n}[s_{女}^2+(\bar{z}_{女}-\bar{x}_{总})^2]$（参见[2]）是等价的.

由此,我们可得到分层抽样的方差计算新公式:设样本中不同层的平均数分别为 $\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_n$,方差分别为 $s_1^2,s_2^2,\cdots,s_n^2$,权重分别为 $p_1,p_2,\cdots,p_n$,则这个样本的方差为 $s^2=\sum_{i=1}^{n}p_i(s_i^2+\bar{x}_i^2)-\bar{x}^2$,$\bar{x}$ 为总样本数据的平均数.(请同学们与以前学习的分层抽样的方差计算公式 $s^2=\sum_{i=1}^{n}p_i[s_i^2+(\bar{x}_i-\bar{x})^2]$ 进行对照、比较)

例 2　现有甲、乙两支田径队,甲队的体重的平均值为 60 kg,方差为 200 kg^2;乙队体重的平均值为 70 kg,方差为 300 kg^2. 又已知甲、乙两队的队员人数之比为 1∶4,求甲、乙两队全部队员体重的平均值和方差.

解　由题意可知 $\bar{x}_{甲}=60$,甲队队员在所有队员中所占权重为 $\frac{1}{1+4}=\frac{1}{5}$;$\bar{x}_{乙}=70$,乙队队员在所有队员中所占权重为 $\frac{4}{1+4}=\frac{4}{5}$.

故甲、乙两队全部队员体重的平均值为 $\bar{x}=\frac{1}{5}\times60+\frac{4}{5}\times70=68$,甲、乙两队全部队员的体重的方差为 $s^2=\frac{1}{5}\times(200+60^2)+\frac{4}{5}\times(300+70^2)-68^2=296$.

或者甲、乙两队全部队员的体重的方差为 $s^2=\frac{1}{5}\times[200+(60-68)^2]+\frac{4}{5}\times[300+(70-68)^2]=296$.

2. 计算线性回归系数 $\hat{b}$

$$\hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}.$$

例 3　某小区物业公司统计了近六年小区私家车的数量,以编号 1 对应 2015 年,编号 2 对应 2016 年,编号 3 对应 2017 年,以此类推,得到相应数据如下:

年份编号 x	1	2	3	4	5	6
数量 y/辆	41	96	116	190	218	275

求该小区私家车的数量 y 关于年份编号 x 的线性回归方程.

附:$\hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$,$\sum_{i=1}^{6}y_i=936$,$\sum_{i=1}^{6}x_iy_i=4\,081$,$\sum_{i=1}^{6}x_i^2=91$.

解　因为 $\bar{x}=\frac{1}{6}(1+2+3+4+5+6)=3.5$,$\bar{y}=\frac{1}{6}\times936=156$,所以 $\hat{b}=\frac{\sum_{i=1}^{6}x_iy_i-6\bar{x}\bar{y}}{\sum_{i=1}^{6}x_i^2-6\bar{x}^2}=\frac{4\,081-6\times3.5\times156}{91-6\times3.5^2}=46$,

$\hat{a}=\bar{y}-\hat{b}\bar{x}=156-46\times4.5=-5$.

所以,y 关于 x 的线性回归方程为 $\hat{y}=46x-5$.

3. 计算相关系数 r

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}\cdot\sqrt{\sum_{i=1}^{n}y_i^2-n\bar{y}^2}}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}\cdot\sqrt{n\sum_{i=1}^{n}y_i^2-(\sum_{i=1}^{n}y_i)^2}}.$$

例 3　我国风云系列卫星可以监测气象和国土资源情况. 某地区水文研究人员为了解汛期人工测雨量 x(单位:dm)与遥测雨量 y(单位:dm)的关系,统计得到该地区 10 组雨量数据如下:

样本号 i	1	2	3	4	5	6	7	8	9	10
人工测雨量 x_i	5.38	7.99	6.37	6.71	7.53	5.53	4.18	4.04	6.02	4.23
遥测雨量 y_i	5.43	8.07	6.57	6.14	7.95	5.56	4.27	4.15	6.04	4.49

并计算得 $\sum_{i=1}^{10} x_i^2 = 353.6$，$\sum_{i=1}^{10} y_i^2 = 361.7$，$\sum_{i=1}^{10} x_i y_i = 357.3$，$\bar{x}^2 \approx 33.62$，$\bar{y}^2 \approx 34.42$，$\bar{x}\bar{y} \approx 34.02$. 试求该地区汛期遥测雨量 y 与人工测雨量 x 的样本相关系数(精确到 0.01)，并判断它们是否具有线性相关关系.

附：$\sqrt{304.5} \approx 17.4$，相关系数

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_1 - \bar{x})^2 \sum_{i=1}^{n}(y_1 - \bar{y})^2}}.$$

解 根据

$$r = \frac{\sum_{i=1}^{10} x_i y_i - 10\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2)(\sum_{i=1}^{10} y_i^2 - 10\bar{y}^2)}},$$

代入已知数据，得

$$r = \frac{357.3 - 340.2}{\sqrt{(353.6 - 336.2)(361.7 - 344.2)}}$$

$$= \frac{17.1}{\sqrt{304.5}} \approx 0.98.$$

所以汛期遥测雨量 y 与人工测雨量 x 有很强的线性相关关系.

四、一点启示

数学学习要全面. 例如对数学公式的学习，不但要理解公式的由来，记住原始形式，还要掌握公式的各种变形，以便根据题设条件灵活运用公式，简洁迅速解决问题. 用联系的观点学习数学，进行反思性学习. 对所学数学知识、方法要进行系统梳理，运用对照、比较等思维方法，找出它们之间的相同点与不同点，进行抽象概括，揭示数学对象之间内在的和本质的联系，可以优化认知结构，提升学习效益，切实减轻学习负担.

参考文献

[1] 单墫，李善良主编. 普通高中教科书数学(必修第二册)[M]. 南京：江苏凤凰教育出版社，2020，1：238－239.

[2] 中华人民共和国教育部. 普通高中数学课程标准(2017 年版 2020 年修订)[S]. 北京：人民教育出版社，2020：126－129.

(上接第 49 页)

在图3中，由 $AB = 3$，$BF = 1$，得 $\angle BAF = \alpha_2$；由 $AE = 1$，得 $\angle BEF = \alpha_4$；由 $FC = 6$，$CD = 3$，得 $\angle EDA = \alpha_3$，$\angle CFD = \angle BEF = \alpha_4$. 又易知 $\angle EFD = 90°$，而 $\angle EAD = 90°$，所以 D，F，E，A 四点共圆. 故 $\angle EDA = \angle EFA = \alpha_4 - \alpha_2$，即 ② 式成立.

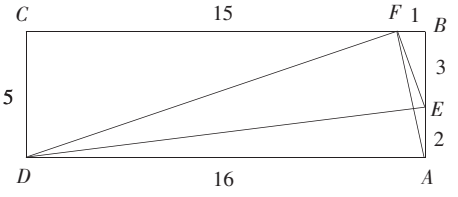

图 4

同样，在图 4 中，由 $AB = 5$，$BF = 1$，得 $\angle BAF = \alpha_1$；由 $AE = 2$，$BE = 5 - 2 = 3$，得 $\angle BEF = \alpha_6$；由 $FC = 15$，$CD = 5$，得 $\angle EDA = \alpha_5$，且 $\angle CFD = \angle BEF = \alpha_6$. 又计算易知 $\angle EFD = 90°$，而 $\angle EAD = 90°$，所以 D，F，E，A 四点共圆. 所以 $\angle EDA = \angle EFA = \alpha_6 - \alpha_1$，即 ③ 式成立.

注解 ① 式更常见的证法是如图 5，$\triangle ABC$ 为等腰直角三角形，而 $\angle CBE = \angle BAE = \alpha_4$，且 $\angle CAE = \alpha_6$，所以 $\alpha_4 + \alpha_6 = \angle BAC = 45°$.

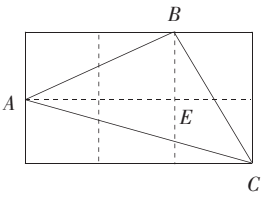

图 5