关于样本方差S^2计算的探讨

唐婷婷

（武警警官学院基础部工程数学教研室 四川 成都 610000）

[摘 要] 本文通过例题解析的形式对样本方差S^2进行计算，在应用同一公式的不同形式$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$和$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$计算样本方差$S^2$时产生两个不同的结果，通过比较分析，找到误差产生较小的计算公式。

[关键词] 样本方差；计算；误差

1.知识准备

设X_1，X_2，...X_n是来自于总体X的一个样本，定义样本平均值$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$；样本方差$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$

2.例题解析

例1 下面是使用铂球测定引力常数（单位：$10^{-11}m^3\cdot kg^{-1}\cdot s^{-2}$）的观察值为

6.661 6.661 6.667 6.667 6.664

（解析）首先将观察值代入样本平均值$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$得$\bar{x}$=6.664，再代入样本方差$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$得$s^2=0.9\times10^{-5}$，同理将观察值代入$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$得$s^2=0.9\times10^{-5}$。

例2 下面是使用金球测定引力常数（单位：$10^{-11}m^3\cdot kg^{-1}\cdot s^{-2}$）的观察值为

6.683 6.681 6.676 6.678 6.679 6.672

（解析）首先将观察值代入样本平均值$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$得，$\bar{x}\approx6.678$

再代入样本方差$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$得$s^2=1.5\times10^{-5}$，同理将观察值代入$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$得$s^2\approx2.686\times10^{-3}$。

3.总结

综上例1，例2针对样本方差，我们采用不同形式的公式$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$和$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$计算出来的结果例1是一致的，例2是不一致的（严格上应该为一致，因为它们是相等的）。那么在例2的样本方差计算中为什么会产生不一致，是什么原因造成这样的结果？并且两个结果中哪个结果更好更精确？

在例1中计算出的的样本平均值$\bar{x}$=6.664，是精确的，无误差的，分别代入样本方差两个公式中计算，由于整个过程都没有误差产生，所以计算的结果都是一致的。而在例2中计算出的样本平均值$\bar{x}\approx6.678$，是近似值，有误差的，分别代入样本方差两个公式中，最后导致样本方差也是有误差的，从而两个公式计算出的样本方差结果不一致，接下来我们比较一下哪个公式的计算结果误差较小。

假设$\bar{X}'=\bar{X}+\delta$，其中$\bar{X}$是准确值，$\bar{X}'$是$\bar{X}$的近似值，且具有限位有效数字，δ是误差。针对具体问题，分别采用公式$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$和$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$计算，得到两个样本方差$S_1^2$和$S_2^2$分别如下

$$S_1^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}')^2=\frac{1}{n-1}\sum_{i=1}^{n}\left[X_i-(\bar{X}+\delta)\right]^2$$
$$=\frac{1}{n-1}\sum_{i=1}^{n}\left[(X_i-\bar{X})^2+2\delta(X_i-\bar{X})+\delta^2\right]$$
$$=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2+\frac{n\delta^2}{n-1}$$

$$S_2^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}'^2\right)=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n(\bar{X}+\delta)^2\right)$$
$$=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2-2n\delta\bar{X}-n\delta^2\right)$$
$$=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)-\frac{2n\delta}{n-1}\bar{X}-\frac{n\delta^2}{n-1}$$

其中$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$，是样本方差精确值。

所以用公式$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$计算出的$S_1^2$产生误差为$\frac{n\delta^2}{n-1}$，用公式$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$计算出的$S_2^2$产生误差为$\frac{2n\delta}{n-1}\bar{X}+\frac{n\delta^2}{n-1}$。$\delta$是样本平均值$\bar{X}$的误差，理应很小，所以误差$\frac{n\delta^2}{n-1}$也就很小，而误差$\frac{2n\delta}{n-1}\bar{X}+\frac{n\delta^2}{n-1}$由于$\bar{X}$比$\delta$大很多，所以整个误差主要由$\frac{2n\delta}{n-1}\bar{X}$决定，加上$\bar{X}$显然比$\delta$大很多，因此最后误差$\left|\frac{2n\delta}{n-1}\bar{X}+\frac{n\delta^2}{n-1}\right|$比$\frac{n\delta^2}{n-1}$大很多。从上面例2来看，使用$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$计算出的$s^2=1.5\times10^{-5}$，同理使用$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$计算出的$s^2\approx2.686\times10^{-3}$，而我们用计算机算出的$s^2=1.5\times10^{-5}$，显然$s^2\approx2.686\times10^{-3}$误差很大，严重失真，而$s^2=1.5\times10^{-5}$误差很小，是合理的。从而可知，用公式$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$比$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$产生的误差小。

4.结束语

综上所述，如果代入的样本平均值$\bar{X}$是准确值（不是近似值），那么公式$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$和$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$计算出的样本方差都是一致的。如果代入的样本平均值$\bar{X}$是有效近似值，那么公式$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$和$S^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)$相比，公式$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$计算的样本方差产生误差较小，因此为了减少计算误差，建议使用$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$来计算样本方差$S^2$。

参考文献

[1]概率论与数理统计（第四版）浙江大学 盛骤，谢式千，潘承毅编，北京：高等教育出版社.

[2]数值分析（第五版）李庆扬，王能超，易大义编，北京：清华大学出版社.