2009年5月20日作成

2011年4月18日修正

典型的な統計モデルにおける標本分散の誤差

Errors of sample variance in typical statistical models

井田　隆

名古屋工業大学セラミックス基盤工学研究センター

はじめに

標本分散の誤差の評価法については別のノート「標本分散の誤差の評価」に詳述しています。平均 m, 標準偏差 σ の統計分布を持つ任意の n 個の標本値 $\{X_j\}$ ($j = 0, ..., n - 1$) に対して，観測された実験データから標本平均，標本分散がそれぞれ

$$\bar{X} \equiv \frac{1}{n}\sum_{j=0}^{n-1} X_j$$

$$V = \frac{1}{n-1}\sum_{j=0}^{n-1}\left(X_j - \bar{X}\right)^2$$

で計算されること，標本平均の分散が

$$\left\langle\left(\bar{X} - m\right)^2\right\rangle = \frac{V}{n}$$

という式で評価される事は比較的良く知られていると思われますが，標本分散の分散が

$$\left\langle\left(V - \sigma\right)^2\right\rangle \sim \frac{1}{n^2}\sum_{j=0}^{n-1}\left(X_j - \bar{X}\right)^4 - \frac{V^2}{n}$$

という式で推定されることが示されています。

このノートでは，ポアソン Poisson 分布と正規分布の場合に，標本分散の相対誤差がいずれも $\sqrt{2/n}$ という値をとることを示します。

粉末回折における粒子統計を評価するために注意すべきと思われることは別のノート「粉末回折測定における粒子統計誤差評価」に述べています。

典型的な確率分布モデルの標本分散の分散

期待値 λ のポアソン分布の場合，

$$m = \lambda$$
$$\sigma^2 = \lambda$$

$$\left\langle \left(x-\langle x\rangle\right)^3 \right\rangle = \lambda$$

$$\left\langle \left(x-\langle x\rangle\right)^4 \right\rangle = \lambda + 3\lambda^2$$

の関係があるので，

$$\left\langle \left(V-\sigma^2\right)^2 \right\rangle \sim \frac{\lambda+2\lambda^2}{n} \sim \frac{2\lambda^2}{n}$$

となり，標本分散の相対統計誤差は

$$\frac{\left\langle \left(V-\sigma^2\right)^2 \right\rangle^{1/2}}{\sigma^2} \sim \frac{\sqrt{2/n}\lambda}{\lambda} = \sqrt{\frac{2}{n}}$$

となります。

　正規分布の場合，

$$\left\langle \left(x-\langle x\rangle\right)^4 \right\rangle = 3\sigma^4$$

なので，やはり標本分散の相対統計誤差は

$$\frac{\left\langle \left(V-\sigma^2\right)^2 \right\rangle^{1/2}}{\sigma^2} \sim \frac{\sqrt{3\sigma^4/n-\sigma^4/n}}{\sigma^2} = \sqrt{\frac{2}{n}}$$

となります。

　これらのことは，ポアソン分布あるいは正規分布に従う実験値から分散を相対誤差10% 以内で評価するためには 200 個以上のデータが必要であることを意味します。相対誤差 5% で評価するためには 800 個のデータが必要です。繰り返し測定によって統計的な誤差の大きさとして意味のある数を導くためには，一般的に数百回以上測定を繰り返す必要があると考えられます。