

方差传递公式估计式的无偏性

浦天舒

(东华大学理学院,上海 201620)

摘要 推广正态样本的均值与样本方差相互独立之定理,证明正态样本 (\bar{x}_1, \bar{x}_2) 与其协方差也是相互独立的。如果假定在直接测量中样本独立同正态分布并且随机误差是小量,那么间接测量的方差传递公式的估计式是方差传递公式的无偏估计式。

关键词 方差传递 无偏估计 正态分布

中图法分类号 O212.1; **文献标志码** A

在测量工作中,用 Bessel 公式可以计算一个直接测量列的方差的无偏估计值。但在间接测量时,待测量是由直接测量的量通过计算而得到的,其方差也是由直接测量量的方差通过方差传递公式计算而得到的。实际上只能用方差传递公式的估计式来估算方差,而关于此估计式是否为无偏估计却并不是显而易见的。如有文献指出^[1],如 x 是真实值 μ_x 的无偏估计量, $U = U(x)$ 是 x 的非线性函数,那么一般地 $U(x)$ 是 $U(\mu_x)$ 的有偏估计,并指出偏差主要来自 x 的方差 σ_x^2 和 U 的二阶导数 $\partial^2 U / \partial x^2$,但对于忽略二阶(及以上)导数时方差传递公式的估计式是否为方差传递公式原始形式的无偏估计,一般文献并未给出证明。因而有必要作一个数学上的论证。

1 若干引理

假设随机变量 x_1 和 x_2 分别服从正态分布 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$, $(x_{1i}, x_{2i}) (i = 1, \dots, n)$ 是来自随机向量 (x_1, x_2) 的简单随机样本; x_{1i} 和 $x_{2i} (i = 1, \dots, n)$ 的均值分别为 $\bar{x}_1 = \frac{1}{n} \sum_{i=1}^n x_{1i}$ 和 $\bar{x}_2 = \frac{1}{n} \sum_{i=1}^n x_{2i}$,

则其方差和协方差估计分别为

$$s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2, s_2^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2 \text{ 和}$$

$$s_{12} = \frac{1}{n-1} \sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)。$$

引理 1^[2] 对任意 n 维向量 $t_1 = (t_{11}, \dots, t_{1n})^T$, 如果 $t_1^T t_1 = t_{11}^2 + \dots + t_{1n}^2 = 1$, 则存在一 $n \times n$ 阶正交矩阵 T (即 $TT^T = I_n$, 这里 I_n 是 $n \times n$ 阶单位矩阵。)使 (t_{11}, \dots, t_{1n}) 恰好是它的第一行。

引理 2^[2] 假设随机向量 $\xi = (\xi_1, \dots, \xi_n)^T$ 的各分量相互独立并且都服从标准正态分布; T 是 $n \times n$ 阶正交矩阵。那么 $\eta = T\xi = (\eta_1, \dots, \eta_n)^T$ 的各分量仍然相互独立并且都服从标准正态分布。

由以上两个引理可以证明^[2] \bar{x}_1 和 \bar{x}_2 分别服从正态分布 $N\left(\mu_1, \frac{\sigma_1^2}{n}\right)$ 和 $N\left(\mu_2, \frac{\sigma_2^2}{n}\right)$ 以及 \bar{x}_1 和 s_1^2 、 \bar{x}_2 和 s_2^2 相互独立。只需稍加推广,便可证明:

引理 3 向量 (\bar{x}_1, \bar{x}_2) 和 s_{12} 相互独立。

证明易见 $x_{1i}^* = \frac{x_{1i} - \mu_1}{\sigma_1}$ 和 $x_{2i}^* = \frac{x_{2i} - \mu_2}{\sigma_2} (i = 1, \dots, n)$ 分别服从标准正态分布。容易验证

$$\bar{x}_1^* = \frac{1}{n} \sum_{i=1}^n \frac{x_{1i} - \mu_1}{\sigma_1}, \bar{x}_2^* = \frac{1}{n} \sum_{i=1}^n \frac{x_{2i} - \mu_2}{\sigma_2},$$

$$\frac{(n-1)s_{12}}{\sigma_1\sigma_2} = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sigma_1\sigma_2} =$$

$$\sum_{i=1}^n x_{1i}^* x_{2i}^* - \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{1i}^* \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{2i}^* =$$

然后,由引理1知,存在正交矩阵

$$T = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nn} \end{bmatrix} \quad (1)$$

考虑正交变换

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nn} \end{bmatrix} \begin{bmatrix} x_{11}^* \\ x_{12}^* \\ \vdots \\ x_{1n}^* \end{bmatrix},$$

$$\begin{bmatrix} y_{21} \\ y_{22} \\ \vdots \\ y_{2n} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nn} \end{bmatrix} \begin{bmatrix} x_{21}^* \\ x_{22}^* \\ \vdots \\ x_{2n}^* \end{bmatrix} \quad (2)$$

则由引理2知 y_{11}, \dots, y_{1n} (以及 y_{21}, \dots, y_{2n}) 独立同标准正态分布,且(1)式、(2)式可见

$$y_{11} = \frac{1}{\sqrt{n}} x_{11}^* + \cdots + \frac{1}{\sqrt{n}} x_{1n}^* = \sqrt{n} \bar{x}_1^*, \quad y_{21} = \frac{1}{\sqrt{n}} x_{21}^* + \cdots + \frac{1}{\sqrt{n}} x_{2n}^* = \sqrt{n} \bar{x}_2^* \quad (3)$$

所以 $\bar{x}_1^* = \frac{y_{11}}{\sqrt{n}}, \bar{x}_2^* = \frac{y_{21}}{\sqrt{n}}$ 都服从正态分布 $N(0, 1/n)$,由此可见 $\bar{x}_1 = \sigma_1 \bar{x}_1^* + \mu_1$ 和 $\bar{x}_2 = \sigma_2 \bar{x}_2^* + \mu_2$ 分别服从正态分布 $N(\mu_1, \frac{\sigma_1^2}{n})$ 和 $N(\mu_2, \frac{\sigma_2^2}{n})$ 。

由(2)式还可得

$$\sum_{i=1}^n x_{1i}^* x_{2i}^* = (x_{11}^*, \dots, x_{1n}^*) \begin{pmatrix} x_{21}^* \\ \vdots \\ x_{2n}^* \end{pmatrix} =$$

$$(y_{11}, \dots, y_{1n}) T^T T \begin{pmatrix} y_{21} \\ \vdots \\ y_{2n} \end{pmatrix} = \sum_{i=1}^n y_{1i} y_{2i},$$

所以

$$\frac{(n-1)s_{12}}{\sigma_1 \sigma_2} = \sum_{i=1}^n x_{1i}^* x_{2i}^* - \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{1i}^* \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{2i}^* = \sum_{i=1}^n y_{1i} y_{2i} - y_{11} y_{21} = \sum_{j=2}^n y_{1j} y_{2j} \quad (4)$$

由于 y_{11}, \dots, y_{1n} (以及 y_{21}, \dots, y_{2n}) 相互独立,向量 $(\bar{x}_1 = \sigma_1 \bar{x}_1^* + \mu_1, \bar{x}_2 = \sigma_2 \bar{x}_2^* + \mu_2)$ 仅依赖于向量 (y_{11}, y_{21}) ,而 $s_{12} = \frac{\sigma_1 \sigma_2}{(n-1)} \sum_{j=2}^n y_{1j} y_{2j}$ 仅依赖于 $(y_{12}, y_{22}), \dots, (y_{1n}, y_{2n})$,可见向量 (\bar{x}_1, \bar{x}_2) 和 s_{12} 相互独立。

2 方差传递公式的原始形式

现设间接测量量 y 与直接测量量 x_1, \dots, x_N 的函数关系为

$$y = f(x_1, \dots, x_N) \quad (5)$$

将式(5)在 x_1, \dots, x_N 的期望值 μ_1, \dots, μ_N 附近按 Taylor 级数展开,忽略二阶及以上项,则有

$$y \approx f(\mu_1, \dots, \mu_N) + \sum_{i=1}^N \left(\frac{\partial f}{\partial x_i} \right)_{\mu_1, \dots, \mu_N} (x_i - \mu_i) \quad (6)$$

两边取期望得

$$E(y) \approx f(\mu_1, \dots, \mu_N) \quad (7)$$

代入(6)式有

$$[y - E(y)]^2 \approx \sum_{i=1}^N \left(\frac{\partial f}{\partial x_i} \right)_{\mu_1, \dots, \mu_N}^2 (x_i - \mu_i)^2 + 2 \sum_{i=1}^N \sum_{j=i+1}^{N-1} \left(\frac{\partial f}{\partial x_i} \right)_{\mu_1, \dots, \mu_N} \left(\frac{\partial f}{\partial x_j} \right)_{\mu_1, \dots, \mu_N} (x_i - \mu_i)(x_j - \mu_j),$$

两边取期望,便可得到方差传递公式的原始形式

$$\sigma^2(y) \approx \sum_{i=1}^N \left(\frac{\partial f}{\partial x_i} \right)_{\mu_1, \dots, \mu_N}^2 \sigma^2(x_i) + 2 \sum_{i=1}^N \sum_{j=i+1}^{N-1} \left(\frac{\partial f}{\partial x_i} \right)_{\mu_1, \dots, \mu_N} \left(\frac{\partial f}{\partial x_j} \right)_{\mu_1, \dots, \mu_N} \text{Cov}(x_i, x_j). \quad (8)$$

但实际使用的则是(7)式和(8)式的估计式。

3 方差传递公式的估计式

显然,若(5)式中的 x_1, \dots, x_N 是测得值或样本

均值,因其期望为 μ_1, \dots, μ_N ,故由(6)式知,在将测得值或样本均值代入由(5)式所决定的函数关系式中并作 Taylor 展开时,在忽略二阶及以上项的情况下, y 的期望的估计值亦即(7)式的(近似无偏)估计值可通过把测得值或样本均值代入(5)式计算而得到。

但是,把(8)式中的方差和协方差分别用样本方差 $s^2(x_i)$ 和样本协方差 $s(x_i, x_j)$ 代替,得到的还不是(8)式的估计式,因为还有偏导数计算的问题。显然,若偏导数用测得值或样本均值代入计算,则偏导数作为测得值或样本均值的函数,在忽略高阶

项的情况下,由(6)式可知其期望即为 $\left(\frac{\partial f}{\partial x_i}\right)_{\mu_1, \dots, \mu_N}$ ($i = 1, \dots, N$)。因此,如若我们把偏导数用样本均值代入计算,便可以把方差传递公式的估计式写成

$$\begin{aligned} s^2(y) &\approx \sum_{i=1}^N \left(\frac{\partial f}{\partial x_i} \right)_{\bar{x}_1, \dots, \bar{x}_N}^2 s^2(x_i) + \\ &2 \sum_{i=1}^N \sum_{j=i+1}^{N-1} \left(\frac{\partial f}{\partial x_i} \right)_{\bar{x}_1, \dots, \bar{x}_N} \left(\frac{\partial f}{\partial x_j} \right)_{\bar{x}_1, \dots, \bar{x}_N} s(x_i, x_j) \end{aligned} \quad (9)$$

但是很明显 $s^2(y)$ 一般并不是 $\sigma^2(y)$ 的(近似)无偏估计,除非直接测量量 x_i 的样本[可设为 $x_i = (x_{i1}, \dots, x_{in})$]是独立同正态分布的。因为此时不仅直接测量量 x_i 的样本均值 \bar{x}_i 和样本方差 $s^2(x_i)$ 相互独立,而且由引理 3 知,(\bar{x}_i, \bar{x}_j) 和样本协方差 $s(x_i, x_j)$ 也相互独立;因而由独立随机变量函数(指 Borel 函数)的独立性^[2] 可知, \bar{x}_i, \bar{x}_j 的函数和 $s^2(x_i), s^2(x_j)$ 及 $s(x_i, x_j)$ 相互独立,于是根据数学期望的性质,由(9)式可得

Unbiasedness of the Estimate of Variance Propagation Formula

PU Tian-shu

(Science College, Donghua University, Shanghai 201620, P. R. China)

[Abstract] An extension of the theorem, which states that a normal sample mean and the sample variance are independent, is used in the proof which states that the normal sample (\bar{x}_1, \bar{x}_2) and its covariance are also independent each other. Assume that in a direct measurement a sample is independent and follows normal distribution; and the random error is a small quantity. Then the estimate of variance propagation formula is an unbiased estimate of the variance propagation formula in an indirect measurement.

[Key words] variance propagation unbiased estimate normal distribution

$$\begin{aligned} E[s^2(y)] &\approx \sum_{i=1}^N \left(\frac{\partial f}{\partial x_i} \right)_{\mu_1, \dots, \mu_N}^2 \sigma^2(x_i) + \\ &2 \sum_{i=1}^N \sum_{j=i+1}^{N-1} \left(\frac{\partial f}{\partial x_i} \right)_{\mu_1, \dots, \mu_N} \left(\frac{\partial f}{\partial x_j} \right)_{\mu_1, \dots, \mu_N} \times \\ &\text{Cov}(x_i, x_j) \approx \sigma^2(y) \end{aligned} \quad (10)$$

在实际进行方差估算时,(9)式中偏导数有时也可直接用一次测得的值代入计算,如果测量只进行了一次的话,只要此时估算的方差与测得值相互独立即可。这个条件通常是容易满足的,因为通常测量仪器的仪器误差(限)都是固定的,与测量值无关。

4 结论

从以上分析可以得出如下结论:方差传递公式的估计式(9)是方差传递公式(8)的(近似)无偏估计的条件,一是随机误差(测得值或其平均值与其期望值之差)为小量,使得 Taylor 展开式中的高阶项可以忽略;二是直接测量量的样本 $x_i = (x_{i1}, \dots, x_{in})$ 独立同正态分布,这也应看成是实际测量工作中多次测量的不确定度评定以正态分布为基础的一个重要理由。注意这里并没有对 x_i 和 x_j (如果它们不独立)的联合分布提出限制。

参 考 文 献

- 1 Willian N. Statistics for engineers and scientists. New York: McGraw-Hill, 2006
- 2 周概率. 概率论与数理统计. 北京:高等教育出版社, 1984; 431; 358; 432; 203