Arbitrary Moments

(Pebay 2008) presents a method calculating arbitrary statistical moments of a set of data. However, part of proof is from the paper. This missing steps are showing below.

The *pth* moment of values x in a set D is expressed as $M_{p,D}=\sum_{x\in D}(x-\mu)^p$ for $p\geq 2$ where μ is the mean of the data (or the 1st moment).

Splitting the data into two subsets $D=D_1\cup D_2$, the *pth* moment of D can be expressed as the sum of the *pth* moment of the subsets.

$$M_{p,D}=\sum_{x\in D_1}(x-\mu)^p+\sum_{x\in D_2}(x-\mu)^p$$

The overall mean μ can be expressed in terms of the combined mean of the two subsets.

$\mu=\frac{\mu_1 n_1+\mu_2 n_2}{n}$ Where $n=n_1+n_2$ and n_x is the number of elements in a set.

The expression for the overall mean can be re-written in two ways to an expression where the first term is the mean of one of the subsets.

$$\mu=\frac{\mu_1 n_1+\mu_2 n_2}{n}$$

$$\mu n=\mu_1 n_1+\mu_2 n_2$$

$$\mu n=\mu_1(n-n_2)+\mu_2 n_2$$

$$\mu n=\mu_1 n-\mu_1 n_2+\mu_2 n_2$$

$$\mu n=\mu_1 n+n_2(\mu_2-\mu_1)$$

$$\mu=\mu_1+\frac{n_2(\mu_2-\mu_1)}{n}$$

$$\mu=\frac{\mu_1 n_1+\mu_2 n_2}{n}$$

$$\mu n=\mu_1 n_1+\mu_2 n_2$$

$$-\mu_1 n_1=\mu_2 n_2-\mu n$$

$$\mu_2 n_1-\mu_1 n_1=\mu_2 n_2-\mu n+\mu_2 n_1$$

$$n_1(\mu_2-\mu_1)=\mu_2 n_2-\mu n+\mu_2 n_1$$

$$\mu n=\mu_2 n_2+\mu_2 n_1-n_1(\mu_2-\mu_1)$$

$$\mu n=\mu_2(n_1+n_2)-n_1(\mu_2-\mu_1)$$

$$\mu n=\mu_2 n-n_1(\mu_2-\mu_1)$$

$$\mu=\mu_2-\frac{n_1(\mu_2-\mu_1)}{n}$$

Substituting these values into the expression for the overall *pth* moment gives:

$$M_{p,D}=\sum_{x\in D_1}\left(x-\mu_1+\frac{-n_2(\mu_2-\mu_1)}{n}\right)^p+\sum_{x\in D_2}\left(x-\mu_2+\frac{n_1(\mu_2-\mu_1)}{n}\right)^p$$

Using the binomial expansion $(x+a)^n=\sum_{k=0}^{n}\binom{n}{k}x^k a^{n-k}$ we find:

$$M_{p,D}=\sum_{k=0}^{p}\binom{k}{p}\sum_{x\in D_1}(x-\mu_1)^{p-k}\left(\frac{-n_2(\mu_2-\mu_1)}{n}\right)^k+\sum_{k=0}^{p}\binom{k}{p}\sum_{x\in D_2}(x-\mu_2)^{p-k}\left(\frac{n_1(\mu_2-\mu_1)}{n}\right)^k$$

In this expression the terms $\sum_{x\in D_1}(x-\mu_1)^{p-k}$ $\sum_{x\in D_2}(x-\mu_2)^{p-k}$ are equivalent to M_{p-k,D_1} and

M_{p-k,D_2} . Substituting this into the expression and combining the binomial terms we find:

$$M_{p,D}=\sum_{k=0}^{p}\binom{k}{p}\left[M_{p-k,D_1}\frac{-n_2(\mu_2-\mu_1)^k}{n}+M_{p-k,D_2}\frac{n_1(\mu_2-\mu_1)^k}{n}\right]$$

This means lower-order moments can be combined to calculate higher order moments.

This is the basis of online algorithms for calculating the variance, skewness and kurtosis of data in one pass, as displayed in RunningStats (Cook).

Bibliography

Philippe Pebau, Formulas for Robust, One-Pass Parallel Computation of Covariances and Arbitrary-Order Statistical Moments, 2008
Cook: John Cook, Computing skewness and kurtosis in one pass, ,
http://www.johndcook.com/skewness_kurtosis.html