Subject: Re: VARIANCE in IDL
Posted by landsman on Tue, 23 Feb 1999 08:00:00 GMT
View Forum Message <> Reply to Message

In article <7au5t5$6on$1@jura.cc.ic.ac.uk>, ashmall@my-dejanews.com (Justin Ashmall) writes...
> Dear All,
>
> I have a question regarding the variance as calculated by IDL - I expect to
> get thoroughly flamed by some statistician types but I'm keen to know if I'm
> wrong!
>
> I always thought the definition of variance was the mean of the squares of the
> differences from the mean, i.e.:
>
> VARIANCE = { SUM [(x - mean_x)^2] } / N
>
> and this is what I *thought* I was getting from IDL - it wasn't until I was
> testing a prog to calculate the means and variances of rows and columns of an
> array that I spotted that IDL's variance has N-1 as the denominator:
>
> VARIANCE = { SUM [(x - mean_x)^2] } / N-1
>
> Now I realise the latter (let's call it Var(n-1)) is the best estimate of
> the variance of the overall population, if my data is a sample from that
> population, but that's not what I want (or expect) from the variance function.
>
Though the documentation to the VARIANCE function should probably include the
formula, I would think that the IDL definition (with N-1 in the denominator)
is the one that, in practice, will be most often used. This is the
formula to use when one has a set of measurements and wants to estimate the mean
and variance from those measurements.

The formula with N in the denominator should be used when one somehow knows
beforehand the true value being measured - perhaps useful for Monte
Carlo experiments or when the mean is known from a different experiment.
Note that more than a keyword must be added to VARIANCE to do this calculation
-- one must also supply the true value of the mean.

In any case, the computation of the variance can be a one-line IDL statement,
if you don't want to use the VARIANCE function.

--Wayne Landsman landsman@mpb.gsfc.nasa.gov