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

Subject: Re: VARIANCE in IDL
Posted by Martin Schultz on Wed, 24 Feb 1999 08:00:00 GMT
View Forum Message <> Reply to Message

Justin Ashmall wrote:
>
> Dear All,
>
> [...]
>
> VARIANCE = { SUM [(x - mean_x)^2] } / N
>
> [vs.]
>
> VARIANCE = { SUM [(x - mean_x)^2] } / N-1
>

as N approaches infinity, the tiny 1 doesn't matter. For samples with low N, it is always questionable what you can learn from the variance. In these cases it is often preferrable to use the absolute difference from the mean as a measure for the scatter of the data.
Example:
a = [-1., 1., 2., 8.]
b = [-1., 1., 4., 6.]

	A	B
MEAN=	2.50000	2.50000
VAR(N)=	11.2500	7.25000
VAR(N-1)=	15.0000	9.66667
ABS.DIFF=	2.75000	2.50000

While the variance differs by more than 50% between the two cases, the absolute difference is only 10% which I would consider a fairer result. Because it is a quadratic measure, the variance is very sensitive to outliers,
and these have a much greater influence on results for small samples.

Martin.

PS: this is a one-liner to compute the absolute difference:
ma = mean(a) & adiff = total(abs(a-ma))/n_elements(a)

--
---------- -------
Dr. Martin Schultz
Department for Engineering&Applied Sciences, Harvard University
109 Pierce Hall, 29 Oxford St., Cambridge, MA-02138, USA

phone: (617)-496-8318
fax : (617)-495-4551

e-mail: mgs@io.harvard.edu

Internet-homepage: http://www-as.harvard.edu/people/staff/mgs/
