Subject: Re: Question about correlate.
Posted by wmc on Sat, 22 Apr 2000 07:00:00 GMT
View Forum Message <> Reply to Message

Eric Kihn <ekihn@ngdc.noaa.gov> wrote:
> This one has me perplexed. I'm using correlate a sample and
> predicted value.
> kpfinite is the sample and prkpfinite is the predicted value.

> IDL> lowkp = where(kpfinite lt 2.0, count)
> IDL> print, correlate(kpfinite(lowkp), prkpfinite(lowkp))
> 0.532239
> IDL> highkp = where(kpfinite ge 2.0, count)
> IDL> print, correlate(kpfinite(highkp), prkpfinite(highkp))
> 0.723756
> IDL> print, correlate(kpfinite, prkpfinite)
> 0.815049

> My question is how is the total correlation gt then the correlation on
> either of the two ranges, when clearly lt and gt 2.0 comprises the
> entire range of Kp? It's
> not clear if this is a stats question or a IDL programming problem on my
> part. Any help appreciated.

Its a stats question. The result you get is exactly what you expect.

Consider:

```
wmc> print,correlate(randomn(seed,1000),randomn(seed,1000))
   0.0400765

wmc>  print,correlate([randomn(seed,1000),100+randomn(seed,1000)],
[randomn(seed,1000),100+randomn(seed,1000)])
    0.999617
```

Which is to say: if your data separate into 2 clumps, one with large and one with small values, then each clump can have zero correlation, but both together can have a very high corr.

-W.

--
William M Connolley | wmc@bas.ac.uk | http://www.nbs.ac.uk/icd/wmc/
Climate Modeller, British Antarctic Survey | Disclaimer: I speak for myself
(yes, BAS has at at last got rid of that irritating "public" in the URL)