Barbara Byrne - NOAA Federal

From:	Barbara Byrne - NOAA Federal
Sent:	Thursday, May 23, 2019 10:17 AM
To:	Stephen Maurano - NOAA Federal
Subject:	Re: Calculating Percentiles in R and Excel
Attachments:	SH salvage measures_steel only.xlsx

Thanks, Stephen! I looked at this youtube video (https://www.youtube.com/watch?v=13Ahbn9o-5I) and got the impression that percentile.exc made the most sense, but when I tried out both methods (see H4:J22 of the "Steel distribution" tab of the attached excel workbook), the percentile.exc function -- for both time periods -- returned the highest value in my dataset, which didn't feel very 90%-ile-y. With only 9 datapoints in a skewed distribution, I think percentiles can just get wonky. I ended up using the percentile.inc function, and just documented it in my proposal so readers can reproduce my results.

On Thu, May 23, 2019 at 9:56 AM Stephen Maurano - NOAA Federal <stephen.maurano@noaa.gov> wrote:

> Hi Barb,
>
> Per the question you mentioned yesterday, there are a number of ways to calculate percentiles in R and Excel (and other statistics software). As described in Wikipedia and ultimately citing to Hyndman, R. J. and Fan, Y. (1996) Sample quantiles in statistical packages, *American Statistician* 50, 361--365. 10.2307/2684934.
>
> - Excel PERCENTILE and PERCENTILE.INC are equivalent to R-7 (the default in R)
> - Excel PERCENTILE.EXC is equivalent to R-6 (the default in other common stat software, Minitab and SPSS)
>
> "The differences between the definitions are most evident for small data sets [n<100] and when there is a large "gap" between one or more adjacent data values" (per this SAS blog).
>
> PERCENTILE.INC works for any value of k between 0 and 1 whereas PERCENTILE.EXC only works if k is between 1/n and 1-1/n, where n is the number of elements in array. Since you're looking for the 90%ile, you only need a dataset of n=9 (although the %ile return will still be sensitive to a large range b/t n and n-1 values)
>
> Put another way, in an array of (1:9) PERCENTILE.EXC can return a 90%ile, but not a 99%ile (whereas PERCENTILE.EXC can do either)
>
> ```
> 9.0   =PERCENTILE.EXC( ,0.9)
> 8.2   =PERCENTILE.INC( ,0.9)
> #NUM! =PERCENTILE.EXC( ,0.99)
> 8.9   =PERCENTILE.INC( ,0.99)
> ```
>
> If you're short on time, I would use the excel default (PERCENTILE). If you have even ten minutes, I would run your data in R to (1) examine the distribution (plot histogram, examine outliers) and (2) iterate a few of the common definitions quickly, and see if the results are different. I suspect the differences will be

trivial. Here's sample code, accomplishing the same task as the Excel example above, although R appears to default back to the n'th value, rather than the #NUM! error, when k>1-1/n. Happy to help further if there's anything I can assist with. The information seems fairly well documented online, so I'm 0.9 confident :-) but I just googled and played around with the numbers, so you can certainly check with other folks to get a more authoritative answer. Thanks for mentioning, it's always educational to learn something new!

```
> nine<-(1:9); ninetynine <-(1:99) #create arrays of 1 to 9 and 99
> quantile(nine,c(.9,.99),type=7) # the default in R and Excel
 90%  99%
8.20 8.92
> quantile(nine,c(.9,.99),type=6) # equivalent to PERCENTILE.EXC
90% 99%
  9   9
```

--

Stephen Maurano
Natural Resource Management Specialist
NOAA Fisheries West Coast Region
U.S. Department of Commerce
Office: (916) 930-3710
Stephen.Maurano@noaa.gov
www.westcoast.fisheries.noaa.gov

--

Barb Byrne
Fish Biologist
NOAA Fisheries West Coast Region
U.S. Department of Commerce
Office: 916-930-5612
barbara.byrne@noaa.gov
California Central Valley Office
650 Capitol Mall, Suite 5-100
Sacramento, CA 95814

Find us online
www.westcoast.fisheries.noaa.gov