Quantile Estimation

When More than the Mean and the Standard Deviation Are Needed

BY THOMAS J. BZIK

Q: Which method of quantile estimation should be used?

A: The estimation of a specific quantile of a data population characteristic is a routine statistical task. Some commonly selected levels for estimation are the first quartile (25 percent), second quartile (50 percent, median), and the third quartile (75 percent). A named quantile, say 25 percent, has 25 percent of the data distribution below the named quantile and (100-25 percent) = 75 percent of the data distribution above the named quantile. Often lower or higher quantile levels such as 1 percent, 5 percent, 10 percent, 90 percent, 95 percent and 99 percent are of interest when the tail regions of a population characteristic are of interest rather than the core of the distribution.

There are two broad approaches to quantile estimation, both of which make use of a set of sample data but in differing manners. One approach is the direct estimation approach. In this approach a given quantile is typically estimated from use of one or two specific elements of the ordered data. Statisticians consider such direct estimates to be nonparametric since they do not rely on parameters from an assumed distribution. A second approach is the distributional approach. In this approach the data is used to estimate parameters from an assumed distributional model, and these parameter estimates allow any selected quantile to be estimated. Each approach provides potential advantages and disadvantages.

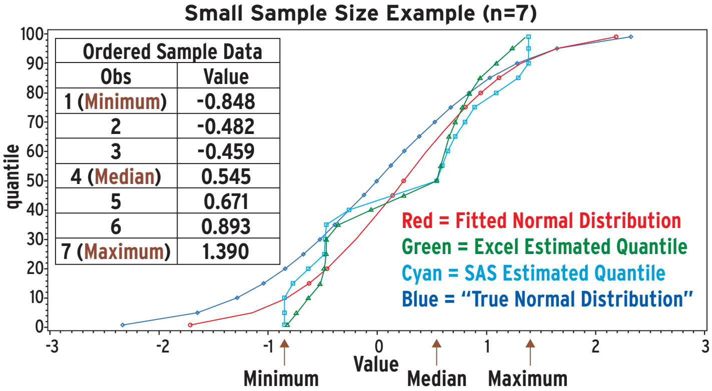
> **Why should quantile estimation be of interest? The short version is that the mean and the standard deviation are often not enough to effectively summarize a distribution.**

Figure 1 was constructed by simulating seven random observations from a standard normal distribution with mean 0 and standard deviation 1. The blue curve shows the true population from which the sample was drawn. Two classes of quantile estimation are illustrated, distributional and direct. The distributional method shows a fitted normal distribution (red curve) using the sample mean and sample standard deviation of the seven measurements. The observed difference between the red fitted distributional model and the blue population model is due to sampling variation. This is part of the price for having very little sample data. Despite the small sample size, a distributional methodology can estimate any desired quantile.

The green and the cyan estimates were generated by two differing direct estimation methodologies (Excel and SAS) that use the ordered sample data to estimate a given quantile. These methodologies use either one sample value or two neighboring ordered sample values to provide any desired estimated quantile. A familiar example is the median, which for an odd sample size is estimated as the middle ordered observation and for even sample sizes is estimated as the average of the two middle ordered observations. While the directly estimated medians are in agreement (quantile = 50 percent), Excel and SAS are generally not in good agreement.

Since the direct estimates are also based on the sample data, they will reflect whatever bias there was in the sample data collection. Hence they track the fitted distributional model more than the true population distribution used in Figure 1. The SAS and Excel estimates do have some things in common. Neither one provides an estimate that is below the observed minimum or is above the observed maximum. This is problematic in that if, for example, one additional sample point beyond seven were to be collected there is a 25 percent = (2/(7 + 1)) chance it will exceed either the prior maximum or fall under the prior minimum. Such direct estimates are not reasonable for estimating a quantile that is beyond the quantile that is likely to be contained in the sample data. In this respect, the Excel method (percentile function) is worse than the illustrated SAS method, but neither is good for small sample estimation of a quantile relatively close to zero or relatively close to one, i.e., the extremes of the distribution. There are many alternative "direct" methodologies, for example, SAS offers the choice of five approaches (PCTLDEF = 4 used herein and is recommended). SAS's default approach, PCTLDEF = 5, is not recommended for small sample sizes. Excel's results do not benchmark to any of these five SAS definitions and appears to be a "unique" definition. Minitab, for example, uses the equivalent of SAS's PCTLDEF = 4 when it reports quartiles in descriptive statistics results. For a small sample size, it is recommended that you do not use the Excel percentile function.

Figure 2 was constructed by simulating 50 random observations from a nonnormal distribution. The blue curve shows the true population from which the sample was drawn. The results of fitting a normal distribution using the sample mean and sample standard deviation is in red. The relatively large observed differences between the red fitted distributional



Ordered Sample Data	
Obs	Value
1 (Minimum)	-0.848
2	-0.482
3	-0.459
4 (Median)	0.545
5	0.671
6	0.893
7 (Maximum)	1.390

Figure 1 — Small Sample Size, Multiple Quantile Estimation Methodologies

model and the blue population model is due to assuming normality when it is not appropriate. The direct estimation methodologies do a much better job than a poorly assumed normal distribution in this example. If any distribution is to be fit, at a minimum, the data should not statistically contradict use of such an assumed model. Additionally, the differences between the SAS PCTLDEF = 4 methodology (cyan) and the Excel methodology (green) have become relatively small other than in the tail regions.

As the sample size becomes large, the use of any fitted distributional model becomes relatively more questionable as a means of quantile estimation. The differences between varying definitions of how to estimate a quantile from an ordered set of data becomes less and less relevant with increasing sample size. Large sample size almost always implies that a direct method of estimation is preferred. The only large sample size caution is that estimating an extreme tail quantile can still be problematic unless the sample data collection is large enough.

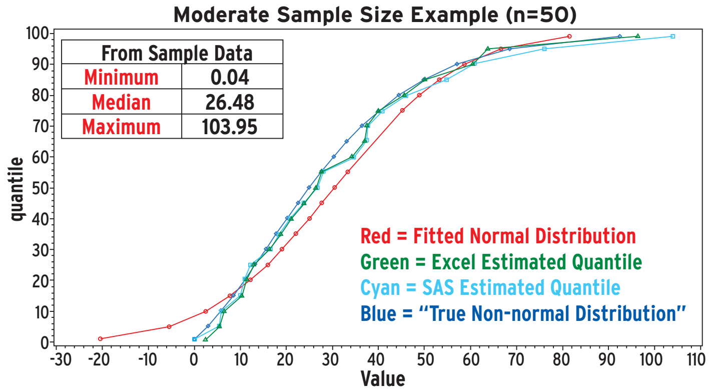


Figure 2 — Moderate Sample Size, Multiple Quantile Estimation Methodologies

Why should quantile estimation be of interest? The short version is that the mean and the standard deviation are often not enough to effectively summarize a distribution. Statistically significant disagreement that is practically meaningful between a quantile estimated directly and one estimated from an assumed distribution implies that use of the distributional model is ill advised. As sample size becomes large, direct estimation will almost always provide better results.

THOMAS J. BZIK, *Air Products and Chemicals Inc., Allentown, Pennsylvania, is the American Statistical Association representative to ASTM Committee E11 on Quality and Statistics; he serves as E11 secretary and as vice chairman of E11.11 on Sampling/Statistics.*

DEAN V. NEUBAUER, *Corning Inc., Corning, New York, is an ASTM International fellow, chairman of E11.90.03 on Publications and coordinator of the DataPoints column; he is immediate past chairman of Committee E11 on Quality and Statistics.*

snonline

Get more tips for ASTM standards development at www.astm.org/sn-tips.

Find other DataPoints articles at www.astm.org/standardization-news/datapoints.