TO MEAN OR MEDIAN; THAT IS THE QUESTION
MEASURES OF CENTRAL TENDENCY
DOCKET NO. 02-057-02
DPU EXHIBIT 6.10

Generally speaking, the sample mean will have better sampling properties than will the sample median. That is, in measuring the central tendency of a distribution, the sample mean will have a smaller sampling error than will the median. Therefore, the sample mean is generally the preferred statistic to summarize or represent the typical value to be expected if a single data point were drawn from a population. In specific cases, however, if the distribution of the population departs significantly from the normal (i.e., symmetric and bell-shaped) distribution, then the median may be the better statistic to summarize the sample's (and the population's) central or typical value.

Therefore, the choice between the two statistics depends on the characteristics of the sample data.

Large Samples (Normal Distributions)

For large samples the Central Limit Theorem implies that the sample mean ($\bar{X}$) is approximately normally distributed[1] with an expected value equal to the mean of the population (μ) and variance

$$Var(\bar{X}) = \frac{\sigma^2}{n} \tag{1}$$

where σ^2 is the population variance and n is the sample size. This is true whether or not the population itself is normally distributed.[2] The sample median ($\tilde{X}$) is also approximately normally distributed for large samples. If we let the sample size n = 2m+1 (i.e., n is an odd number), then the sample median will be approximately normally distributed with an expected value equal to the population median ($\tilde{\mu}$) and variance

$$Var(\tilde{X}) = \frac{1}{8*[f(\tilde{\mu})]^2 * m} \tag{2}$$

where f(·) represents the probability density function of the population from which the sample is drawn. If the sample is drawn from a normal population, then the population mean and median are equal ($\mu = \tilde{\mu}$). In this case the sample median will be approximately normally distributed with an expected value equal to the population mean and variance approximately equal to

---

[1] The normal distribution is bell-shaped and symmetric around the mean (μ):

[2] See John E. Freund, *Mathematical Statistics*, 5th Ed. Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1992, pp. 294-298.

$$Var(\tilde{X}) = \frac{\pi\sigma^2}{2(n-1)} \cong 1.57 * \frac{\sigma^2}{n-1} = 1.57 * \left(\frac{n}{n-1}\right)\left(\frac{\sigma^2}{n}\right) = 1.57 * \left(\frac{n}{n-1}\right) * Var(\bar{X}) \quad (3)$$

Thus, in large samples drawn from normally distributed populations, the sample median has a larger variance than does the sample mean.[3] And the sampling error of the median will be more than 25% larger than that of the sample mean:

$$\sqrt{Var(\tilde{X})} \cong 1.25 * \sqrt{\left(\frac{n}{n-1}\right)}\left(\frac{\sigma}{\sqrt{n}}\right) = 1.25 * \sqrt{\left(\frac{n}{n-1}\right)}\sqrt{Var(\bar{X})} \quad (4)$$

Therefore, in large samples drawn from normal populations, the mean is a better measure of central tendency than the median.

Small Samples (Approximately Normal Distributions)

In small samples, as long as the sample does not contain extreme values, it can be shown using numerical techniques that the sample mean is generally still a better measure of central tendency. One such technique is bootstrapping.

Bootstrapping is commonly used to estimate probability ranges for sample statistics. For example, bootstrapping can be used to estimate the range or interval "a" to "b" such that, the probability that the sample mean is less than "a" or greater than "b" is equal to 5 percent – the probability that the sample mean is between "a" and "b" is ninety percent. If we estimate the ninety percent range for the sample median and the sample mean, then the relative size of these two ranges is an indication of which sample statistic, the mean or the median, is more accurate, or in other words, has a smaller sampling error. In the case where there are no outliers in the sample, we would expect the range for the sample mean to be smaller than that for the sample median.

To demonstrate, consider the sample of ROE estimates produced in Dr. Williamson's testimony (Exhibit QGC 5.3): {14.83, 14.84, 15.85, 13.80, 12.62, 11.54, 12.3, 14.64, 13.49}. In this sample of nine observations, there are no extreme values as measured under standard box-plot techniques. And, as expected, the estimated ninety percent range for the sample median is greater than that for the sample mean.

[3] Ibid., pp. 322-325.

Table 1: Bootstrapping Results for Questar's List of Companies (Original Sample)

	Sample Mean	Sample Median	Sample Size	
Original Sample Statistics	13.77	13.80	9	
Estimated (Bootstrapped) 90% Confidence Interval (Sample Size 9; 1,000 Replications)				
	Lower Bound	**Upper Bound**	**Range**	**Midpoint**
Sample Mean	13.09	14.39	1.31	13.74
Sample Median	12.62	14.83	2.21	13.73
Relative Size (Med/Mean)			1.69	

Indeed, the range for the sample median is almost 70% greater than that for the sample mean. Thus we conclude that the sample mean is a better estimate of the typical value when the sample does not contain outliers.[4]

Non-Normal Distributions

If there are outliers in the sample, the relative accuracy of the sample mean and median are likely to change; the median will be a better estimate of the central tendency of the distribution if there are outliers in the data. This can also be demonstrated using bootstrapping techniques.

Consider again the sample listed above. If we add five to the two largest values in the sample, then the two resulting values will be outliers: {11.54, 12.30, 12.62, 13.49,13.80, 14.64, 14.83, 19.84, 20.85}. In this case, the estimated 90% range for the sample mean is approximately 36% larger than the estimated range for the sample median. (See Table 2) Thus, in the face of outliers, the sample median is the better sample statistic to represent the typical or expected value.

[4] For an explanation of the bootstrapping and box-plot techniques, see DPU Exhibits 6.11 And 6.12 Respectively.

Table 2: Bootstrapping Results, Modified Sample

	Sample Mean	Sample Median	Sample Size	
Modified Sample Statistics	14.88	13.80	9	
Estimated (Bootstrapped) 90% Confidence Interval (Sample Size 9; 1,000 Replications)				
	Lower Bound	**Upper Bound**	**Range**	**Midpoint**
Sample Mean	13.27	16.27	3.00	14.77
Sample Median	12.62	14.83	2.21	13.73
Relative Size (Med/Mean)			0.74	