ECONOMICS BULLETIN

Median as a weighted arithmetic mean of all sample observations

SK Mishra
Dept. of Economics, NEHU, Shillong, India

Abstract

This paper shows how median may be computed as a weighted arithmetic mean of all sample observations, unlike the conventional method that obtains median as the middle value (odd observations) or a simple mean of the two middlemost values (even observations). Monte Carlo experiments have been carried out to investigate the bias, efficiency and consistency of the alternative methods.

Citation: Mishra, SK, (2004) "Median as a weighted arithmetic mean of all sample observations." *Economics Bulletin,* Vol. 3, No. 18 pp. 1−6
Submitted: June 3, 2004. **Accepted:** June 12, 2004.
URL: http://www.economicsbulletin.com/2004/volume3/EB−04C10011A.pdf

Median as a Weighted Arithmetic Mean of All Sample Observations

SK Mishra
Dept. of Economics
NEHU, Shillong (India)

1. Introduction: Innumerably many textbooks in Statistics explicitly mention that one of the weaknesses (or properties) of ***median*** (a well known measure of central tendency) is that it is not computed by incorporating all sample observations. That is so because if the sample $x = (x_1,\ x_2, ..., x_n)$, where the variate values are ordered such that $x_1 \le x_2 \le ... \le x_n$ then $median(x) = (x_k + x_{n+1-k})/2;\ k = \text{int}((n+1)/2).$ Here int(.) is the integer value of (.). For example int(10 ≤ (n+1)/2 < 11) = 10. This formula, although queer and expressed in a little roundabout way, applies uniformly when n is odd or even. Evidently, $median(x)$ is not obtained by incorporating all the values of x, and so the alleged weakness of the median as a measure of central tendency.

2. The Median Minimizes the Absolute Norm of Deviations: It is a commonplace knowledge in Statistics that the statistic $\bar{x}$ (the arithmetic mean of x) minimizes the (squared) Euclidean norm of deviations of the variate values from itself or explicitly stated, it minimizes $S = \sum_{i=1}^{n} \left| x_i - c \right|^2$ since S attains its minimum when $c = \bar{x}$. To obtain this result, one may minimize $\sqrt{S}$ (the Euclidean norm per se) also. On the other hand the *median* minimizes the Absolute norm of deviations of the variate from itself, expressed as $M = \sum_{i=1}^{n} |x_i - c|$ which yields $c = median(x)$. In a general framework, we obtain arithmetic mean or median by minimizing the general Minkowski norm $\left[\sum_{i=1}^{n} |x_i - c|^p \right]^{1/p}$ for p=2 or p=1 respectively. This view of the arithmetic mean and the median gives them the meaning of being the measures of central tendency.

3. Indeterminacy of Median when the Number of Values in the Sample is Even: When in the sample $x = (x_1,\ x_2, ..., x_n)$, the number of observations, n, is odd, the value of $median(x) = (x_k + x_{n+1-k})/2;\ k = \text{int}((n+1)/2)$ is determinate; $x_k = x_{n+1-k}$ minimizes the absolute norm, M. However, when n is an even number, x_k and x_{n+1-k} are (very often) different. As a matter of fact, any number z for which the relationship ($x_k \le z \le x_{n+1-k}$) holds, minimizes the absolute norm of deviations. Thus, the median is indeterminate. It has been customary, therefore, that in absence of any other relevant information, one uses the ***principle of insufficient reason*** and obtains $median(x) = (x_k + x_{n+1-k})/2.$ However, it remains a truth that any number z for which the relationship ($x_k \le z \le x_{n+1-k}$) holds, is the value of the median as much as $z = (x_k + x_{n+1-k})/2.$

4. Median as a Weighted Arithmetic Mean of Sample Observations: If $x=(x_1,\ x_2,...,x_n)$ are ordered such that $x_1 \le x_2 \le ... \le x_n$, it is possible to express median as a weighted arithmetic mean $\left(\sum_{j=1}^{n} x_j w_j \Big/ \sum_{j=1}^{n} w_j\right)$ where $w_j = w_{n+1-j} = 0.5$ for $j = \text{int}((n+1)/2)$ else $w_j = 0$ for $j \ne \text{int}((n+1)/2)$. However, this is trivial.

Now we present a non-trivial alternative algorithm to obtain $median(x)$. In order to use this algorithm it is not necessary that the values of x be arranged in an ascending (or descending) order, that is $x_1 \le x_2 \le ... \le x_n$ condition is relaxed. The steps in the algorithm are as follows:

(i) Set $w_i = 1\ \forall\ i = 1,2,...,n.$ Obviously, $\sum_{i=1}^{n} w_i = n.$

(ii) Find $v_1 = \dfrac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$, the weighted arithmetic mean of $(x_1,\ x_2,...,x_n)$.

(iii) Find new $w_i = 1\big/|d_i|$ if $d_i = |x_i - v_1| \ge \varepsilon$ ($\varepsilon > 0$ is a small number, say 0.000001), else $w_i = 0.000001$ or any such small number; $i = 1,2,...,n.$

(iv) Find $v_2 = \dfrac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$ using the weights obtained in (iii) above.

(v) If $|v_1 - v_2| \ge \tau$ (where τ is a very small number, say, 0.00001 or so, controlling the accuracy of result) then v_1 is replaced by v_2 (that is, v_2 is renamed as v_1) and go to step (iii); **else**

(vi) Median is v_2 and $w=(w_1, w_2,...,w_n)$ are the weights associated with $(x_1,\ x_2,...,x_n)$. Stop.

This algorithm yields non-trivial weights $w=(w_1, w_2,...,w_n)$. It yields the median identical to that obtained by the conventional formula if n is odd. If n is even, it gives a number z : ($x_k \le z \le x_{n+1-k}$), which is median as mentioned in section 2.

5. Some Monte Carlo Experiments: We have conducted some Monte Carlo experiments to study the performance of the alternative method (weighted arithmetic mean representation) vis-à-vis the conventional method of obtaining median. Three sample sizes (of n = 10, 21 and 50) have been considered. Samples have been drawn from five distributions (Normal, Beta_1, Beta_2, Gamma and Uniform). In each case 10,000 experiments have been carried out. A success of the alternative estimator is there if it obtains median identical to that obtained by the conventional method in case n is odd and obtains median = z : ($x_k \le z \le x_{n+1-k}$) in case n is even. The summary of results is presented in table 1.

We find that when n is odd, irrespective of the distribution or the sample size both the methods yield identical results. When the distribution is skewed (i.e. there is a significant divergence between median and mean) and n is even, the alternative median is slightly pulled by the mean (its inclination is towards the mean). This appears justified because it is expected that the values lying between x_k and x_{n+1-k} (for $x = (x_1,\ x_2, ..., x_n)$: $x_1 \le x_2 \le ... \le x_n$; $k = \text{int}((n+1)/2)$) must be more densely distributed in the side of the mean. The conventional method, however, considers them uniformly distributed in want of information. The alternative method appears to exploit the information contained in the sample.

6. Analysis of Inclination of Computed Medians to Mean Value: We have seen that when n is an even number, the values of median estimated by the two methods differ and the one estimated by the alternative (weighted arithmetic mean) method appears to be pulled towards the mean value, $\bar{x}$. Then, a question arises : is the median estimated by the alternative method biased (towards the mean)? To investigate into this question, we generate some n_a values (in our experiment 80) of v such that ($x_k \le v \le x_{n+1-k}$), and v follows the distribution identical to that of x. We do it again and again for a large number of times (in our experiment, 10,000). We count as to how many times the v_i< the median values obtained by the two competing methods. The probability of v_i =computed medians is very small (in our experiment we never encountered equality). Table-2 clearly shows that in case of Gamma and Beta_2 distributions both medians are pulled by mean, though the median obtained by the alternative method is more inclined to mean. The pull is stronger in case of the Gamma distribution, since it is more skewed than the Beta_2 distribution. In case of normal distribution we find the opposite tendency (push). In case of uniform distribution no pull or push force is observed, while in case of Beta_1 distribution a mixed observation is there.

7. Relative Efficiency and Consistency of the Competing Methods: Now suppose we generate a large (in our experiment 5001) number of variate values following a specified distribution. Let us call the collection of these values U or the Universe. We may obtain the Median(U) = μ , say. This value may not be the true median of the distribution (or if U were very large) , but it is likely to be very close to that.

From U we may draw some n (in our case 10, 50 and 90) random values, say $x = (x_1,\ x_2, ..., x_n)$, compute medians ($m_0$ and m_1) by the two competing methods (respectively) again and again. In our case, *ntrial*=1000, with replacement. In each draw, the computed medians will differ from μ . From this, we may obtain the norms for each median. These norms would suggest which median is most frequently closer to μ . Symbolically,

$$Norm_t = \left(\sum_{i=1}^{ntrial} \left| m_{i,t} - \mu \right|^p \right)^{1/p} \ ; \ \begin{pmatrix} t = 0 \ \ for \ \ traditional \\ t = 1 \ \ for \ \ alternative \end{pmatrix}$$

We have used the absolute norm (p =1 in the formula defining norm). The results of the experiments are given in table 3. We observe that for Uniform, Normal and Beta_1 distributions $norm_1$ is smaller than $norm_0$. For Gamma and Beta_2 distributions the opposite is true. We also observe that the norms are smaller for larger values of n, indicating to consistency.

8. Asymmetry of Distribution and Efficiency of the Competing Methods: It is well known that the Gamma distribution is severely skewed for small shape parameters, but with the increasing value of that parameter, the distribution tends to become symmetric.

Table 4 shows the relative norms for the competing methods due to increasing values of the shape parameter of the Gamma variate. We observe that norm_1 becomes uniformly smaller (than $norm_0$) while the shape parameter reaches 16. This experiment reinforces our conclusion that the alternative method of obtaining median is better than the conventional method while the distribution is less asymmetric.

9. Conclusion: This study establishes that median may be expressed as a weighted arithmetic mean of all sample observations. If the conventional formula does not incorporate all sample values, it is the property of the specific method of computation and not of median per se, as often alleged to it. If our experiments convey something, then we may also state that for relatively more symmetric distributions the alternative formula (weighted mean) performs better than the conventional method. But for heavily asymmetric distributions the conventional method of computing median performs better, although both the methods yield biased estimates.

The alternative algorithm of computation is easily extended to other median type estimators - such as Least Absolute Deviation (LAD) estimator of the regression model $y = X\beta + u$ - as shown by Fair (1974) and Schlossmacher (1973).

References

Fair, RC (1974). "On the Robust Estimation of Econometric Models", *Annals of Economic and Social Measurement*, 3 (667-677).

Schlossmacher, EJ (1973). "An Iterative Technique for Absolute Deviations Curve Fitting", *Journal of the American Statistical Association*, 68 (857-859).

Table 1. Performance of the Alternative Method to obtain Median

Distribn.	Sample Size = n	Arithmetic Mean	Median (Conventional)	Median (Alternative)	Inclination to Mean	Success Rate (%)
Uniform	10	50.00490	50.04172	50.00207		100
	21	49.98979	50.02352	50.02352		100
	50	49.99520	50.05171	50.07819		100
Gamma	10	2.50630	1.35170	1.57886	Yes	100
	21	2.50647	1.23411	1.23415		100
	50	2.50656	1.17245	1.22250	Yes	100
$Beta_1$	10	251.66043	251.45144	251.47560		100
	21	251.63371	252.62551	252.62551		100
	50	251.64002	253.49261	252.82187		100
$Beta_2$	10	3343.83487	614.92200	741.89412	Yes	100
	21	3346.41080	526.19264	526.19264		100
	50	3346.43339	500.91137	519.64713	Yes	100
Normal	10	0.00062	-0.00149	0.02596		100
	21	0.04474	0.39990	0.39990		100
	50	0.07170	-0.11195	-0.14733		100

Table 2. Inclination of the Competing Methods to the Mean Value						
Distribn.	Sample Size = n	n_a (no. of v values generated)	Median (Conventional)	Inclination to Mean	Median (Alternative)	Inclination to Mean
Uniform	10	80	0.49932	No	0.49975	No
	50	80	0.49913	No	0.50818	No
Gamma	10	80	0.93670	+ Yes	0.97439	+ + Yes
	50	80	0.93670	+ Yes	0.98637	+ + Yes
$Beta_1$	10	80	0.50140	No	0.50178	No
	50	80	0.50137	No	0.46743	- Yes
$Beta_2$	10	80	0.98186	+ Yes	0.98721	+ Yes
	50	80	0.98188	+ Yes	0.98743	+ Yes
Normal	10	80	0.88420	- - Yes	0.80256	- Yes
	50	80	0.88433	- - Yes	0.78008	- Yes
+ pull; + + stronger pull; - push; - - stronger push						

Table 3. Efficiency of the Competing Methods to obtain Median						
Distribn.	Sample Size = n	True Median U(5001)	Computed Median (m_0)	$\frac{1}{1000} Norm_0$ (ref. m_0)	Computed Median (m_1)	$\frac{1}{1000} Norm_1$ (ref m_1)
Uniform	10	49.54680	50.24744	1128.71303	50.15183	948.46047
	50	49.54680	49.77801	109.58568	49.80445	99.67029
	90	49.54680	49.59458	47.33987	49.62526	43.96779
Gamma	10	1.18282	1.37910	64.02190	1.60125	75.13577
	50	1.18282	1.21661	6.04453	1.26538	6.29757
	90	1.18282	1.20275	2.62366	1.23218	2.71264
$Beta_1$	10	251.40387	246.36445	8190.69425	246.39085	6644.16110
	50	251.40387	253.87199	885.00476	253.43050	786.28595
	90	251.40387	248.03851	373.58662	248.16928	344.18679
$Beta_2$	10	505.64720	819.56830	53555.96607	1058.20626	72762.71516
	50	505.64720	574.58766	4062.13108	608.48262	4391.75660
	90	505.64720	523.30939	1578.37985	539.86140	1654.75995
Normal	10	-1.17553	-0.68537	2909.56588	-0.82465	2667.77153
	50	-1.17553	-0.90968	280.48175	-0.84421	268.16238
	90	-1.17553	-0.81813	122.07105	-0.81020	117.49905

Table 4. Asymmetry of Distribution and Efficiency of the Competing Methods

Distribution Gamma (shape parameter)	Sample Size = n	True Median U(5001)	Computed Median (m_0)	$\frac{1}{1000} Norm_0$ (ref. m_0)	Computed Median (m_1)	$\frac{1}{1000} Norm_1$ (ref m_1)
Gamma(0.5)	10	1.18282	1.37910	64.02190	1.60125	75.13577
	50	1.18282	1.21661	6.04453	1.26538	6.29757
	90	1.18282	1.20275	2.62366	1.23218	2.71264
Gamma(1.0)	10	3.47279	3.63562	115.39901	3.97688	125.82501
	50	3.47279	3.54220	11.72396	3.63471	12.23126
	90	3.47279	3.43676	4.85865	3.48243	5.02049
Gamma(2.0)	10	8.33893	8.63084	186.36717	9.05013	191.56218
	50	8.33893	8.40070	18.70702	8.53297	8.98361
	90	8.33893	8.39305	8.18323	8.48076	8.39214
Gamma(4.0)	10	18.24966	18.70355	287.69966	19.14927	284.44678
	50	18.24966	18.41913	26.19773	18.59760	26.36324
	90	18.24966	18.33831	11.55354	18.45101	11.67358
Gamma(8.0)	10	37.60569	38.42688	431.66793	38.93380	411.37192
	50	37.60569	37.54826	37.21978	37.81617	36.79947
	90	37.60569	37.44675	15.26999	37.61967	15.29079
Gamma(16.0)	10	81.39965	80.72528	502.29528	80.59877	451.81319
	50	81.39965	81.30942	40.20729	81.15754	39.05575
	90	81.39965	81.13149	17.31381	81.00336	17.21622
Gamma(50.0)	10	249.13221	248.12446	862.73501	248.99141	836.98511
	50	249.13221	248.70702	82.86870	248.98742	80.24412
	90	249.13221	248.89959	38.98871	249.11508	38.14233