

# Are We Being Mean to the Median?

A closer look at the meaning of averages

By Paul R. Bynum

Market Analyst, Executive Broker, Mount Data Real Estate

*This report discusses the differences between using the mean and median averages in summarizing data for Real Estate. It evaluates the rationale for both approaches. The idea of percentile grouping is presented as an alternative to a simplistic reporting of just the mean or median. Square-footage is uncovered as a hidden source of value adjustments that is rarely reported. Finally, the annual reports of 2006 and 2007 for residential Northwest Arkansas are summarized and compared using percentile group means and square-footage.*

## Contents

(Words in blue hyperlink to the Mount Data glossary)

[THE MEDIAN UNDER ATTACK](#)

[WHAT DOES IT ALL MEAN?](#)

[The idea of distributions](#)

[Mean, Median, Mode](#)

[INTRODUCING PERCENTILE GROUPS](#)

[REALITY AND THE BUYING POOL](#)

[Academic and real world concerns](#)

[Buyer demand and income level](#)

[The Domino effect](#)

[A HIDDEN SOURCE OF VALUE: SQFT](#)

[PUTTING IT ALL TOGETHER](#)

[Price and Sq-Ft adjustments](#)

[Total gain and losses](#)

[IS THE MEDIAN WORTH IT?](#)

[CHARTS AND GRAPHS](#)

[Frequency Graph](#)

[Sale Count Line](#)

[Sales by Percentile Chart](#)

[Percentile comparisons 2006 2007](#)

[Gains and Losses by Percentiles](#)

[Gains and Losses Chart](#)

## The Median Under Attack

Back in my College days, we had fun discovering "paradoxes" in mathematics. We could discover, (or invent), [formulas](#) that give seemingly real results and yet the answers were *absurd* or *contradictory*.

In geometry it was possible for us to show that all triangles were greater than 180 degrees. In algebra we could demonstrate that any number was equal to any other number. The fun was in discovering the "mistake" in the other guy's presentation which led to silly results. *Invariably one assumed something one should not have and this led to the error.*

In the past several weeks, articles have been written blasting the use of the [Median](#) in [statistical](#) reporting. Demonstrations were given trying to convey mistrust for its use. Sort of a "More Lies with Statistics" approach. Once again those statistical "men in black" were out to do mischief and bamboozle the real estate public. Later in this article I will show what these articles found to be "mistrustful" about the Median.

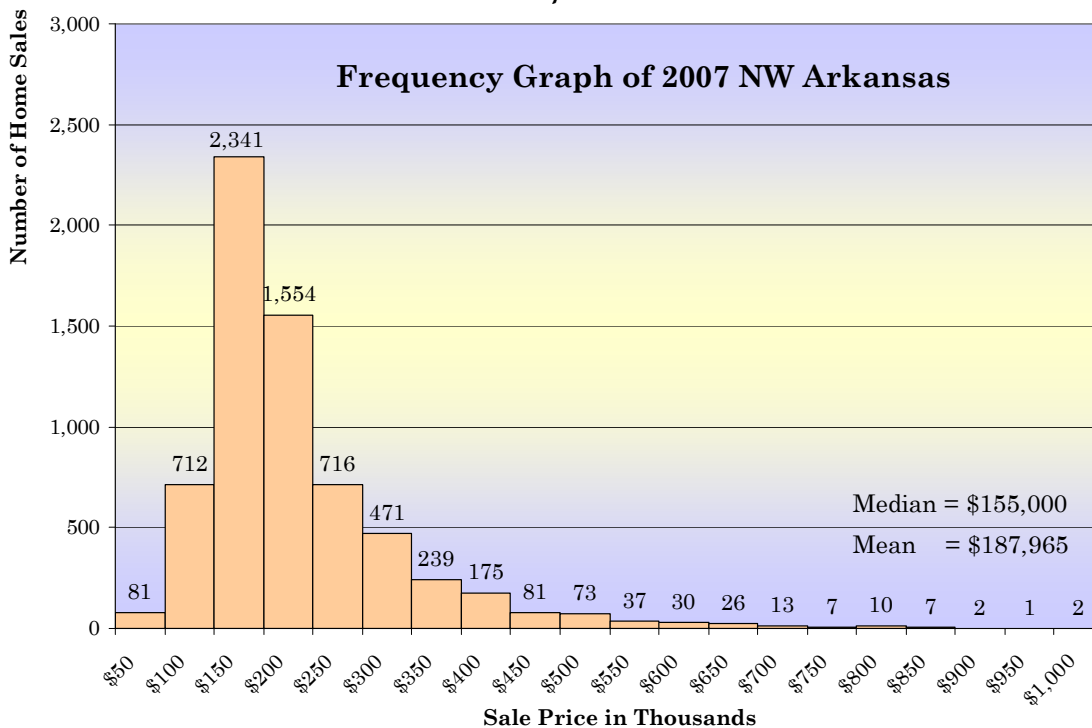
As far as my reports go, I like notoriety as well as the next guy, but I refuse to stoop to fear tactics and the "only I speak the truth" approach to [data](#) reporting.

*I would much rather enlighten you...*

### What Does it all *Mean*?

Market Analysts (I reserve the word [statisticians](#) for people actually *paid* to do this stuff), are often called upon to deal with more than just one particular [number](#). Often, there are many numbers needed to be analyzed in the form of a [series](#). Examples are: the monthly [volume](#) of sales of a product, the hourly temperature readings throughout the day, and the number of annual home sales in a real estate market. Our hope to discover within a series, meaningful [patterns](#) in events that single numbers will not reveal. Series that involve the passage of time are called "[Time Series](#)" and are very common in real estate.

#### *The Idea of Distributions*



The above graph is the Frequency [Distribution](#) of Residential Northwest Arkansas

for the year 2007. It categorizes the number of sales by price ranges. For example: there were 2,341 sales from \$100,000 to \$150,000. Notice how the columns that represent the price ranges rise quickly and then drop off slowly.

To analyze a distribution we find a count of the total number of data points. We are also interested in the "Average," the "[Spread](#)," the "[Standard Deviation](#)" and the "[Skew](#)". In the above *Frequency Distribution* the *spread* ranges from a low sale of \$9,000 (!) to a high of \$1 *million*. (There were actually 11 sales reported over 1 million dollars, but to simply the chart, I left them off).

The *skew* is a measure of lop-sidedness. In the above distribution the frequency peak comes far to the left of the mid-point of the value ranges (\$500,000).

A critical value of any distribution is the average. Three common "averages" are the [Mean](#), [Median](#), and [Mode](#). Their meaning all spring from the idea of the "center". More specifically, a tendency for a series of numbers to cluster about a "[central location](#)."

### *Mean, Median and Mode*

#### Sale Count Line for 2007

Type of Average Sale Price		<i>Mode</i>	<i>Median</i>		<i>Mean</i>	$\frac{1}{2}$ Vol		<i>Totals</i>
		\$145,000	\$150,000		\$187,965	\$199,900		
Accumulated Units	0	←————→	3,307		4,369	4,648	½ of the	6,614
Percent of Units	0%		50%		66%	70%	total volume	100%
Volume (Millions)	\$0		\$387		\$568	\$675	lies in the	\$1,243
Percent of Volume	0%		31%		46%	54%	upper 30%	100%
							of unit sales	

The above table shows 4 different types of "[averages](#)" for the 2007 real estate market in Northwest Arkansas. The red line represents the count of closed units beginning with [zero](#) and ending with **6,614 sales**. It also indicates the total volume (\$1,243 *Million*), and different percentages along the line. Let's start with the "Mean."

The mean is what most people *commonly call* the "average." **Add up the entire series of numbers and divide by the count of the individual members and you have it.** In the above table, we divide the total volume by the total sales and get **\$187,965** as the mean average.

Its strength lies in the ability of almost everyone to grasp it. The mean uses *all* the numbers in the series to calculate the result. Also, if we know the mean and the individual count, a simple multiplication will restore the sum of the series. In fact, **from knowing any 2 of the values (Sum, Count and Mean), we can calculate the remaining value.**

Well, if it's so commonly understood, why do we need the others? Because the mean has *several flaws* that make *other choices* appealing.

- The mean is a mathematical construct and ***does not necessarily represent a real world number.***
  - Example: The census bureau tells us that the average (mean) family consists of **2.7** people. Another example: there was ***no actual sale*** for the amount of the "average" of ***\$187,965*** found above for 2007.
- It can exhibit "***skew***" which ***can lead to false impressions of a distribution that are not in keeping with the true picture.***
  - Example: Notice in the table above, counting from the beginning of the ordered number of sales, the mean lies at number 4,369-considerably to the right of the middle of the count (3,308).

This brings us to the discussion of the ***Median*** average. The Median is that point in a ***distribution*** or series of numbers, at which there are just as many individual values ***above*** the point as there are ***below***. It is the half-way point of an ***ordered*** series. Ordered in this case meaning ***arranged from lowest to highest in value.***

It ***does*** represent a real world number (Well, almost. If there are an odd number of data points, it does. If an even number, we add the middle two numbers and divide by two). The median is ***not*** affected by ***skew***. Making the highest individual number twice as large, will ***not*** affect the position or value of the Median.

In the above table the Median for 2007 was ***\$155,000*** which lies at the mid-point (3,308), of the count. Notice that the median lies at the ***50%*** point of the count, while the mean (in 2007), lies at the ***66%*** point.

So what are the limits of the mean?

- It best represents a distribution with a large quantity of data points. At least a hundred data points is ideal.
- The Median is found by counting, not calculating. Therefore, the Median cannot reconstruct the sum of the data points in a distribution as can the Mean.

Still, the ***Median*** has the advantage of eliminating skew, since it is based on a ***count alone*** and ***not***, as in the case of the ***Mean***, the ***ratio of the volume to the count.*** Finding the median does ***not*** require a series of numbers to be ***summed***, while finding the mean ***does***.

For 2007 the Median tells us there were just as many sales below ***\$155,000*** as there were above. It is automatically, by definition, at the ***50%*** location on the sale line. In my opinion, it is a more accurate picture of the ***availability*** of homes in a region. The ***Mean*** answers nothing about its position on the sale line, and thus nothing about the number of homes available above or below it. In 2007 it ***happens*** to be at ***66%***, but that is ***not a permanent part of its nature.***

The third "**average**" we will consider, is the **Mode**. The Mode is simply the individual sale price with the **highest frequency** or count. Some series have **no** Mode because there is no single data point value having the most occurrences. Some series may actually have more than one Mode, if two or more data points share the same number of occurrences.

In 2007, the Mode is the **\$145,000** sale. There were **67** of them! The mode by itself **doesn't tell us much**. There were 27 separate sale values that had counts over 30. The lowest value was **\$100,000** with 42 sales. The highest was **\$250,000** with 30.

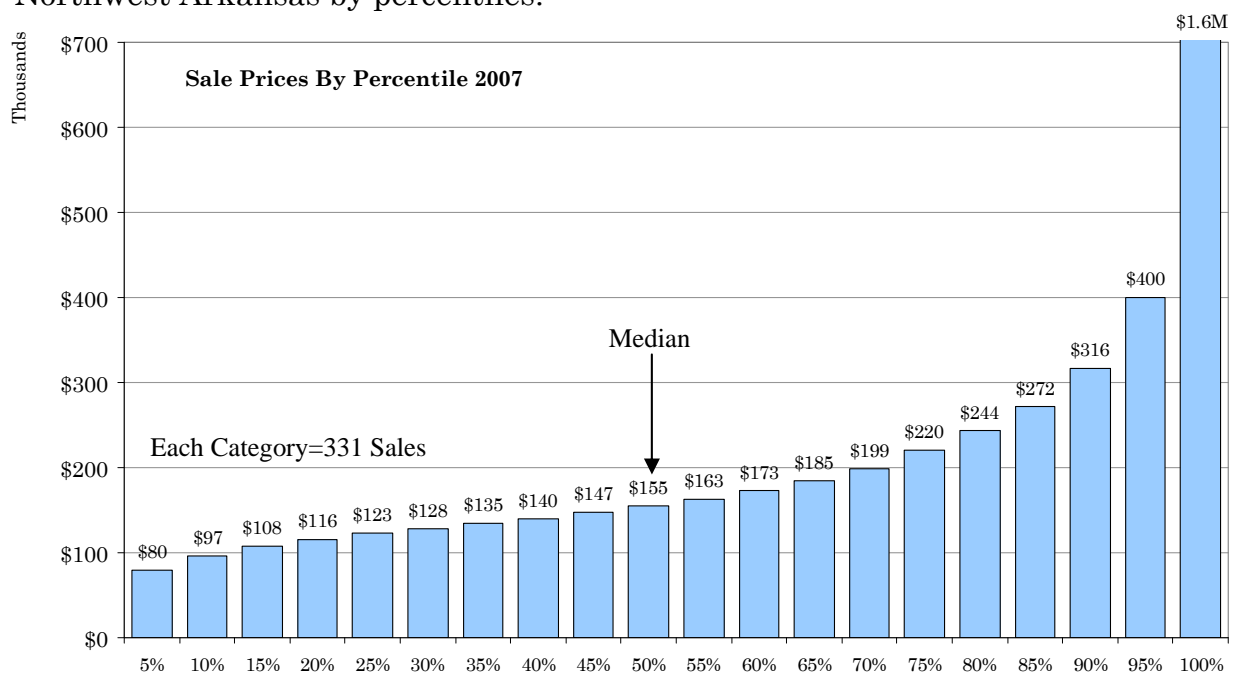
The last "average" we will discuss is really no average at all, but some people may think of it as one. This is the point on the sale line representing  $\frac{1}{2}$  of the volume. **For 2007 in the ordered series it turns out to be at number 4,648, which is 70% of the units and 50% of the volume.** Or, put another way, 30% of the upper-end sales make up 50% of the total volume for 2007. Notice that **both** the **Mean** and the **Median** lie **below** this point.

### Introducing Percentile Groups

You are probably familiar with price categories. The frequency graph shown above has price categories of **\$50,000** each. **Categories help to summarize the sales and give insights into different "sub-markets" of value.**

Another common way of categorizing is using **percentiles** of sales. Instead of dividing up the sales by sold prices, we divide up the sales by count. Using 2007 as an example, the **6,614** sales can be divided in **20** categories of about **331** buyers each. **This makes each category 5% of the buyers.**

**Each category shows the upper-end price associated with that particular category.** Take a look at the chart below which represents the sales in 2007 in Northwest Arkansas by percentiles.



Above is a **percentile chart** for the sales of 2007. Notice above each column is the upper-end sale values. For example: Out of **6,614** buyers, the lowest 5% (or 331) paid up to **\$80,000** for their home. The **next** 5% paid between **\$80,000** and **\$97,000**, and so on. We see the **50%** percentile where buyers paid between **\$147,000** and **\$155,000**. In fact, at **exactly** the **upper 50% point**, they paid **exactly \$155,000**. This is the **Median**.

It is interesting that **70%** of all buyers paid **less than \$199,000** for their home. Notice also, that only **331** buyers, those in the **last 5%** of sales, **paid above \$400,000** for a home in 2007, with **\$1,600,000** being the highest price paid.

## Reality and the Buying Pool

### *Academic and Real World Concerns*

Take a look at the following 5 sales for 2006 in a small community: **\$105,000; \$123,500; \$144,250, \$234,000, and \$275,800**. Here, the **Median** is the middle value or **\$144,250**.

Now suppose in 2007 the sales in the same community were: **\$105,500, \$124,000, \$144,250, \$214,000 and \$260,000**. The Median is the same: **\$144,250**. The question is: "Does this indicate the average home is unchanged over 2006?" Or put another way, **can we rely upon identical median values to imply identical home values for all sale categories?**

Recently, the median has been brought under attack because it is possible to construct two series, such as the ones above, **having the same median but neglecting major changes in the upper-end values of homes**. It is time for us to retreat from the "fun with paper and pencil approach" using only 5 data points and actually enter the real world dealing with thousands of sales.

### *Buyer Demand and Income Level*

In very broad terms, you can think of the frequency sales chart shown earlier as a **reflection of the frequency chart of the income level of the buying pool in a given community**. In general, buyers buy homes in keeping with their income level. **It is no accident that there is a close correlation between Median prices of a community and Median income.**

This pattern is dynamic and extends to all price categories. **A general increase in wage income can increase buyer demand pushing prices up**. This will generally affect all percentile groups.

Now homes in the upper 5% group represent people with very high incomes. This category can change against the rest of the categories because home choices are ***not as reliant*** on their income as in other categories.

### *The Domino Effect*

One of the ***main dynamic mechanisms*** at work in the market to keep prices adjusted is called the ***Domino Effect***. Here's how this works. Let us say there is an ***excess*** of inventory in the ***\$300,000*** price range and owners decide to drop prices by ***10%*** to ***move them***. This puts them at about \$275,000.

But ***there exists homes already in this price category*** that are now in ***competition*** with these new homes that were priced higher. A \$275,000 home ***cannot compete*** with a \$300,000 home. If it could, ***it would have been priced at that level to begin with***.

So the only way for these homes to compete is for the owners to also lower ***their*** price. The process now starts over again, spreading down the price ranges to ***eventually encompass the entire inventory***.

We can see it works in the other direction also. If there were a sudden ***increase in demand*** in homes priced at say, ***\$125,000***, prices would rise in response. But there are already homes on the market at the increased price level. ***Since they offer more value owners can increase their prices in response***. This continues up the latter eventually affecting all prices.

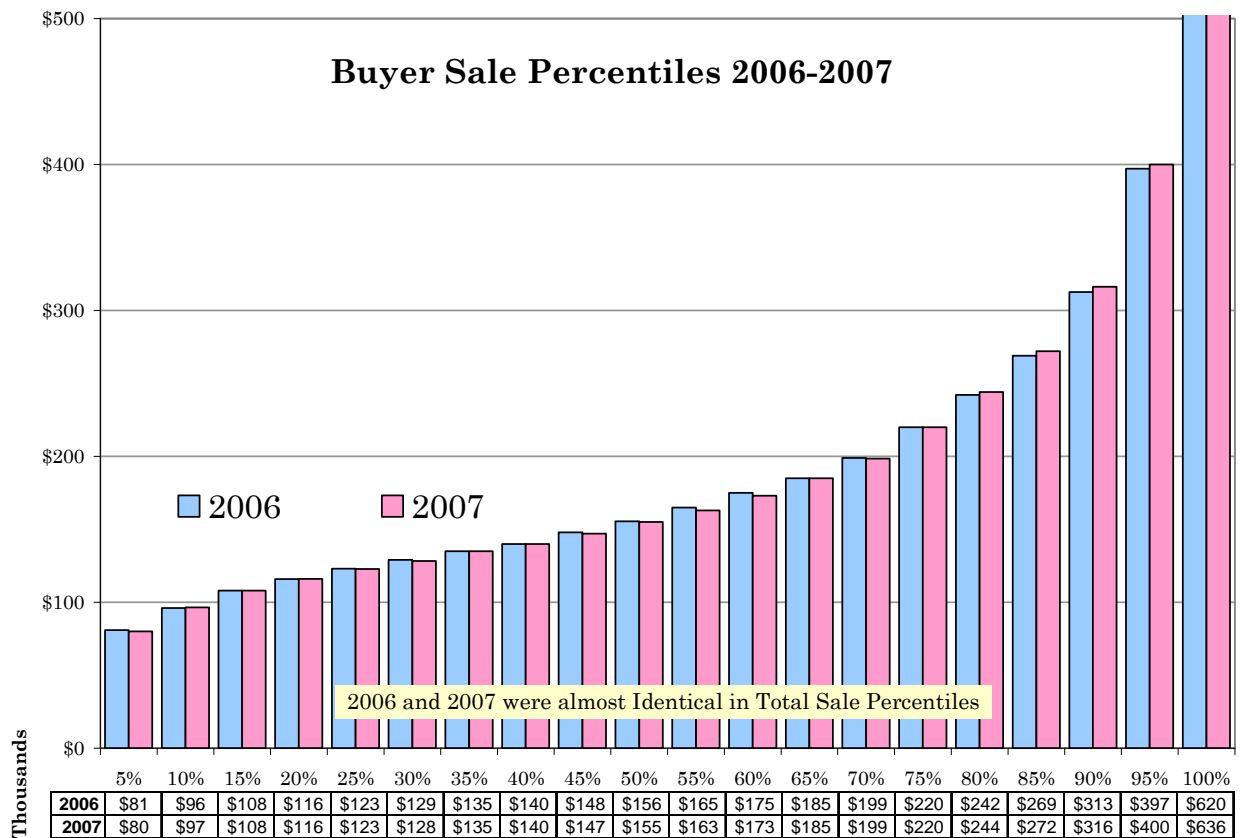
Actually, the dynamic is a ***bit more complex*** than this. A general demand is not usually changing in one particular price range ***but in many at the same time***. Buyers moving into the area have a certain level of income which automatically becomes demand at a corresponding price level. Gluts in the market at other levels bring price reductions adding to a downward shift in prices in other categories.

The point is: ***The real estate market is dynamic and organic***. It is composed of many forces pushing and pulling it in different directions at any given time. ***One particular price range or percentile group may be affected by a temporary demand, but eventually the entire market will feel the impact***.

***Although it is possible to invent imaginary market scenarios behind closed doors and marvel at our creation, in the real world, the real estate market has a mind of its own and often refuses to function as our fantasy dictates***.

Again, one of the arguments against using the median price to compare annual markets is the belief that one particular percentile group or category may be the same as last year, while others may differ significantly.

Let's look at Northwest Arkansas broken into annual percentile groups:



In 2007 Northwest Arkansas had **6,614** residential sales. The Median sale was \$155,000. In 2006 there were **7,838** sales and the Median sale was **\$155,500**. Now obviously, the volume was down in 2007. But the question we want answered is: ***did home owners loose value in their home in 2007 or did they gain?***

The values listed under the percentages in the categories above are the upper-end of what buyers paid for homes. ***A glance at the chart would seem to indicate that the market changed very little from 2006 to 2007.*** The Median values are very close, while many of the other percentiles have identical upper values.

***The largest changes seem to lie in the upper 20% group where buyers actually paid more for homes in 2007 than they did in 2006!***

Even to a critical observer, one would have to admit that the market appears pretty "flat" in comparing the two years. But there is one essential factor we are leaving out. Square Footage.

#### A HIDDEN SOURCE OF VALUE: SQUARE-FOOTAGE

Consider two homes, both selling for **\$150,000**. Are they equally valuable? We could say as long as they were both offered on the market for a reasonable time period and enough buyers looked at them, then, perhaps yes, they are equally valuable. (Although not all buyers will see them as equal).

But suppose one had a square footage of **1,625** while the other had a square footage

of 1,700. Assuming similar ages, neighborhoods, conditions and amenities, *the home with the higher square-footage was offered at a better price. Or, the same money bought a bigger home.*

When we are evaluating the upper-end values of the percentile groups, we must also hold in mind, that if a buyer in the same percentile group paid the same money this year as last year, but received more square-footage, *they actually got a price reduction and the price for that one home went down.*

## Putting it all Together

### Price and Square-Footage Adjustments

So in order to really get as accurate as possible in answering our question of home values rising or falling, we must include square-footage into our calculations.

Here was the approach I took:

- For each percentile group, I found the *mean average* of the prices paid.
- A change in the price of a percentile group shows either an *average* gain or loss *due to pricing*.
- For the same group, I found the *mean average* of the square-footage.
- A change in square-footage in a percentile group shows either an *average* gain or loss *due to square-footage*.
- The *combined* gain or loss indicates the *total gain or loss for a given percentile group*.
- The average gain or loss of the entire percentile series gives the *gain or loss of the 2007 market over 2006*.

### Annual Comparison • Median Percentile Sales and Square Footage • Gains and Losses

Percent of Sales	Average of Sale Percentiles		Average SqFt			Average P/SqFt		Total Gains and Losses			
	2006	2007	2006	2007	G/(L)	2006	2007	By Price	By Sq/Ft	Total	Percent
0%	0	0	0	0	0	0	0	0	0	0	0
5%	\$61,687	\$59,752	1,120	1,108	12	\$55.1	\$53.9	-\$1,935	\$671	-\$1,264	-2.0%
10%	\$89,158	\$88,502	1,179	1,187	-9	\$75.6	\$74.5	-\$657	-\$669	-\$1,326	-1.5%
15%	\$101,817	\$102,097	1,240	1,250	-10	\$82.1	\$81.7	\$279	-\$853	-\$574	-0.6%
20%	\$112,258	\$112,353	1,288	1,330	-42	\$87.2	\$84.5	\$95	-\$3,672	-\$3,577	-3.2%
25%	\$119,491	\$119,652	1,341	1,368	-27	\$89.1	\$87.5	\$161	-\$2,399	-\$2,238	-1.9%
30%	\$126,095	\$125,346	1,393	1,395	-2	\$90.5	\$89.9	-\$749	-\$165	-\$913	-0.7%
35%	\$131,716	\$131,201	1,449	1,447	2	\$90.9	\$90.7	-\$516	\$207	-\$308	-0.2%
40%	\$137,703	\$137,700	1,515	1,515	1	\$90.9	\$90.9	-\$3	\$60	\$57	0.0%
45%	\$144,402	\$143,588	1,589	1,548	40	\$90.9	\$92.7	-\$814	\$3,653	\$2,839	2.0%
50%	\$151,613	\$151,329	1,679	1,676	3	\$90.3	\$90.3	-\$283	\$306	\$22	0.0%
55%	\$160,262	\$159,452	1,687	1,691	-4	\$95.0	\$94.3	-\$810	-\$410	-\$1,220	-0.8%
60%	\$169,463	\$167,562	1,807	1,751	55	\$93.8	\$95.7	-\$1,902	\$5,171	\$3,269	1.9%
65%	\$180,549	\$178,828	1,877	1,848	29	\$96.2	\$96.8	-\$1,721	\$2,789	\$1,068	0.6%
70%	\$191,941	\$191,402	1,959	1,956	3	\$98.0	\$97.9	-\$538	\$258	-\$280	-0.1%
75%	\$208,811	\$208,365	2,094	2,104	-10	\$99.7	\$99.0	-\$446	-\$1,001	-\$1,447	-0.7%
80%	\$230,303	\$230,717	2,219	2,283	-64	\$103.8	\$101.1	\$414	-\$6,629	-\$6,215	-2.7%
85%	\$254,885	\$256,941	2,405	2,430	-25	\$106.0	\$105.7	\$2,056	-\$2,641	-\$585	-0.2%
90%	\$288,227	\$292,089	2,648	2,763	-115	\$108.9	\$105.7	\$3,862	-\$12,493	-\$8,631	-3.0%
95%	\$351,300	\$355,702	2,991	3,129	-138	\$117.4	\$113.7	\$4,402	-\$16,179	-\$11,778	-3.4%
100%	\$547,701	\$546,677	3,887	3,861	27	\$140.9	\$141.6	-\$1,024	\$3,745	\$2,721	0.5%

Figures in red are negative and reflect 2007 average home losses over 2006

### *Total Gain and Losses*

This table summarizes the **gain or loss** for each percentile group. Notice the headings along the top. In the first column are the percentile groups reported by 5% increases. In 2007 there were **331** buyers in each percentile and in 2006 there were **392**.

The second and third columns report the **average home sale price** for each group and for both years. Remember, these are not the upper-end values. **These are the averages within the percentile groups.**

The fourth and fifth columns report the average square-footage for each group and each year. Columns six and seven calculate the price per square-foot for each year, by **dividing the average sale price by the average square-footage in each percentile group.**

Columns eight through eleven show the gain or loss for each percentile group. Column eight calculates the gain or loss by average sale price. Column nine does the same for average square-footage. Column ten totals both gains and losses, and column eleven **shows the final gain or loss in percentages.** Calculations use 2006 figures to compare gains or losses. **Figures in red indicate a loss in home value to the seller.**

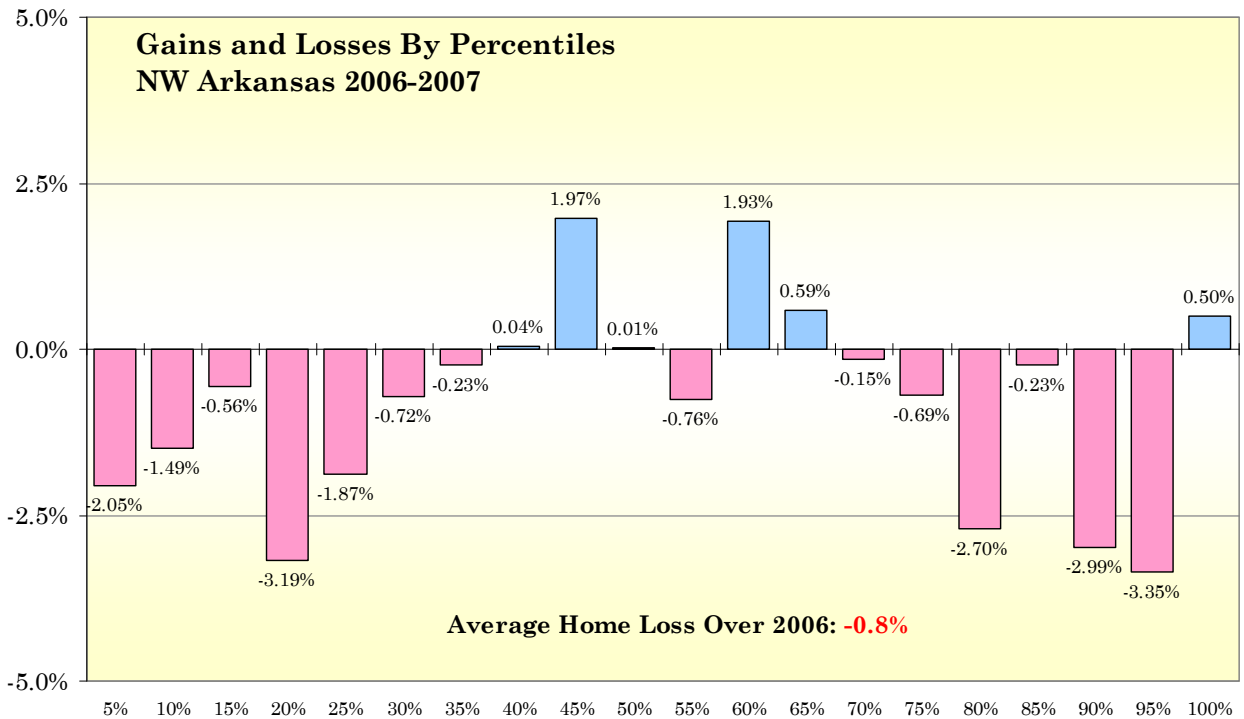
Let's follow an example. In the **20%** category column 8 tells us there was only a **\$95** difference between both years, and the average price was higher in 2007 than 2006. This is an advantage to the seller. However, the average square-footage increased by **42**. Using the 2006 figure of **87.2** dollars a square-foot and multiplying this by **42** gives a gain to the buyer, but a loss to the seller of **\$3,672** over 2006. In other words, homes in the **20** percentile group gave an average of **42** square-feet in concessions over 2006.

Adding the **\$95 gain** due to **increased prices** paid by the buyer to the **\$3,672 loss suffered by the seller** due to **increased square-footage concessions**, gives a net loss of **\$3,577**. Dividing the average price in this group in 2006 into this loss, gives us a net loss of **3.2%**

Finally, taking an **average of all the gains and losses** from **all** the percentile groups gives us a net loss of **.8%** for the **entire market.**

Now frankly, this is **statistically insignificant.** That is, the difference in home values between 2006 and 2007 for the entire market is **practically unchanged.** Make no mistake, the volume is down. **But the volume being down by 16% is due to a decrease of offers, not a general decrease in home values.**

## Gains and Losses Chart



Six categories had a gain over 2006. The percentile group that suffered most was the 95% group. This was due, not to a decrease in prices, but square-footage concessions that averaged 138 sq-foot per home.

### Is the Median Worth It?

So, have we answered the question about the Median? Is it worth considering?

Using the Median alone and comparing the Median of 2007 at \$155,000 to the Median of 2006 at \$155,500, we would say the market decreased by .3% while using the more involved, but perhaps the more accurate method show by the table gives a .8% decrease in the market.

***Believe me; no market analyst would ever see the difference as that significant.***

Until next time and by the numbers

Paul R. Bynum  
 Market Analyst, Principal Broker, GRI, CRS  
 Mount Data Real Estate

Sources:

*All data was from the regional MLX system, and is accurate as of 1/18/08. Late reportings will make small changes to some of the figures, but not enough to change the conclusions of this report.*