# Section U

STATISTICS AND THE APPRAISAL PROCESS

PREFACE

Like many of the technical aspects of appraising, such as income valuation, you have to work with and use statistics before you can really begin to understand what they tell you about your data. The point is that just because you are not familiar with these tools, do not be hesitant in trying a few simple ones as you will soon gain mastery thereof and seek out new and better tools.

STATISTICS AND THE APPRAISAL PROCESS

**INTRODUCTION**

Statistics offer a way for the appraiser to qualify many of the heretofore qualitative decisions which he has been forced to use in assigning values. In the process, he can learn more about how the data he uses behaves as well as how it relates to the property valuation at fair market.

This brings us to the definition of that word "STATISTICS". A statistical measure or "statistic" is a tool that helps you better describe the characteristics of a set of data, such as the relationship of sale price to appraised value.

While useful, a far more technical and comprehensive definition is appropriate rather than the more simplistic one given above, namely, "statistics is the theory and method of analyzing quantitative data obtained from samples of observations in order to study and compare sources of variance of phenomena, to help make decisions to accept or reject hypothesized relations between the phenomena, and to aid in making reliable inferences from empirical observation." The preceding, from FOUNDATIONS OF BEHAVIORAL RESEARCH by Fred N. Kerlinger, states very well what statistics are, their usefulness, and implications for our work. His book is highly recommended to all who wish to gain an understanding of many statistical tools and the requisite knowledge of the "scientific method" of constructing cases for analysis. A somewhat less advanced text for the beginner is AN INTRODUCTION TO BUSINESS AND ECONOMIC STATISTICS by John R. Stockton.

It is not our intent to try and present a programmed text to teach statistics but we will hopefully indicate which are useful where and what they tell the property appraiser about his values.

## STATISTICS AND THE APPRAISAL PROCESS

Sales offer the only real set of data which can be established as indicating market value for properties. Appraisals which are done to supplement sales as parcels to which one may relate for purposes of comparison are merely attempts to predict what the sales price would be that parcel should actually sell. It is our belief that surrogates for actual sales are needed only when parcels (for a class) show a statistically insignificant number of sales.

Particularly for single family residential properties sales are usually always available and are in most cases legitimate arm's length transactions.

The most frequently asked question is usually "Where am I in relation to market?" There are ways of describing this relationship; each of which will help you understand "where" you are in relation to the market.

Level of assessment in relation to market is one part of the answer. It is usually expressed as a ratio of appraised values to sale values. Common measures of this ratio, overall, for a county are called "MEANS", "MEASURES OF CENTRAL TENDENCY", or "AVERAGE".

## SIMPLE OR UNWEIGHTED MEAN

This measure is found by dividing the sum of all individual sales by the number of sales. That is, given the following hypothetical list of sales, compute the means:

| OBSERVATION NUMBER | SALEPRICE | APPRAISED VALUE | SALES RATIO |
|---|---|---|---|
| 1 | $22,600. | $21,500. | 95 % |
| 2 | 31,000. | 28,600. | 92 |
| 3 | 37,800. | 34,000. | 90 |
| 4 | 38,400. | 33,000. | 86 |
| 5 | 34,300. | 29,500. | 86 |
| 6 | 20,000. | 16,000. | 80 |
| 7 | 13,000. | 9,800. | 75 |
| 8 | 18,700. | 13,500. | 72 |
| 9 | 26,900. | 17,200. | 64 |
| 10 | 40,800. | 24,500. | 60 |
| | $283,500. | $227,600. | 800 % |

Mean Sale Ratio = 800/10 = 80%.

Mean Appraised Value = $227600/10 = $22,760.

Mean Sales Price = $283500/10 = $28,350.

As you can see, there are several "MEANS" which may be computed; each of which is an expression of central tendency.

There is another type of mean called a WEIGHTED MEAN which reflects the impact of the dollar magnitude of the values in the calculation of the mean. It is obtained by dividing the total of all appraised (or assessed) values by the total of all sales prices. For example:

$227,600/$283,500 = 8.3%

or in the previous example:

TOTAL ASSESSED VALUE/TOTAL SALES PRICE = weighted mean

This measure is affected by large values which have a proportionately greater impact on the ratio than smaller values. As a general rule, this measure is, therefore, somewhat less useful for sales ratio work than the un-weighted mean.

A highly useful statistic is the MEDIAN. It is a measure which is least influenced by extreme values as it is based upon position rather than on level. That is, it is the value half-way from either end of a list of values when the list is arrayed in ascending (or descending) order. If the list contains an odd number of sales then the median is the middle value in the list. However, if there are even numbers of sales in the list then it is the average of the two values on either side of the theoretical mid point in the list. Using our example it is:

MEDIAN = (TOTAL NUMBER OF SALES + 1) / 2 + (10 + 1) / 2 + 5.5th item in the list

| That is in our list: | Sales | Sales Ratio |
|---|---|---|
| | 1 | 95% |
| | 2 | 92 |
| | 3 | 90 |
| | 4 | 86 |
| | 5 | 86 |
| Median 5.5 Sales----------> | | |
| | 6 | 80 |
| | 7 | 75 |
| | 8 | 72 |
| | 9 | 64 |
| | 10 | 60 |

The median is, therefore, halfway between the ratio 86 and 80 or:

MEDIAN = (86 + 80) / 2 = 166 / 2 = 83%

This statistic is generally not usable in more advanced mathematical manipulations; however, it is useful because it does enter into the total concept of data and is useful in judging uniformity and level of assessment. (Note: you may also calculate a median sales value as well as a median appraised value.)

**MODE**

The mode is a measure of central tendency that is easy to understand. It is the value in the set of observations which occurs most frequently. In our example, the mode of sales ratios would be 86% (occurs 2 times).

## MEASURES OF VARIABILITY

A classic example of reliance on the use of the mean only as a method of description may be rather graphically illustrated by the following:

If you were fired upon one time and were missed by 100 yards and were fired upon a second time and were hit, you could conclude that you were missed by an average of 50 yards.

The point is, the mean does not tell the whole story about the data. Other tools are needed to better describe the data. These tools are measures of how much you miss the mean (in general) or in more technical terms, measures of dispersion.

## RANGE

The range is simply the lowest and highest value in your set of observations subtracted from one another; although it may be reported as the minimum and maximum values themselves. In our example, you could say the range (for the sales ratios) is:

35% or from 60% to 95%

As a general statement it is not too useful in analysis due to its obvious dependence on extreme values.

## MEAN DEVIATION & MEDIAN DEVIATION

This measure is the average of the difference between the mean (or median) and the individual observations.

MD = [d] / N or [x] / N

That is, the mean or median deviation is the sum of the absolute value of the differences between the mean (or median) and each observation divided by the number of observations. (Absolute value means the signs are ignored, that is assumed to be positive, when accumulating [x] or [d].)

For our example:

| SALES RATIO | - | MEAN | = | [x] ([d] is used for the median) |
|---|---|---|---|---|
| 95 | - | 80 | = | 15 |
| 92 | - | 80 | = | 12 |
| 90 | - | 80 | = | 10 |
| 86 | - | 80 | = | 6 |
| 86 | - | 80 | = | 6 |
| 80 | - | 80 | = | 0 |
| 75 | - | 80 | = | 5 |
| 72 | - | 80 | = | 8 |
| 64 | - | 80 | = | 16 |
| 60 | - | 80 | = | 20 |

Hence: MD = 98 / 10 = 9.8%

This ratio expresses the average amount by which the data varies from the mean (or median) in a particular set of data. It is influenced by extremes as is the mean and even when computed about the median, it is likewise influenced. It also is not useful in making further statistical analysis of the data.

## STANDARD DEVIATION

To overcome the handicaps of the mean deviation, the standard deviation is used. It is a numerical measure of the degree of dispersion, variability, or non-homogeneity of the data to which it is applied.

In calculation, it is similar to the average deviation but differs in its method of averaging differences from the mean. It does this by squaring each difference and eventually summing all squared differences averaging them and taking the square root thereof giving an "average deviation" from the mean.

In practice it is quite easy to compute using a handy "working formula" to make the task easier. First the formal formula:

STANDARD DEVIATION = $\sqrt{\frac{\Sigma(X-U)^2}{N}}$ or $\sqrt{\frac{\Sigma(x-u)}{N-1}}$ Where u = "mu" (arithmetic mean)

$$\sqrt{\frac{\text{Sum of the individual differences squared}}{\text{Number of observations}}}$$

The second formula using N-1 is most often used when dealing with sample data and is used in our sales ratio reports.

In our example, using sales ratios it would be:

| Observation | X | (X-u) | $(X-u)^2$ |
|---|---|---|---|
| 1 | 95% | 15 | 225 |
| 2 | 92 | 12 | 144 |
| 3 | 90 | 10 | 100 |
| 4 | 86 | 6 | 36 |
| 5 | 86 | 6 | 36 |
| 6 | 80 | 0 | 0 |
| 7 | 75 | 5 | 25 |
| 8 | 72 | 8 | 64 |
| 9 | 64 | 16 | 256 |
| 10 | 60 | 20 | 400 |

X = 800% $(X-u)^2$ = 1286

Arithmetic Mean (u) Sales Ratio = 800 / 10 = 80%

Hence: SD = $\sqrt{\frac{(X-u)^2}{N}}$ OR SD = $\sqrt{\frac{\Sigma(X-u)^2}{N-1}}$

$\sqrt{\frac{1286}{10}}$ = $\sqrt{\frac{1286}{10-1}}$

$\sqrt{128.6}$ = $\sqrt{142.89}$

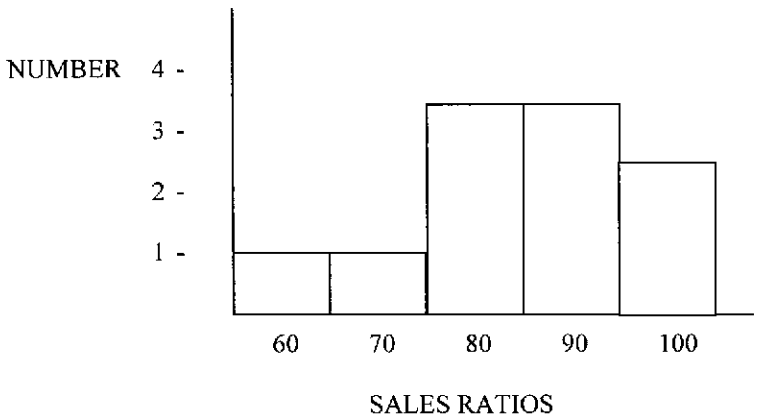
$\sqrt{11.34}$ = $\sqrt{11.95}$

The standard deviation is useful in that it is logical mathematically and may hence be used satisfactorily in further calculations.

## FREQUENCY DISTRIBUTIONS

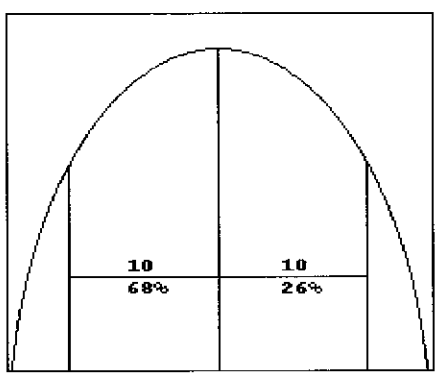
This is a good time to discuss distributions. All frequency distributions are an arrangement of numerical data according to size or magnitude. Distributions are normally presented as tables or graphs. The following table and graph is taken from our example:

| SALES RATIO CLASS INTERVAL | NUMBER OF OCCURENCES |
|---|---|
| 91 - 100 | 2 |
| 81 - 90 | 3 |
| 71 - 80 | 3 |
| 61 - 70 | 1 |
| 51 - 60 | 1 |
| | 10 |




When describing our observations, we really are trying to use numbers [mean, median, mode, standard deviation, average deviation, etc.] to give a mental picture of what our frequency distribution would look like if we drew it on a graph.

A particularly shaped distribution is the one from which we depart when trying to visualize the shape of a distribution when given such statistics as the mean, median and mode for information. The reference point is what is called the "NORMAL DISTRIBUTION". It has some particular features by which it is characterized and referred to. This is what it looks like:




"Normal" Distribution Showing the Percentage of the Area Included Within One Standard Deviation Measured Both Plus and Minus About the Arithmetic Mean.

The MEAN, MEDIAN, and MODE are all equal. It also possesses some traits which make it statistically useful in making decisions about differences in distributions.

One of these properties is that one may determine what percent of the observations lie within; one, two, or three times the calculated standard deviation by using pre-computed tables. (In fact, any fractional part of the standard deviation may also be used.)

The way it would likely be useful to you is in making a statement about the uniformity of your values which is in part what it measures. For instance, if you have a set of sales with a mean of 87% and a Standard Deviation of 10%, you could conclude that 95.46% of all sales would fall between the limits of 75.46% and 115.46%. Extrapolating that sales represent the rest of the parcels in your county (we leave the question of the validity of this assumption up to you), you could then have some mental picture of how your county roll values would distribute themselves in relation to the market values of the parcels.

For all the statistically astute, we do include two things: (1) remember that the distribution must be normal or approximately so for this to be true and (2) if there is ever a source of disagreement, sales ratio studies are surely prime material. However, we will let the relative merits of the case go untouched in this text.

One final word on the description of a distribution. When you first begin to work with these tools, please get a simple straight forward text such as one of the "cram course" texts on statistics available in any college bookstore with an appealing title such as STATISTICS MADE SIMPLE, etc. You will find it most useful in attacking problems. One we recommend is available from Barnes & Noble in their college outline series titled "STATISTICAL METHODS".

**RELATIVE MEASURE OF VARIATION**

Handy statistical tools are the relative measures. They are ways of relating back to the mean or median in discussing the degree of variance in a set of observations. Three common ones are:

$$\frac{\text{AVERAGE DEVIATION ABOUT THE MEAN}}{\text{MEAN}} \times 100 = \text{Coefficient of dispersion of the average deviation}$$

$$\frac{\text{STANDARD DEVIATION}}{\text{MEAN}} \times 100 = \text{Coefficient of dispersion of the standard deviation}$$

$$\frac{\text{STANDARD DEVIATION ABOUT THE MEDIAN}}{} \times 100 = \text{Coefficient of dispersion of the median deviation}$$

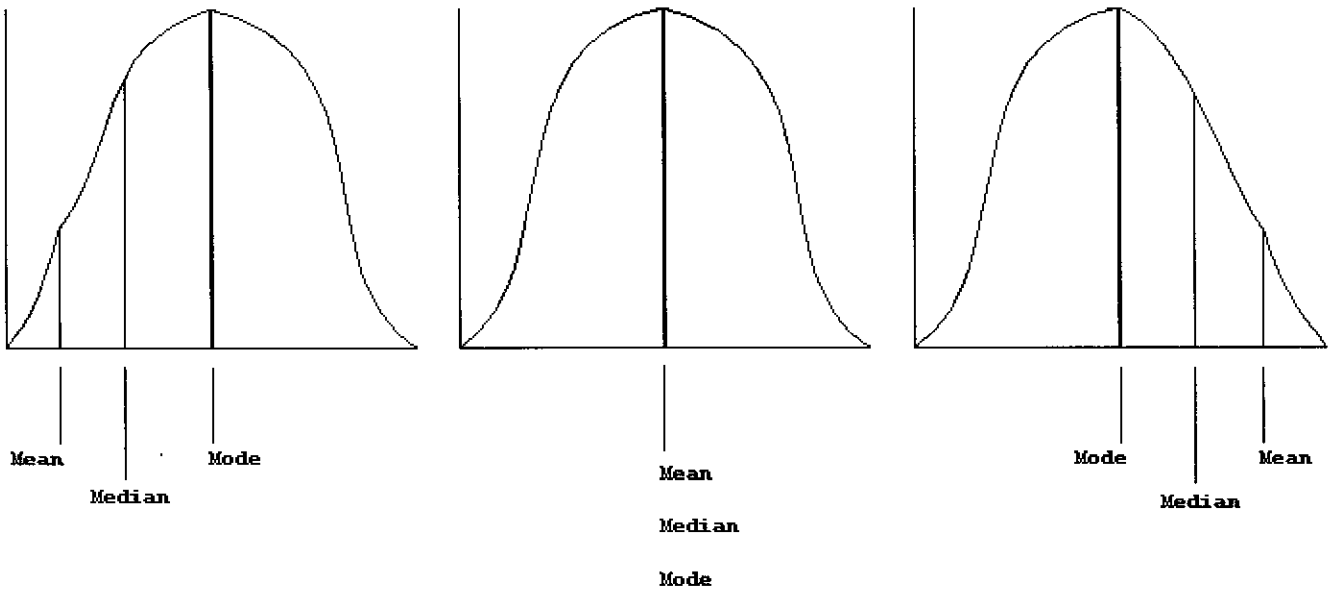
The last two yield the most useful statistic in that the standard deviation is significant in appraising in relationship to the level as there are few who would want a ratio to go consistently over 100% (which is one use of the standard deviation) or whom would want a mean of 70% with a relative error of 35% on 68% of all parcels.

**SHAPE**

How do you describe the shape of a distribution? Well, we have used the mean, median, mode, average and standard deviation. We also would like to be able to tell the extent to which our values were consistently biased, either high or low. The statistics measuring this are the coefficients of skewness. That is, a measure of the degree to which the distribution departs from the normal distribution.

There are three, more or less, classic shapes a distribution may take (although it may look like anything!) They are:

SKEWED LEFT NORMAL SKEWED RIGHT




Skewness is a term for the degree of distortion from symmetry exhibited by a frequency distribution. What this means is that if you were to graph the sales ratios you would expect that all errors should be random and hence symmetrical and not biased either low or high for certain properties. This can be checked by using the common measures of degree of skewness.

$$SK_1 = \frac{3\,(\text{MEAN-MODE})}{\text{STANDARD DEVIATION}}$$

Note: (Pearson's Coefficient of Skewness)

and

$$SK_2 = \frac{(Q3 - \text{MEDIAN}) - (\text{MEDIAN} - Q1)}{(Q3 - Q1)}$$

The second measure uses a "QUARTILE" which is something like the median (in fact, the median is the Q2 or second quartile or quarter, EG 50% of the way through the list, item) but is the item 25% (Q1) down the list and the 75% (Q3) item down the list of ordered observations and may be determined much as is the median.

**NON PARAMETRIC STATISTICS**

This class of statistics is useful in that unlike many statistical tools, they do not depend on having normally distributed values to be meaningful.

The most usable is the chi-squared statistic. It is simple and is very useful in testing a number of common questions or hypotheses which you pose formally or informally in appraising.

Suppose, for instance, you have collected a set of observations of the sale parcels in an area and you wish to compare the distribution of these sales with the distribution of all parcels for the area to see if the distributions match up and will give you some assurance that the sales are comparable to the universe of all parcels. To do this let us assume you use a single method of classification, age, and restrict the discussion to only a single exterior wall type (a good discriminator).

How do you proceed? First classify the sale parcels into groups of 5 years although the greater of lesser intervals could have been selected depending on our data. For example:

TABLE OF ACTUAL FREQUENCIES
FOR SALE PARCELS

| AGE (in years) INTERVAL | FREQUENCY IN NUMBER | PERCENT OF TOTAL |
|---|---|---|
| 1 - 5 | 10 | 13.2 |
| 6 - 10 | 22 | 28.8 |
| 11 - 15 | 17 | 22.4 |
| 16 - 20 | 10 | 13.2 |
| 21 - 25 | 7 | 9.2 |
| 26 - 30 | 10 | 13.2 |
| | 76 | 100.0% |

Then classify all parcels for the area into groups of a like interval used with the sale parcels. For example:

TABLE OF ACTUAL FREQUENCIES
FOR SALE PARCELS

| AGE (in years) INTERVAL | FREQUENCY IN NUMBER | PERCENT OF TOTAL |
|---|---|---|
| 1 - 5 | 128 | 12.2 |
| 6 - 10 | 234 | 22.4 |
| 11 - 15 | 355 | 33.9 |
| 16 - 20 | 139 | 13.3 |
| 21 - 25 | 87 | 8.3 |
| 26 - 30 | 104 | 9.9 |
| | 1,047 | 100.0% |

The question we really want to ask is are the two distributions the same (in the sense that the distribution of parcels by age makes them equal for purposes of judging similarities) or are the distributions different. To answer this, we must consider the element of chance. It is possible that the sales are distributed like the total area but show difference in cell frequencies due to chance alone, for as you may observe, the percentages of the total by age are indeed different.

We would expect the sales to be distributed in like frequencies as the total area was distributed unless the sales do not represent the area under study.

The use of a very handy tool, the statistic known as the CHI-SQUARE ($X^2$) test, is worth learning. It is useful in that it does not require that one have normally distributed data to be valid; hence it is non parametric. It is used by taking an expected frequency and comparing it to the actual or observed frequency. In our case, it is the area parameters projected upon the sales data.

We would expect the number of sale parcels per age group to be the same as the frequencies observed for the total of all parcels in the hypothetical area under consideration. Therefore, we use the percentages for the total to generate the expected number of sales for each age interval.

The CHI-SQUARE statistic expressed as a formula is:

$$x^2 = \sum [(fo-fe)2/fe]$$

where fo = frequency observed
fe = frequency expected

Example:

| PERCENT OF TOTAL PARCEL | x | TOTAL SALES | = | EXPECTED NUMBER OF SALES IN EACH INTERVAL |
|---|---|---|---|---|
| 12.2 | | 76 | | 9.3 |
| 22.4 | | 76 | | 17.0 |
| 33.9 | | 76 | | 25.8 |
| 13.3 | | 76 | | 10.1 |
| 8.3 | | 76 | | 6.3 |
| 9.9 | | 76 | | 7.5 |
| 100.0% | | | | 76. |

The actual number of sales in each interval is set down. One then subtracts the estimated number from the observed number of sales, interval by interval, squaring the result and dividing by the expected number.

Example:

| GROUP | OBSERVED FREQUENCY | EXPECTED FREQUENCY | OBSERVED MINUS EXPECTED | SQUARED RESULT | DIVIDED BY EXPECTED |
|---|---|---|---|---|---|
| 1 | 10 | 09.3 | 0.70 | 00.49 | 0.053 |
| 2 | 22 | 17.0 | 5.00 | 25.00 | 1.471 |
| 3 | 17 | 25.8 | 8.80 | 77.44 | 3.002 |
| 4 | 10 | 10.1 | 0.10 | 00.10 | 0.010 |
| 5 | 07 | 06.3 | 0.70 | 00.49 | 0.053 |
| 6 | 10 | 07.5 | 2.50 | 06.25 | 0.833 |
| | | | | $X^2$ = | 5.422 |

The number 5.422 is the chi-square for this comparison. It is evaluated based upon what is known as DEGREES OF FREEDOM of the problem and the use of a table of chi-square values common to most statistics texts. We may say here that "degrees of freedom" means the latitude of variation a statistical problem has. It is the number of groups (Nk) minus 3 or V = ($N_k$ - 3). In this case V = 3.

Consulting our table, we find that the probability of having a chi-square due to chance of 5.42 is approximately .75 or sufficiently different from .95 for us to state that the sales do differ significantly from the actual distribution of all parcels. Hence, we would conclude that we should be careful in the extrapolation of sale parcel statistics to the entire distribution of all parcels.