POLI 300 ***Handout #6*** N. R. Miller

MEASURES OF CENTRAL TENDENCY (AVERAGES)

A *measure of central tendency* is a *univariate* statistic that indicates, in one manner or another, the *average* or *typical* observed value of a variable in a data set or, put otherwise, the *center* of the frequency distribution of the data. We here consider three measures of central tendency. (there are others) appropriate for different levels of measurement (nominal, ordinal, and interval/ratio).

The Mode

The *mode* (or *modal value*) of a variable in a set of data is the value of the variable that is observed *most frequently* in that data (or, in a continuous frequency density, is at the point of *greatest density*).

Note. The mode is the value that is observed most frequently, not the frequency itself. (I see this error too frequently on tests.)

The mode is defined for *every* type of variable [i.e., *nominal*, *ordinal*, *interval*, or *ratio*]; but the mode may be ill-defined if we have either:

(i) a small number of cases; or
(ii) a precisely measured continuous variable and a finite number of cases;

because in either event it is likely that no value will be observed more than once in the data.

In any event, the mode is *unstable* in that

(i) small changes in the data can result in large changes in the modal value; and

(ii) changes in coding the variable (for example, RELIGION), or changes in class intervals, can change the modal value.

How To Calculate the Mode

(a) *Given a list of observed values (raw data)*: Construct a frequency table and go to (b).

(b) *Given a frequency table (or bar graph)*: Observe which value in the table (or graph) has the *greatest* (absolute or relative) *frequency*; the most frequent *value* (*not* the *frequency* itself) is the mode.

(c) *Given a continuous frequency curve*: The mode is the value under the *highest point* of the frequency curve (the point with the greatest density of observations).

The Median

The *median* (or *median value*) of a variable in a data set is the value in the *middle* of the observations, in the sense that (no more than) *half* of the cases have *lower* values and (no more than) half of the cases have *higher* values or, more generally, such that (no more than) half of the cases have values that lie on either side of the median value. (That is, the median value stands at the 50th percentile. See *Percentiles, The Median, and Ranges* on the back page of this handout.)

The median is defined if and only if the variable is *at least ordinal* in nature [i.e., *ordinal*, *interval*, or *ratio*], and we can therefore rank all (non-missing) observations in terms of lower to higher values or some other natural (e.g., liberal to conservative) ordering.

Note. The median should be clearly distinguished from another (infrequently used) measure of central tendency that is defined only for variables that are interval in nature. This is the *midrange* value, which is the value in the middle of the observations in the (different) sense that it lies exactly halfway between the *minimum* (lowest) and *maximum* (highest) observed values (i.e., at the *midpoint* of the range of values), i.e., (*min* + *max*)/2.

How To Calculate the Median

(a) *Given a list of observed values (raw data)*: Rank order the cases in terms of their observed values (e.g., from lowest to highest); the value of the case at the *middle of this rank-ordered list* is the median; *or* construct a frequency table and go to (b).

Note. If the number of cases is even, there is no observation at the exact middle of the list. Rather we look at the two observations closest to the middle of the list. If they have the same value, that value is the median. If they have different values, every value in the *interval* bounded by these two values meets the definition of a median; but conventionally the median in this event is defined as the *midpoint* of the interval. (See for example, *Percentiles, The Median, and Ranges* on the last page of this handout.)

(b) *Given a frequency table (or bar graph)*: Calculate the cumulative frequencies (in either direction); the median value is the *first value whose cumulative frequency exceeds 50%*. (You will get the same answer cumulating in either direction, unless the number of cases is even and you run into the problem discussed in the preceding note.)

(c) *Given a continuous frequency curve*: Draw a vertical line such that *half* of the area under the frequency curve lies *on one side* of the line and *half on the other side*. The median value of the variable lies on this line.

The Mean

The *mean* (or *mean value*) of a variable in a set of data is the *result of adding up all the observed values of the variable and dividing by the number of cases* (i.e., the "average" as the term is most commonly used).

The mean is defined if and only if the variable is *at least interval* in nature [i.e., *interval* or *ratio*].

Suppose we have a variable *X* and a set of cases numbered 1,2,...,*n*. Let the observed value of the variable in each case be designated x_1, x_2, etc. Thus:

$$mean\ of\ X = \bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum x}{n}\ .$$

How to Calculate the Mean

(a) *Given a list of observed values (raw data)*: Following the formula above, add up the observed values of the variable in all cases and divide by the number of cases.

(b) *Given a frequency table (or bar graph)*: Take each value and multiply it by its *absolute* frequency, add up these products over all values, and divide this sum by the number of cases; *or* (and this is usually easier) take each value and multiply it by the *decimal fraction* (between zero and one) that represents its *relative* frequency, and add up these products over all values. (Do not divide by the number of cases — you have already done this as a result of multiplying each value by a decimal fraction).

(c) *Given a continuous frequency curve* : The mean is the "center of gravity" of the distribution. Determine (by "eyeball" approximation) the value of the variable such that the density "balances" at that point; this value is the mean.

Average Values

1. For a *quantitative* variable, the *range of observed values* of a variable in a set of data is the interval extending from the *minimum* observed value to the *maximum* observed value. For *every* measure of central tendency, the *average* value of the observations lies *somewhere in this range*, often (but certainly not always) somewhere near the midpoint of this range. The modal or median observed value can equal the minimum or the maximum value, but the mean observation can do so only in the very special case in which all cases have identical observed values (so there is no *dispersion* [next topic] in the data). Once again, remember that the median is not (necessarily) the midpoint of the range.

2. If a quantitative variable is *discrete* (if all of its values are whole numbers), the modal or median observed value is always a whole number also (with the possible exception noted above for the median when the number of cases is even), but the mean observation is almost never a whole number. (Thus, the "average" family may have 2.374 children, even though *no* individual family can have that [or any fractional] number of children.)

Deviations from the Average

Unless all cases have the same observed value, some (in the case of mean, probably all) observed values will be different from an average value. (This consideration previews the topic of *dispersion*, to be discussed in Handout #7.) If the variable is (at least) interval in nature, we can speak of the *deviation from the average* in each case, i.e., the "distance" or "interval" from the observed value to the average value. Such a deviation is *positive* if the observed value is greater than the average value, *negative* if the observed value is less than the average value, or *zero* if the observed value is equal to the average value.

Since we have three different types of averages, we also have three different types of deviations from the average. The deviations from each type of average have different properties.

(a) There are *fewer* non-zero deviations from the modal value than from any other value of the variable.

(b) No more than *half* of the deviation from the median are *positive* and no more than *half* are *negative*. Unless several cases have the median value (and perhaps even then), the *number of positive deviations equals the number of negative deviations* — that is, the positive and negative deviations balance out with respect to the *number* of deviations.

(c) The *sum* (or mean) of the *absolute deviations* (i.e., ignoring whether the deviations are positive or negative) *from the median* is *less* than the sum (or mean) of the absolute deviations from any other value of the variable.

(d) The sum (or mean) of the (non-absolute, i.e., taking account of whether deviations are positive or negative) *deviations from the mean is zero*, i.e., $\sum(x - \bar{x}) = 0$. That is, the positive and negative deviations balance out with respect to the *magnitude* of the deviations.

(e) The *sum* (or mean) of the *squared deviations from the mean*, i.e., $\sum(x - \bar{x})^2$, is *less* than the sum of the squared deviations from any other value of the variable.

Median vs. Mean Values

The mode is rarely used to describe the central tendency in quantitative data, because (as we saw) it may be undefined or unstable. However, both the mean and the median are commonly used with quantitative (i.e., interval or ratio) data. While the median (unlike the mean) *may* be used with variables that are merely ordinal, the median *may also* used (along with the mean) with interval (or ratio) variables. For example, it is common to report both median and mean income or wealth, median and mean test scores, median and mean prices (of cars, houses, etc.). It is important to recognize that, while the median and the mean are both proper and useful measures of central tendency, they have different definitions and properties and may (depending on the distribution of the data) give very different answers to the question "what is the average value in this data?" Recall the distinction between a *symmetric* and *skewed* frequency distribution introduced at the end of Handout #5. It has important implications for the median and mean.

1. If the distribution of the data is *symmetric*, the median and mean values are the same (and if the distribution is "almost" symmetric, the median and the mean are "almost the same"). For example, test scores are typically distributed approximately symmetrically, so median and mean test scores are typically approximately the same.

2. However, if the distribution of the data is *skewed*, the mean is pulled (relative to the median) in the direction of the long thin tail. For example, we saw in Handout #5 that income is distributed in a highly skewed fashion, with a long thin tail in the direction of higher income. Thus mean income is typically considerably higher than median income. (The distributions of most ratio variables are typically skewed upward, because values less than zero cannot occur while there may be no definite upper limit on possible values.)

3. More generally, the median (unlike the mean) is "resistant to outliers," where an *outlier* (in univariate data) is a case with an extreme (very high or low) value. That is, adding some outliers to the data (or removing them) may have a big impact on the mean value but usually has little impact on the median value.

4. If the observed values in some distribution of data changes, the median value changes only if the value (or identity) of the median case changes, whereas the mean value most likely is affected by any change in the data. For example, if "the rich get richer while everybody else stays about the same," mean income increases while median income stays the same.

5. However, if changes in the data do not change the sum of all observed values, i.e., if $\sum x$, remains constant, the mean remains constant, while the median may change. For example,

if Congress has decided agrees that a tax cut will total $100 billion but has not decided how this fixed sum should be divided up among the nation's 100 million households, the median benefit will depend on the specifics of the legislation but the mean tax benefit per household will be $1000 in any event.

6. If observed values are *polarized* with approximately half the cases having high values, half having low, and none having medium values, the *median is unstable* — that is, the median value will be high or low depending on whether slightly more than half the cases have high values or slightly more than half the cases have low values, whereas the mean value will barely change as a result of such small shifts in the data.

7. To "see" median and mean values behave in different ways, go to the course web page => On-Line Statistical Demonstrations: Statistical "Applets" => Mean and Median and play around with some examples.

Politics and Averages

Suppose members of a club must decide to how much money to spend on a particular project. Some members are intensely interested in the project and want to spend a lot, while others don't much care about the project and want to spend little if any money on it. Members write down their preferred expenditures on ballots on the understanding that the club will spend the *average* of the reported preferences. But there are two kinds of averages: the mean and the median. If the rule were *to spend the mean reported preference*, members who want to spend a high amount have an incentive to report on their ballots even larger expenditures than they really want, since such misreporting of their preferences pulls the mean upwards and closer to what they truly want. (Members who want to spend only a little can try the mirror-image gambit, but they are restricted by the fact that they [presumably] cannot report a preference of less than $0.) But suppose that the rule is *to spend the median reported preference*. Then club members have no incentive to report anything other than their true preferences. Marking high ballots even higher (or low ballots even lower) than you truly want has no effect on the median, so members with such preference cannot gain by misreporting them. And the voter who casts the median ballot is getting exactly what he or she wants and certainly has no incentive to misrepresent.

Now suppose that a club, committee, or legislature has to collectively choose the value of a variable, e.g., how much to spend on some activity, and members disagree on how much it should be. Members favor proposals that are closer to their most preferred amounts to proposals that deviate more from their most preferred amounts. The only proposal that cannot be beaten by some other proposal in a majority vote is the *median* proposal. This is called the *Median Voter Principle*, and it helps explain why, in a two-party system, both parties tend to advocate moderate (close to the median) preferred policies.

PERCENTILES, THE MEDIAN, AND RANGES

```
(1)   (2)   (3)    (4)                                              (1)    Rank (approximate due to lack of precision in
                                                                           recording variable values)
  1   FL    17.8    99 <=============                               (2)    Name of case (state postal code)
  2   PA    14.8    97                    \                         (3)    Value of variable
  3   IW    14.8    95                     \                        (4)    Percentile (approximate due to small
  4   RI    14.7    93 _____                \                              number of cases and to lack of precision
  5   AR    14.6    91      |                \                             in recording variable values)
  6   SD    14.0    89 _____|<=======         \
  7   WV    13.9    87                  \      \
  8   MO    13.8    85                   \      \
  9   NB    13.8    83                    \      \                       [Mean = 12.1]
 10   MA    13.7    81                     \      \
 11   OR    13.7    79                      \      \
 12   KS    13.6    77                       \      \
 13   CN    13.4    75 <===========           \      \
 14   ME    13.4    73             |           \      \
 15   ND    13.3    71             |            \      \
 16   WS    13.2    69             |             \      \
 17   NJ    13.0    67             |              |      \
 18   NY    13.0    65             |              |       \
 19   OK    12.8    63             |              |        \
 20   AZ    12.7    61             |              |         \
 21   MN    12.6    59             |              |          \
 22   MT    12.5    57             |              |           \
 23   OH    12.5    55             |              |            \
 24   AL    12.4    53 __          |              |             \
 25   TN    12.4    51   | Median | IQ Range     | ID Range     \  Range
 26   KY    12.3    49 __| = 12.35 | = 13.4 −10.7 | = 14.3 −9.85 |  = 17.8 − 3.6
 27   IL    12.1    47             |  =  2.7      | =   4.45     /   =  14.2
 28   IN    12.1    45             |              |             /
 29   MS    12.1    43             |              |            /
 30   VT    11.9    41             |              |           /
 31   NC    11.8    39             |              |          /
 32   WA    11.8    37             |              |         /
 33   DL    11.6    35             |              |        /
 34   ID    11.5    33             |             /        /
 35   MI    11.5    31             |            /        /
 36   NH    11.5    29             |           /        /
 37   LA    10.8    27             |          /        /
 38   MD    10.7    25 <===========          /        /
 39   SC    10.7    23                      /        /
 40   NV    10.6    21                     /        /
 41   CA    10.6    19                    /        /
 42   VA    10.6    17                   /        /                      [SD = 2.16]
 43   HW    10.1    15                  /        /
 44   GA    10.1    13 _____           /        /
 45   NM    10.0    11      | <=======         /
 46   TX     9.7     9 _____|                 /
 47   CO     9.2     7                       /                      The cases are the fifty states.
 48   WY     8.9     5                      /                       The variable is PERCENT ELDERLY (65+)
 49   UT     8.2     3                     /                        The data is from Handout #5, Table 1
 50   AK     3.6     1<============
```