

Sociology 541
Thursday February 1, 2001

Describing the Center of a Data Set

Although a tabular or graphical summary of the data is useful, any further analysis of the data requires the properties of the data to be summarized numerically. One feature of the distribution of our data that we would like to describe is its center. We would like a single value to represent this.

We will use the following data sets to illustrate measures of center.

DATA SET 1 The following are test scores from a class of 20 students:

96 95 93 89 83 83 81 77 77 77 71 71 70 68 68 65 57 55 48 42

DATA SET 2 The same 20 test scores are arranged in a grouped frequency distribution. To illustrate how to obtain measures of central tendency for grouped data, we'll assume that all we know about the data is what is presented in the table. We no longer know the original scores.

The frequency distribution for these test scores is:

Class	Mid-point	Frequency	Cumulative frequency
90-99	94.5	3	20
80-89	84.5	4	17
70-79	74.5	6	13
60-69	64.5	3	7
50-59	54.5	2	4
40-49	44.5	2	2

Notation:

- Σ Summation sign
- n Total number of sample observations
- X A variable or a measured characteristic
- X_i Any one value for a variable

MODE

The mode of the sample is the value of the variable having the greatest frequency.

Example: Obtain the mode for Data Set 1

77

For a grouped frequency distribution, the modal class is the class having the greatest frequency if the class intervals are equal and the mode is the midpoint of the modal class.

Example: Obtain the mode for Data Set 2

Modal interval is 70-79. For a more exact number, we say that the mode is the midpoint of the interval or 74.5

Properties of Mode

1. Appropriate for all types of data.
2. Not necessarily unique.
3. Note that the mode doesn't have to be near the center of the distribution.

MEDIAN

The median of a set of observations is the value of the variable such that half the values are less than the median and half are greater than the median. When observations are ordered from lowest to highest, the median is the number that divides the sample so that an equal number of cases fall above and below it.

For discrete observations, the median is found by first ordering the observations from smallest to largest, and then if the number of observations, n , is odd, taking the middle observation $(n+1)/2$ and if n is even, the median is the average of the observations at position $n/2$ and $(n/2)+1$ in the ordered arrangement.

Example: Obtain the median for Data Set 1

10th and 11th scores are 77 and 71. Average of these two scores is 74.

For a grouped frequency distribution, the median lies with the class (the median class) containing the value with a cumulative frequency of $(n+1)/2$. It can be determined from a cumulative frequency polygon or numerically by interpolation within the median class.

Example: Obtain the median for Data Set 2

We're looking for the interval that contains the middle observation (10.5th value). Looking at cumulative frequency numbers, that would be interval 70-79 (critical interval).

Whereabouts in the interval is the 10.5th value (median) or that score in the 50th percentile?

$$Md = l_i + i [(n/2 - f_m)/f_i]$$

Md = median

l_i = exact lower limit of interval i

i = width of interval i

$n/2$ = identifies interval in which median is located

f_m = number of observations in intervals below interval i

f_i = number of observations in interval i

$$\text{Median: } 69.5 + [(10.5 - 7)/6]*10 = 75.33$$

Properties of Median

1. Appropriate for interval level data and ordinal data
2. Median is not affected by extreme scores.
3. Median is insensitive to distances of measurements from the middle.
4. The median is problematic with binary data.

Quartiles and other Percentiles

Percentile: p th percentile is a number such that $p\%$ of the scores fall below it and $(100-p)\%$ fall above it.

Lower quartile: 25th percentile - median for observations that fall below median

Upper quartile: 75th percentile - median for observations that fall above median

MEAN

The sample mean of a sample of n observations x_1, x_2, \dots, x_n , denoted by \bar{x} , is the sum of the n observations divided by n if the data are not grouped.

$$\begin{aligned}\bar{x} &= \frac{\text{sum of all observations in the sample}}{\text{number of observations in the sample}} \\ &= (x_1 + x_2 + \dots + x_n) / n = (\Sigma x) / n\end{aligned}$$

where n is the number of observations, x_1 is the first sample observation, x_2 is the second sample observation, \dots , x_n is the n -th (last) sample observation.

The population mean, denoted by μ , is the average of all x values in the entire population.

Example: Obtain the sample mean for Data Set 1

Sum up all scores and divide by 20: $1466/20 = 73.3$

Weighted Mean

$$\bar{x} = (n_1 \bar{x}_1 + n_2 \bar{x}_2) / (n_1 + n_2)$$

Example:

Class 1: $n_1=20$; $\bar{x}_1= 73.3$

Class 2: $n_2=15$; $\bar{x}_2= 81$

$$[20(73.3) + 15(81)] / [20+15] = 76.6$$

Mean for Grouped Frequency Distribution

Example: Obtain the mean for Data Set 2

Use formula to calculate mean but substitute in midpoints of class intervals in place of the actual score values and weight by the frequency.

$$[3(94.5)+4(84.5)+6(74.5)+3(64.5)+2(54.5)+2(44.5)] / 20 = 73$$

Properties of Mean

1. Appropriate only for interval level data and above.
2. Mean is very sensitive to extreme scores, called outliers. An additional measurement at outer points would pull up or down the mean. So mean may not be representative of the measurements in the sample (particularly with small samples).
3. Mean is the center of gravity or point of balance for frequency distribution.
4. The sum of the deviations from the mean equal 0.

The deviation indicates the distance and direction of any raw score from the mean.

$$\text{Deviation} = X - \bar{x}$$

$$\text{Sum of the deviations} = \Sigma (X - \bar{x}) = 0$$

Example

1, 3, 4, 4, 9, 15

$$\bar{x} = 6$$

Deviations: $X - \bar{x}$

$$(1-6)=-5$$

$$(3-6)=-3$$

$$(4-6)=-2$$

$$(4-6)=-2$$

$$(9-6)=3$$

$$(15-6)=9$$

$$\Sigma (X - \bar{x}) = 0$$

Shape of Distribution

Mean, mode and median are identical for a unimodal, symmetric distribution, such as bell-shaped distribution.

The mean and median are identical if the histogram is symmetric. (e.g. Bimodal distribution)

If the histogram is unimodal with a long right hand tail (positively skewed) the mean lies above the median.

1, 2, 3, 4, 100
Mean = 22; Median = 3

If the histogram is unimodal with a long left hand tail (negatively skewed), then the mean is smaller than the median.

1, 97, 98, 99, 100
Mean = 79 Median = 98

Mean is influenced by extreme scores so you should be cautious about using the mean with highly skewed distribution. MEDIAN is not affected much, if at all, by changes in extreme scores.

With a bi-modal distribution, it's best to characterize the distribution by both modes. Using median or mean would obscure important features of the distribution.

The measures of central tendency do not give us any information about the spread (variability) of the observations. In the next class, we will look at describing variability in a data set.