

Social Studies 201**October 2, 2006****Median for grouped data** – text, section 5.3, pp. 171-183.**Introduction**

In the previous examples of obtaining the median, with a small number of cases, all values could easily be listed and ordered, and the middle value readily obtained. In determining the median for data grouped into a frequency or percentage distribution, it would take too much time and effort to list all cases, order them, and find the middle value. The notes in this section examine ways of determining the median for data that have been grouped into categories or intervals, in a frequency or percentage distribution.

For grouped data, the median is defined in the same manner as earlier, that is, the median is the middle value of the variable, with one-half of cases less than or equal to this value and the other one-half of cases greater than or equal to this value. While various methods can be used to obtain the median for grouped data, the method adopted here uses the cumulative percentage distribution. Cumulative percentages provide a running total of the percentage of cases in each category up to the largest value in the category. Proceeding from the smallest value of the variable toward larger values, the median is the value of the variable where the cumulative percentage of cases first exceeds fifty per cent.

The cumulative percentage distribution thus provides a relatively straightforward way of determining the median, as well as being useful for determining percentiles. If you are not familiar with cumulative distributions, study the following notes on cumulative percentage distributions prior to proceeding with the discussion of how to obtain the median for grouped data.

Cumulative percentage distributions

See text, pp. 175-176.

A cumulative percentage distribution provides a running total of the percentage of cases in a distribution. More exactly, a cumulative percentage distribution specifies the percentage of cases from the lowest value of a variable up to and including a specific value of the variable. A regular percentage distribution for a variable gives the percentage of cases at each value of X or within each interval (see text – section 4.6). Using these percentages, a

cumulative percentage is obtained by adding the percentage of cases in the categories or intervals up to and including those taking on the specific value of the variable X . This process is illustrated in the following examples.

Example – Student views about respect for government and politicians. Table 1 gives frequency and percentage distributions of responses of students (from SSAE98) to the question “In general, how much respect do you have for Government and for politicians?” Responses vary from 1, indicating no respect, to 5, indicating great respect.

Table 1: Frequency and percentage distributions of student views about respect for Government and for politicians

Response (X)	Degree or respect			
	For Government		For politicians	
	Number	%	Number	%
1 (No respect)	51	7.4	107	15.7
2	137	19.8	235	34.5
3	289	41.8	276	40.5
4	183	26.4	57	8.4
5 (Strongly agree)	32	4.6	7	1.0
Total	692	100.0	682	100.0

Question. From the distributions in Table 1, obtain the cumulative percentage distribution of responses for each of the two questions concerning respect.

Answer. For the variable respect for government, the cumulative percentages for the distribution of degrees of respect are contained in the last column of Table 2 and the distribution of responses concerning respect for politicians is in 5. An explanation of how to construct this column is as follows.

The second and third columns of Table 2 give the number and percentage of respondents with each degree of respect (X). The

Table 2: Frequency, percentage, and cumulative percentage distributions of degrees of respect for Government

Response (X)	Number	%	Cumulative %
1 (No respect)	51	7.4	7.4
2	137	19.8	27.2
3	289	41.8	68.9
4	183	26.4	95.4
5 (Great respect)	32	4.6	100.0
Total	692	100.0	100.0

final column of this table provides the cumulative percentage for each value of X . These are obtained as follows.

For $X = 1$, no respect, there are 51 respondents, or $(51/692) \times 100 = 7.4\%$ of all respondents. The cumulative percentage of respondents with a response of no respect is also 7.4%, since respondents cannot have a view less than 1.

For the second category, $X = 2$, there are 137 respondents, or $(137/692) \times 100 = 19.8\%$ of respondents. The cumulative percentage for $X = 2$ is the 7.4% with a response of 1 plus the 19.8% with a response of 2, or $7.4 + 19.8 = 27.2\%$.

For $X = 3$, the cumulative percentage is the percentage of respondents with responses of 3 or less. This is all those responding 1, 2, or 3, that is, $7.4 + 19.8 + 41.8 = 68.9\%$ of respondents (using the rounding in the SPSS table).

Rather than adding all the percentages from the lowest value of X , another way to obtain cumulative percentages is to produce a running total of percentages. This involves adding the percentage of respondents in the next higher category of X to the cumulative percentage for the previous value of X . For example, the cumulative per cent at $X = 4$ is all those with a response of $X = 3$ or less plus all those with a response of 4. The cumulative per-

centage at $X = 3$ was 68.9%. Adding all those with a response of 4, another 26.4%, gives a total of $68.9\% + 26.4\% = 95.4\%$, again using the rounding from the SPSS table. This is the cumulative percentage of respondents for $X = 4$.

Finally, the cumulative percentage for $X = 5$ includes all respondents, or 100%. That is, all respondents have a response of 5 or less, since 5 is the maximum response possible.

Example – Grade distribution. A grade distribution for undergraduate students (from SSAE98) is given in the first three columns of Table 3.

Table 3: Frequency, percentage, and cumulative per cent distribution of grade point averages for 573 undergraduate students

Grade	Number	Per cent
Less than 60	14	2.4
60-64	40	7.0
65-69	93	16.2
70-74	146	25.5
75-79	138	24.1
80-84	119	20.8
85 plus	23	4.0
Total	573	100.0

Question. Obtain the cumulative percentage distribution of grades.

Answer. The cumulative percentage distribution for Table 3 is in Table 4.

The right column Table 4 provides the cumulative percentages. This is obtained as follows.

Table 4: Frequency, percentage, and cumulative per cent distribution of grade point averages for 573 undergraduate students

Grade	Number	%	Cumulative %
Less than 60	14	2.4	2.4
60-64	40	7.0	9.4
65-69	93	16.2	25.6
70-74	146	25.5	51.1
75-79	138	24.1	75.2
80-84	119	20.8	96.0
85 plus	23	4.0	100.0
Total	573	100.0	

Since the first interval contains all the lowest values of grades, less than 60, for this first category, the cumulative percentage (2.4%) is the same as the percentage of with those grades less than 60 (2.4%). For the 60-64 category, there are another 7.0% of cases, giving a cumulative percentage of $2.4 + 7.0 = 9.4\%$ of cases up to and including the cases in the 60-64 interval. For the next interval, there are another 16.2% of students with grades between 65 and 69. When all the grades in the 65-69 interval are considered, there are 25.6% of all students ($2.4 + 7.0 + 16.2 = 25.6\%$). The cumulative percentages for the other categories can be obtained in the same way.

The cumulative percentage column gives the percentage of students with grades less than or equal to the grade at the upper end of each interval. For example, for a grade of 70-74, the cumulative percentage is 51.1% of students. This means there are 51.1% of students with grade point averages of less than 75. Similarly, there are 75.2% of students with grades less than 80. Finally, once all those with grades of up to 85 plus are included, all 100% of students are included.

Cumulative percentage distributions – Summary. The cumulative percentage of respondents provides a running total of the percentage of respondents with values of the variable up to and including each specific value. The following notes demonstrate how to use the cumulative percentage distribution to obtain the median.

Obtaining the median using cumulative percentages

Using cumulative percentages, the median value of the variable is in the category or interval where a cumulative percentage of 50 per cent is first reached. Proceeding in order from the lowest to the highest value of the variable, the middle value, or fifty per cent point is not encountered until the cumulative percentage is fifty per cent or more. It is the category or interval where the variable first crosses the fifty per cent point that contains the median or middle value of the variable.

Example – Respect toward Government and politicians.

For Government, in (Tables 1 and 2), the median attitude is 3, since this is where a cumulative total of fifty per cent is first reached. That is, the response of $X = 1$ accounts for only 7.4% of cases, and by the time all those with responses of $X = 1$ and $X = 2$ are encountered, there are only 27.2% of cases. But when all the responses of $X = 3$ are considered, that is, all those responding 1, 2, or 3, this accounts for 68.9% of all cases. This means that one-half, or fifty per cent, of all respondents have responses of 1, 2, or 3, and the fiftieth per cent is at $X = 3$. As a result, the median degree of respect for Government is 3, or a neutral response.

Table 5 contains the frequency, percentage, and cumulative percentage distributions for the responses concerning respect for politicians. From this table, the median response to this question is 2, an attitude of little respect for politicians. At $X = 1$, there are only 15.7% of responses but then there are 34.5% of responses at $X = 2$, bringing the cumulative percentage to just over fifty per cent. As a result, one-half of respondents have a respect level of 2 or less and the other one-half have a respect level of 2 or more.

Table 5: Frequency, percentage, and cumulative percentage distributions of degrees of respect for politicians

Response (X)	Number	%	Cumulative %
1 (No respect)	107	15.7	15.7
2	235	34.5	50.1
3	276	40.5	90.6
4	57	8.4	99.0
5 (Great respect)	7	1.0	100.0
Total	682	100.0	100.0

Note that the distribution of responses for the two questions indicates that while the respect level for governments is not high (median of 3 or neutral), the respect level for politicians is even lower (median of only 2).

Example – Grade point averages. The grade point averages of five hundred and seventy three students is given in Table 4. From the table, it can be seen that the median grade is in the interval 70-74. At any grade below 70, there are not yet fifty per cent, or one-half, of all cases taken into account. Including all those with grades up to 70, that is, including the 65-69 category, there are only 25.6% of all students (cumulative percentage column). Proceeding to the next interval and including students with grades of 70-74, an additional 25.5% of students, the cumulative percentage reaches 51.1%. As a result, the fifty per cent point is crossed in the interval from 70 and 74. The median grade is in the grade range from 70-74.

In order to determine a specific value for the median grade, it is necessary to interpolate within the 70-74 interval. This is explained in the next section.

Summary

To determine a median category in the case of data grouped into tables, first make sure that the variable has at least an ordinal level of measurement. Then obtain the percentage distribution and cumulative percentage distributions. The median is the value of the variable for the category or interval where the fifty per cent point of the cumulative percentage distribution is first reached.

Note that the median is not the cumulative percentage itself, but the value of the variable at which the cumulative percentage first reaches fifty per cent.

Where the data are discrete, as in the example of attitudes above, the median is simply the value of the variable where the cumulative percentage distribution first reaches fifty per cent or more. For a continuous variable, the same method is used to determine the interval containing the median. In order to determine a more exact, or specific, value of the median, the method of interpolation is used. This is described in the following notes.

Interpolation to obtain median.

See text, section 5.3.3, pp. 178-183.

For data where the variable is numeric and grouped into intervals, the median is obtained by interpolating across the interval containing the median, or fifty per cent point. As noted above, the cumulative percentage is used to identify the interval containing the median. Then the following formula is used to determine a more exact value of the variable for the median (text, p. 182).

$$\text{Median} = \begin{array}{c} \text{Value of variable} \\ \text{at lower end of} \\ \text{interval} \end{array} + \left[\frac{50 - \begin{array}{c} \text{Cumulative per cent} \\ \text{at lower end of interval} \end{array}}{\begin{array}{c} \text{Per cent of cases in interval} \end{array}} \right] \begin{array}{c} \text{[Interval]} \\ \text{[width]} \end{array}$$

That is, the median is the value of the variable at the lower end of the interval containing the median, plus the distance along this interval required to meet the fifty per cent point.

The numerator of the large expression in the middle of the formula represents the percentage of cases between the lower end of the interval and the

median, or fifty per cent point. Dividing this by the percentage of cases in the interval gives the proportion of the distance along the interval required to reach the fifty per cent point. Multiplying this proportion by the interval width, in units of the variable, indicates the position along the interval where the median is located. That is, the product of the two expressions in brackets is the distance from the lower end of the interval to the median, or fifty per cent point. Adding this distance to the value of the variable at the lower end of the interval gives the value of the median.

Example – Median grade point average. Obtain the median grade point average for the five hundred and seventy-three students in Table 3 using the method of interpolation.

Answer. As demonstrated earlier, from the cumulative percentage distribution in Table 4, the median grade is in the interval 70-74. The real class limits should be used in this case, in order to ensure the proper interval width of 5 grade points, and to ensure that none of the grade points are left out. The real class limits for the 70-74 interval are 69.5 and 74.5.

First using the formula,

$$\text{Median} = \text{Value of variable at lower end of interval} + \left[\frac{50 - \text{Cumulative per cent at lower end of interval}}{\text{Per cent of cases in interval}} \right] \left[\text{Interval width} \right]$$

the lower end of the interval is 69.5 and the interval width is five units of grades (69.5 to 74.5). The cumulative percentage of cases at the lower end of the interval is 25.6% and the percentage of cases in the 70-75 interval is 25.5%. Entering these numbers in the formula gives:

$$\begin{aligned} \text{Median} &= 69.5 + \left(\frac{50 - 25.6}{25.5} \times 5 \right) \\ &= 69.5 + \left(\frac{24.4}{25.5} \times 5 \right) \\ &= 69.5 + (0.957 \times 5) \\ &= 69.5 + 4.78 \\ &= 74.28. \end{aligned}$$

Rounding to the nearest tenth of a grade point, the median grade point average for this group of students is 74.3.

The diagrammatic presentation is provided in Figure 1. Beginning at the lower end of the interval, there are 25.6% of cases with values less than or equal to a grade of 69.5. In this example, the median is almost at the upper end of the interval, since the cumulative percentage of cases at the upper end is 51.1%. The fifty per cent point lies just to the left of this. In percentages, the distance from the lower end of the interval to the fifty per cent point is $50 - 25.6 = 24.4\%$ of cases. There are 25.5% of students with grades in the interval from 69.5 to 74.5. As a fraction of the distance across the interval required to reach the fifty per cent point, this represents the proportion $24.4/25.5 = 0.957$ of the distance along the interval.

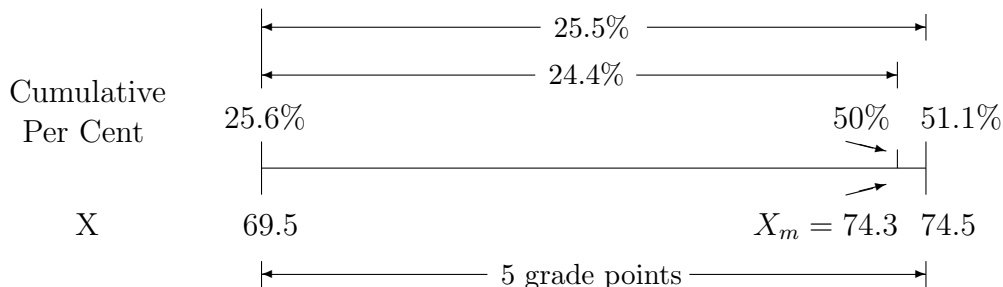
The next step is to convert this proportion into units of the variable X . The distance across the interval is from 69.5 to 74.5, or 5 units of grade point average. The distance from the lower end of the interval to the median is the proportion 0.957 of these ten units, or $0.957 \times 5 = 4.78$ grade points.

The median is thus 4.78 units above the lower end of the interval. Adding this to the value of the variable at the lower end of the interval, $X = 70$, gives the median value of $69.5 + 4.78 = 74.28$. Using this method gives the same results as the formula, that is, a median grade point average of 74.3.

Recap and summary of the median

- **Ordinal level of measurement.** To calculate the median, it is necessary to order the values of the variable. To do this, the variable must have an ordinal, interval, or ratio level of measurement. That is, variables that have a nominal scale of measurement, but no more than nominal, cannot be ordered. For example, the median cannot be determined for a variable such as political party preference, sex, religious preference, or ethnicity. These represent different characteristics of individuals, but such characteristics cannot be ordered as less than

Figure 1: Diagram for calculating median grade point average



or greater than other characteristics – they are just different characteristics. Values of variables such as attitudes or opinions (ordinal), and income or age (interval and ratio), can be ordered as less than, greater than, or equal to other values, so the median can be obtained for these variables.

- **Methods.** In the case of ungrouped data, list the values in order (low to high or high to low) and determine the middle value. In the case of data grouped into discrete categories, the median is at the value of the variable where the fifty per cent point of the cumulative percentage distribution occurs. When the variable is continuous so that the data are grouped into intervals, the median is in the category where the cumulative percentage distribution reaches the fifty per cent point. Interpolation using the formula or diagram can then be used to obtain a more precise estimate of the median.
- **Middle or typical value.** The median denotes the middle value of a set of values or of a distribution. As a result, the median is sometimes considered a **typical** value, in the sense that it is near the centre or middle of a distribution. Consider the following from Statistics Canada.

Statistics Canada reports that the median total income of persons 15 years of age and over, for the year 2000, was \$24,064 in Regina and \$19,636 for Saskatchewan.

Source: <http://www12.statcan.ca/english/profil01/PlaceSearchForm1.cfm>

This indicates that the typical income for Canada as a whole is greater than the typical income in Regina by about \$4,500. In Regina, one-half of individuals have incomes equal to less than \$19,636 and the other one-half have incomes greater than or equal to this.

From the 2001 Census of Agriculture, Statistics Canada reports that in the year 2001,

for the farm operator population, the median age is 49 years
... for the entire labour force, it is 38. **Source:** Statistics Canada,
The Daily, November 20, 2002.

<http://www.statcan.ca/Daily/English/021120/d021120a.htm>

From this result, it can be concluded that a typical farm operator is older than a typical member of the labour force. One-half of farm operators are age 49 or less and the other half are age 49 or more. In contrast, for the entire labour force, the age that divides the members of the labour force in two is 38 years.

- **Not sensitive to extremes.** The median is not sensitive to changes in the extremes of a distribution. This can be observed from student ages (Example of September 27, 2006). Whether the last student is aged 71 or 51 years of age would make no difference in determining the median – a change in the extreme values would not alter where the centre of the distribution is located. This property of the median is an advantage if a data analyst wishes to portray the position of the centre of a distribution. It can be a disadvantage if the analyst is concerned with the extremes of a distribution.

Last edited October 2, 2006.