Social Studies 201

September 29, 2004

Median for grouped data – text, section 5.3, pp. 171-183.

Note: The examples in these notes may be different than used in class on September 29. However, the examples are similar and the methods used are identical to what was presented in class.

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 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 determine 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 is that of the cumulative percentage distribution. Cumulative percentages represent a running total of the percentage of cases. Once the half way point, or the cumulative percentage of fifty per cent, is reached, that is the location of the median.

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 comulative 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 the variable X through 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 3.9 – Male student views about issue of "Help themselves". Table 1 gives a frequency and percentage distribution of responses of a group of male students (from SSAE98) to the question "The more money spent helping pepole, the less they will help themselves." As in earlier examples using this variable, responses vary from 1, indicating strongly disagree, to 5, indicating strongly agree.

Table 1: Frequency and percentage distributions of male views on "Help themselves,"

Response (X)	Number	%
1 (Strongly disagree)	19	7.3
2	45	17.2
3	76	29.2
4	75	28.7
5 (Strongly agree)	46	17.6
Total	261	100.0

Question. From the distribution in Table 1, obtain the cumulative percentage distribution of responses.

Answer. The cumulative percentates for the distribution of attitudes concerning the variable "Help themselves" is contained in the last column of Table 2. An explanation of how to construct this column is contained following the table.

The second and third columns of Table 2 give the number and percentage of males with each view (X). The final column of this table provides the cumulative percentage for each value of X. These are obtained as follows.

For $X = 1$, an opinion of strongly disagree, there are 19 respondents, or $19/261 \times 100 = 7.3\%$ of all respondents. The cumulative percentage of respondents with with a response of strongly disagree is also 7.3%, since respondents cannot have a view less than 1.

Table 2: Frequency, percentage, and cumulative percentage distributions of male views on "Help themselves,"

Response (X)	Number	%	Cumulative %
1 (Strongly disagree)	19	7.3	7.3
2	45	17.2	24.5
3	76	29.2	53.7
4	75	28.7	82.4
5 (Strongly agree)	46	17.6	100.0
Total	261	100.0	100.0

For the second category, $X = 2$, there are 45 respondents, or $45/261 \times 100 = 17.2\%$ of respondents. The cumulative percentage for $X = 2$ is the 7.3% with a response of 1 plus the 17.2% with a response of 2. The cumulative percentage of respondents with a response of up to 2, is $7.3 + 17.2 = 24.5\%$.

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, $7.3 + 17.2 + 29.2 = 53.7\%$ of respondents.

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 or 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 percentage at $X = 3$ was 53.7%. Adding all those with a response of 4, another 28.7%, gives a total of $53.7\% + 28.7\% = 82.4\%$. 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 3.10 – 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
60-65	40	7.0
65-70	93	16.2
70-75	146	25.5
75-80	138	24.1
80-85	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 **??**.

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

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 60-65 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 60-65 interval. For the next interval, there are another 16.2% of students with grades between 65 and 70. When all the grades in the 65-70 interval are considered, there are 25.6% of all students (2.4+ 7.0+16.2=25.6%). The cu-

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
60-65	40	7.0	9.4
65-70	93	16.2	25.6
70-75	146	25.5	51.1
75-80	138	24.1	75.2
80-85	119	20.8	96.0
85 plus	23	4.0	100.0
Total	573	100.0	

mulative 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 70-75, the cumulative percentage is 51.1% of students. This means there are 51.1% of students with grade point averages of 75 or less. Similarly, there are 75.2% of students with grades up to 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 3.11 – Male student views about "Help themselves". In Example 3.9, containing the distributions of male student attitudes about the variable "Help themselves" (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 7.3% of cases, and by the time all those with responses of $X = 1$ and $X = 2$ are encountered, there are only 24.5% of cases. But when all the responses of $X = 3$ are considered, that is, all those responding 1, 2, or 3, this accounts for 53.7% 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 response of the two hundred and sixty-one male students is 3, or a neutral response.

Example 3.12 – Grade point averages. In Example 3.10, grade point averages of five hundred and seventy three students (Table 4), the median grade is in the interval 70-75. 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 65-70 category, there are only 25.6% of all students (cumulative percentage column). Proceeding to the next interval and including students with grades of 70-75, 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 75. The median grade is in the grade range from 70-75.

In order to determine a specific value for the median grade, it

is necessary to interpolate within the 70-75 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 cumulatie 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).

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

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 varaible, 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 3.14 – Median grade point average**. Obtain the median grade point average for the five hundred and seventy-three students in Table 3 using the two methods of the formula and diagram.
>
> **Answer**. As demonstrated earlier, from the cumulative percentage distribution in Table 4, the median grade is in the interval 70-75.
>
> First using the formula,

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

the lower end of the interval is 70 and the interval width is five units of grades (70 to 75). 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:

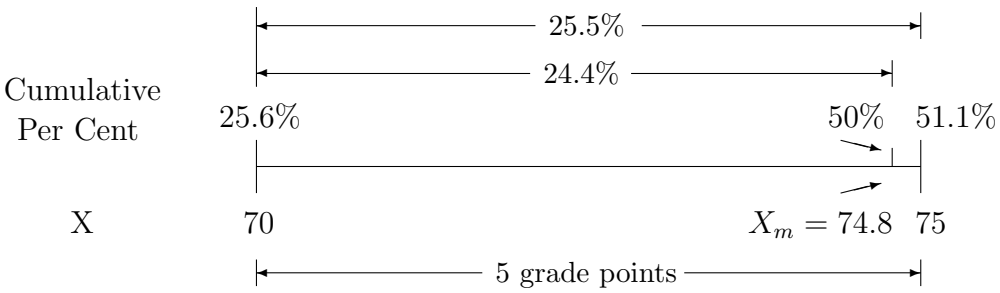
$$\begin{aligned}\text{Median} &= 70 + \left(\frac{50 - 25.6}{25.5} \times 5\right)\\ &= 70 + \left(\frac{24.4}{25.5} \times 5\right)\\ &= 70 + (0.957 \times 5)\\ &= 70 + 4.78\\ &= 74.78.\end{aligned}$$

Rounding to the nearest tenth of a grade point, the median grade point average for this group of students is 74.8.

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 70. 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 70 to 75. 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 70 to 75, or 5

Figure 1: Diagram for calculating median grade point average



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 $70 + 4.78 = 74.78$. Using this method gives the same results as the formula, that is, a median grade point average of 74.8.

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 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. Where data are grouped into continuous intervals, the median is in the category where the cumulative percentage distribution reaches the fifty per cent point. Interpolation using the formula or diagram presented in these notes can be used to obtain a more precise value 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 in Example 3.8. 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.