Social Studies 201

September 27, 2004

Mean for grouped data – text, section 5.4.2, pp. 188-200.

For grouped data, where data are grouped into categories or intervals and presented as diagrams or tables, the definition of the mean is unchanged, but the method of obtaining it differs from that used for ungrouped data. The mean is the sum of values of the variable divided by the number of cases, but it is necessary to make sure that the sum is properly obtained. In order to obtain the sum or total value, each individual value must be multiplied by the number, or percentage, of cases that take on that value, and these products are added together. The mean is then obtained by dividing this sum by the number, or percentage, of cases.

A formal definition and examples follow – again, if you are unfamiliar with the notation below, please read the text, pp. 97-100 and pp. 190-92 on notation.

> **Definition of the mean for grouped data**. For a variable X, taking on k different values $X_1, X_2, X_3, ...X_k$, with respective frequencies $f_1, f_2, f_3, ...f_k$, the mean of X is
>
> $$\bar{X} = \frac{f_1X_1 + f_2X_2 + f_3X_3 + ... + f_kX_k}{n}$$
>
> where
>
> $$n = f_1 + f_2 + f_3 + ... + f_k.$$
>
> Using summation notation,
>
> $$\bar{X} = \frac{\Sigma(fX)}{n}$$
>
> where $n = \Sigma f$.

Notes on calculating the mean with grouped data

1. **Products**. The first step in calculating the mean is to multiply each value of the variable X by the frequency of occurrence of that value. This produces the k products f_iX_i or, without the i indexes, simply fX.

2. **Sum**. All the k products, $f_i X_i$, are added together, to produce the sum of values of the variable across all cases in the data set. This sum is $\Sigma(fX)$ or, without the i indexes, $\Sigma(fX)$. This forms the numerator for calculating the mean (see note 4).

   Note that $\Sigma(fX)$ is the sum of the individual products fX. It is **not** the sum of the fs, multiplied by the sum of the Xs.

3. **Number of cases**. The denominator used to obtain the mean is the number of cases. This is obtained by adding the frequencies of occurrence for each value of the variable, the fs. That is, the total number of cases in the data set is $n = \Sigma f_i$, or simply $n = \Sigma f$.

4. **Mean**. In order to determine the mean, divide the sum (ΣfX) in item 2 by the number of cases (n) in item 3.

5. **Weighted mean**. A mean of the form presented here is sometimes referred to as a weighted mean. That is, each value of the variable X is weighted by the frequency of occurrence of that variable, the f values.

Steps in calculation of $\bar{X}$. Data are often presented in the form of tables of the frequency distribution of the variable. Table 1 provides a generic format for such a table. In this table, the variable X has k categories and $\Sigma f = n$ cases.

Table 1: Format of table for calculaing mean of grouped data

X	f	fX
X_1	f_1	$f_1 X_1$
X_2	f_2	$f_2 X_2$
.	.	.
.	.	.
.	.	.
X_k	f_k	$f_k X_k$
Total	$\Sigma f = n$	ΣfX

When data are presented in tabular form, the mean is obtained as follows:

- First create a table with the values of X in the first column and the frequencies of occurrence f in a second column.
- Create a third column for the products of the fs and the Xs, fX. Multiply each f by its corresponding X value and enter these products in the third column of the table.
- Sum the products fX in the third column to obtain the column total ΣfX. Again, note this is the sum of the individual products in this column, **not** the sum of the fs multiplied by the sum of the Xs.
- Divide the sum in step 3 by n, the sum of the frequencies in second column. This produces the mean of the variable X, $\bar{X} = \Sigma(fX)/n$.

The examples that follow demonstrate how to obtain the mean for grouped data, first with variables that are discrete and then with variables whose values are grouped into intervals. In the latter case, the midpoints of the intervals are used as the appropriate X values for obtaining the mean. In addition, there is an example of how to obtain the mean for a percentage distribution.

Example – Credit hours of students. A sample of fifteen University of Regina undergraduate students gives the frequency distribution of Table 2. Obtain the mean credit hours for this sample of students.

Table 2: Distribution of credit hours for fifteen University of Regina undergraduate students

Number of credit hours	Number of students
3	1
9	3
12	4
14	1
15	4
17	2
Total	15

Answer. The data of Table 2 are reorganized into a tabular format using the X and f notation in Table 3. The variable is "credit hours" and is labelled X – values for X are in the first column of Table 3. The number of students with each number of credit hours is the frequency of occurrence. These frequencies are labelled f and placed in the second column of Table 3. The entries in the third column of Table 3 are obtained by multiplying each f by its corresponding X value to obtain the values fX, the product of the entries in the first two columns.

Proceeding through Table 3, for the first row, there is 1 student taking 3 credit hours, for a product of $1 \times 3 = 3$. That is, $X_1 = 3$ and $f_1 = 1$, so $f_1X_1 = 1 \times 3 = 3$. For the second row, those with $X = 9$ credit hours, there are $f_2 = 3$ students, so $fX = 3 \times 9 = 27$. The remaining rows are calculated in a similar manner. The sum of the f column is the number of cases, that is, $n = \Sigma f = 15$.

The sum of the third column is ΣfX, that is, the sum of the products of the frequency times the value of the variable. In this

Table 3: Table for calculation of mean credit hours for fifteen University of Regina undergraduate students

X	f	fX
3	1	3
9	3	27
12	4	48
14	1	14
15	4	60
17	2	34
Total	15	186

table, this sum is $\Sigma(fX) = 186$. This is the total credit hours taken by all the students in this sample.

The mean number of credit hours is obtained by dividing the total credit hours by 15, the number of students. That is, the mean is $186/15 = 12.4$ credit hours. In symbols, the total credit hours for all student in the sample is $\Sigma(fX) = 186$ and the number of cases is where $n = \Sigma f = 15$. The mean is thus

$$\bar{X} = \frac{\Sigma(fX)}{n} = \frac{186}{15} = 12.4.$$

The mean credit hours taken by this sample of fifteen students is 12.4 hours.

Midpoint of the interval

Where values of a variable X are presented in interval format, the specific values of X used in the formula

$$\bar{X} = \frac{\Sigma(fX)}{n}$$

are the midpoints of the intervals. The midpoint of the interval is the sum of the two endpoints of the interval, divided by two. It does note matter whether you use apparent or real class limits to do this. The midpoint of the interval will be same, regardless of which class limits used. See text, section 4.7.3, beginning on p. 134.

> **Example – Anticipated earnings of undergraduates**. A frequency distribution of anticipated annual earnings for a sample of ninety-two University or Regina undergraduate students is given in Table 4. These are earnings the students expect to receive after graduation.

Table 4: Frequency distribution of anticipated annual earnings (in thousands of dollars) of a sample of 92 undergraduate students

Anticipated earnings	Frequency
0-20	30
20-30	27
30-40	14
40-60	19
60-80	2
Total	92

> The data are adapted from the Canadian Undergraduate Survey Consortium, *Graduating Students Survey: 2003*, p. 93. The whole report is available in the Office of Resource Planning section of the University of Regina web site

(http://www.uregina.ca/presoff/orp).

From these data, obtain the mean anticipated earnings of this sample of students. Briefly comment on the likely accuracy of this mean.

Answer. The variable here is anticipated earnings in thousands of dollars, a variable measured at the interval and ratio scale of measurement.

The data of Table 4 are reorganized in Table 5 using the tabular format and notation introduced in this section. A new column, X is introduced, to denote the midpoint of each interval in thousands of dollars. The frequency column is give the symbol f and a final column, with the products of fX is added to the table.

Table 5: Calculations for mean anticipated annual earnings of sample of 92 undergraduate students

Anticipated earnings	X	f	fX
0-20	10	30	300
20-30	25	27	675
30-40	35	14	490
40-60	50	19	950
60-80	70	2	140
Total		92	2,755

The first interval is from 0 to 20 thousand dollars, so the midpoint is $(0 + 20)/2 = 10$. Multiplying the frequency $f = 30$ by this midpoint $X = 10$ gives a product $30 \times 10 = 300$ and this product is entered in the last column. This product represents the total anticipated annual earnings for the thirty respondents in the first row of the table.

The second row of the table has 27 respondents who anticipate earnings between 20 and 30 thousand dollars. The midpoint of

this second interval is $(20 + 30)/2 = 50/2 = 25$ and the fX product is thus $27 \times 25 = 675$.

The midpoint of the third interval, 30-40, is $X = 35$, and when this is multiplied by its corresponding frequency of $f = 14$, the product is is $14 \times 35 = 490$, again entered into the last column.

Each of the succeeding rows is produced in the same manner – for the fourth row, $19 \times 50 = 950$ and finally $2 \times 70 = 140$.

The sum of the last column, the product column, is $\Sigma fX = 2,755$. The sum of the f column is the sample size for this sample of students, that is, $n = \Sigma f = 92$.

Using the sums in the last row of the table gives a mean of

$$\bar{X} = \frac{\Sigma(fX)}{n} = \frac{2,755}{92} = 29.946$$

The mean anticipated annual earnings are 29.9 thousand dollars, or \$29,900, rounded to the nearest hundred dollars.

There are several reasons why this mean may not be all that precise. First, this is a sample of only ninety-two students who were asked to state their anticipated earnings. While their anticipations may be more or less correct, these are anticipations only, and may not be accurate judgments of actual future earnings. Second, while the calculations for the mean reported here are correct, the data come with considerable uncertainty. That is, the reader of the report is not given the anticipated earnings of each student surveyed, but the data are grouped into a table. There are thirty students in the first interval, from 0 to 20, and the midpoint of this interval, 10 thousand dollars, is used in the formula. It may be that if more detailed information were available about these thirty respondents, the average for these thirty would not be exactly 10 thousand dollars. But the midpoint of 10 is the best that can be done in terms of estimating an appropriate average for this first interval.

As a result of these considerations, it is best to round the mean to the nearest thousand dollars, and report that mean anticipated earnings of this sample of undergraduates is approximately \$30,000.

Recap and summary of the mean

- **Mean**. Regardless of how data are presented, the mean is the sum of the values of the variable across all cases, divided by the number of cases.

- **Average**. While the term "average" can be used for any of the mode, median, or mean, many people have the mean in mind when referring to an average. Someone talking about average age or average temperatute likely has in mind the mean age or temperature. The mean as a measure of average is embedded in popular usage.

- **Arithmetic mean**. There are several different means used in scientific work – geometric mean, harmonic mean, etc. The mean used here is referred to as the arithmetic mean – the sum of all values divided by the number of cases. When working with grouped data, this mean is sometimes referred to as the weighted mean or, more properly, the weighted arithmetic mean.

- **Ungrouped and group methods**. For ungrouped data, the mean is simply the sum of all values divided by the number of cases. For grouped data, the sum of all values is obtained by multiplying the frequency or percentage of occurrence by the value of the variable. In the case of data grouped into intervals, the value of the variable is the midpoint of the interval or, in the case of open-ended interval, an estimate of the midpoint of the interval.

- **Centre**? A mean may not represent the centre of a distribution. If the values of a variable are 4, 6, 8, and 32, the mean is $(4 + 6 + 8 + 32)/4 = 50/4 = 12.5$. This is the correct mean for these four values but some would argue that this mean is an artificial construction, not really representative of the sample. While that argument may be correct, the issue is not so much whether the mean is artificial. It is more a matter of properly interpreting what a mean indicates. Work through some of the exercises following this section and this may help you understand how to interpret the mean and other measure of centrality.