Social Studies 201

January 24, 2005

Mean as a measures of centrality

Text, section 5.4, pp. 183-207.

Introduction

The mean is the most commonly used measure of central tendency or centrality, and is often referred to as simply the average. Two examples, referring to students at the University of Regina, are as follows.

> For full-time first year University of Regina students, entering in the Winter 2003 semester, the admission average was 76.7.
>
> Between Fall 1998 and Fall 2002, the average age of students at the University of Regina declined from 26.5 years to 26.1 years.
>
> Data from University of Regina, *Fact Book 1999-2003*, p. 252 and p. 74. (Available at http://www.uregina.ca/presoff/orp/FACT-BOOK/1999-2003.pdf).

In each of these examples, the average used is the mean, not the median or mode. In the first quote, the average admission grade of entering students is a mean of 76.7%; the second quote reports the mean age of students at the University of Regina. A quick definition of the mean is as follows, although different methods are used to obtain the mean, depending how data are presented and organized.

> **Definition**. The mean is the sum of all the values of a variable divided by the number of cases.

For example, in the first quote above, the mean grade of entering students is the sum of the grade point averages of all full-time first year students, divided by the number of such students.

Before giving a formal definition of the mean and discussing methods of calculating and interpreting the mean, take note of the following general statements about averages and means.

Average. In ordinary usage, the term "average" could refer to any of the three common measures of centrality, the mode, median, or mean. When someone refers to an average, this usually implies the "mean." However, when encountering new data, where an average is used, it is important to consider which of the three measures is being used.

Ungrouped and grouped data. As was the case for the mode and median, the method of calculating the mean differs, depending whether the data are ungrouped (a simple list of values) or grouped (a frequency or percentage distribution table or histogram). See pp. 94-5 of the text and the examples following if you need to review this distinction.

Interval or ratio level of measurement. Since it is necessary to sum all the values of a variable to obtain the mean, the variable must be numerical. It would not be possible to obtain the mean for a list of values such as strongly agree, agree, or disagree without attaching numerical values to these categories of response.

In addition, the variable should be measured at the interval or ratio level to obtain the mean of a variable. There should be a well-defined unit of measure for the variable, with numerical differences between values of a variable being meaningful. For example, height in centimetres has the centimetre as the unit of measure, and a difference of ten centimetres is meaningful in terms of this unit.

Variables such as political party supported, sex, ethnicity, and other variables having no more than a nominal scale of measurement, do not have a unit of measure. In addition, differences between categories of variables with no more than a nominal scale are are not well-defined, so the mean cannot be meaningfully calculated, even if numbers (usually codes) are attached to the categories.

For variables such as attitude and opinions, where responses are ranked on an ordinal scale, the mean is sometimes obtained.

While there is not a well-defined unit of measure for these variables, researchers commonly treat such ordinal scales as having an interval level of measurement. It is common to report the mean opinion about a social or political issue. For example, the mean opinion for undergraduate students on the issue of increasing corporate taxes is 3.8, where responses are measured on a five-point scale from 1, indicating strongly disagree, to 5, denoting strongly agree. A mean of 3 implies a neutral response (midway between 1 and 5) so a mean of 3.8 indicates moderate agreement with increasing corporate taxes. A mean of 4.5 would indicate stronger agreement and a mean of, say, 2.3 would indicated disagreement.

The notes in this section discuss methods for obtaining the mean for data presented in different formats. The mean is the most widely used measure of centrality, and the measure used most extensively throughout the rest of the course. As a result, you should make sure you understand how to obtain and interpret the mean.

Mean for ungrouped data

For ungrouped data, that is, a list of values of a variable, the mean is the sum of the values divided by the number of cases. For this section and later parts of Social Studies 201, it is useful to have this stated in symbolic terms. A formal definition of the mean for a variable X is as follows.

Note: If you have difficulty with the notation below, before proceeding please read pages 186-8 of the text, where summation notation is discussed.

Definition of the mean. For a variable X, the mean is the sum of the values of X divided by the number of cases.

Using algebraic notation, if there are n values for the variable X, that is, $X_1, X_2, X_3, ...X_n$, the mean of X is

$$\bar{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n}.$$

Alternatively, this formula can be written more compactly using the summation notation.

$$\bar{X} = \frac{\Sigma_{i=1}^{n} X_i}{n}$$

or, if the index i is dropped, as simply

$$\bar{X} = \frac{\Sigma X}{n}.$$

Notes on calculating the mean

1. **Bar** X. In statistical work, whenever a horizontal line appears above a variable, this indicates the mean of the variable. In the definitions above, the variable is X, so that $\bar{X}$ is the mean of the variable X. If a variable was labelled y, the mean of this variable would be indicated by $\bar{y}$.

2. **Sum**. The first step in calculating the mean is to add together all the values of the variable. This is the sum $X_1 + X_2 + X_3 + ... + X_n$ or ΣX. This forms the numerator for calculating the mean.

3. **Number of cases**. The denominator for the formula determining the mean is the number of cases summed, that is, n. This involves merely counting the number of cases in the data set.

4. **Mean**. To determine the mean, divide the sum (ΣX) in item 2 by the number of cases (n) in item 3. That is, the mean is $\bar{X} = \Sigma X/n$.

5. **Units**. The mean is expressed in units of the variable X. For example, if X represents income in dollars, mean income is also in dollars.

Example. A student takes five classes during her first semester at the University of Regina. The grades obtained are 64, 74, 68, 79, and 85. What is the mean grade of the student?

Answer. The variable is grade, presumably in per cent, so this variable has at least an interval level of measurement.

The sum of the five grades is $64 + 74 + 68 + 79 + 85 = 370$. The mean is this sum divided by 5, the number of grades. The mean is thus 370 divided by 5 or 74. That is $\bar{X} = 74$.

Illustrating this process using symbolic notation, the variable X is the grade, there are $n = 5$ grades obtained by the student, and the individual grades are labelled as in Table 1. That is, the first grade is $X_1 = 64$, the second grade is $X_2 = 74$, and so on. The sum of these five values is $\Sigma X = 64 + 74 + 68 + 79 + 85 = 370$. The mean is thus

$$\bar{X} = \frac{\Sigma X}{n} = \frac{370}{5} = 74.$$

Table 1: Labels for calculating mean grade when using symbolic notation

Label	Grade (X)
X_1	64
X_2	74
X_3	68
X_4	79
X_5	85
Sum or ΣX	370

Example. For a sample of eleven University of Regina undergraduates, the ages are 27, 19, 18, 18, 18, 29, 20, 39, 24, 18, and 19 years.

1. Obtain the mean age of these eleven students.
2. If a senior, aged 71 years, decides to return to university and joins the original eleven students, what is the mean age of the twelve students?

Answer. The variable is age in years, a variable that has an interval and ratio level of measurement.

1. The number of cases is $n = 11$. If the age of students is given the symbol X, the sum of the eleven values is

   $$\Sigma X = 27+19+18+18+18+29+20+39+24+18+19 = 249.$$

The mean is

$$\bar{X} = \frac{\Sigma X}{n} = \frac{249}{11} = 22.636.$$

The mean age is 22.6 years, rounded to the nearest tenth of a year.

2. If a 71 year old is added to this group, there are now $n = 12$ students and the sum of the 12 ages is

$$\Sigma X = 27{+}19{+}18{+}18{+}18{+}29{+}20{+}39{+}24{+}18{+}19{+}71 = 320.$$

The mean is

$$\bar{X} = \frac{\Sigma X}{n} = \frac{320}{12} = 26.667.$$

Rounded to the nearest tenth of a year, the mean age is 26.7 years.

Note that that the addition of this one senior student, with a much older age than the others, raises the mean age by approximately four years. This was in contrast to the situation for the median, where the median age changed very little when the older student was added to the original group.

Notes on the mean

- **Ordering not necessary**. Unlike the case of the median, where the values must be placed in order from low to high, or high to low, to determine the middle value, such ordering is not necessary to obtain the mean. To calculate the mean, all that is required is to add all the values and divide by the number of cases.

- **Units**. The mean has the same units as the variable. In the example of the five grades for a first semester student, the mean is 74 per cent, assuming grades are reported in units of per cent; for age, the mean for the eleven students is 22.6 years, where the unit for age is the year.

- **Rounding**. In the above examples, the mean is rounded to one decimal place, given that the original data (grades and ages) may only be accurate to the nearest integer. Do not report too many decimals, but report a sufficient number of decimals to indicate the accuracy associated with the values of the variable and the guidelines in Chapter 4, section 4.7.2. Use the examples and exercises in this section as a guide concerning appropriate procedures for rounding. You may wish to refresh your memory on this issue by reading the text, section 4.7.2, pp. 126-134.

- **Unusual cases**. A data set such as the second part of the example of student ages (one much older student added to a group of younger students) demonstrate that the mean may not represent a "middle" or "typical" value of the variable. The mean is obtained by adding all the values, so each case is considered on a par with any other case in determining the mean. Where there is one or more unusual values in a set of data, this can create a mean either lower or higher than most values in the sample. That can make the mean somewhat unrepresentative, and in these cases, the mode or median may be preferred. For example, one or two very high incomes, among a group of people most of whom are middle income, can lead to a mean income that is unrepresentative of the whole group. In this case, the median may represent a more typical income than does the mean.

- **Obtaining total from the mean**. The mean is the total value of a variable divided by the number of cases. As a result, if the mean and the number of cases for a data set are known, the total can be obtained from this. Merely multiply the mean by the number of cases, to obtain the total. That is, if $\bar{X} = \Sigma X/n$, then $\Sigma X = \bar{X} \times n$.

  For example, if the mean income across one hundred households is estimated to be \$52,000, then an estimate of the total income for all one hundred households is $\$52,000 \times 100 = \$5,200,000$.

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_iX_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 2 provides a generic format for such a table. In this table, the variable X has k categories and $\Sigma f = n$ cases.

Table 2: Format of table for calculaing mean of grouped data

X	f	fX
X_1	f_1	f_1X_1
X_2	f_2	f_2X_2
.	.	.
.	.	.
.	.	.
X_k	f_k	f_kX_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 3. Obtain the mean credit hours for this sample of students.

Table 3: 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 3 are reorganized into a tabular format using the X and f notation in Table 4. The variable is "credit hours" and is labelled X – values for X are in the first

column of Table 4. 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 4. The entries in the third column of Table 4 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.

Table 4: 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

Proceeding through Table 4, 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 $f_1 X_1 = 1 \times 3 = 3$. 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, 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 5. These are earnings the students expect to receive after graduation.

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.

Table 5: 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 of Table 5 are reorganized in Table 6 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 6: 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	2,555

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 $\Sigma fX = 2,555$. 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

$$\bar{X} = \frac{\Sigma(fX)}{n} = \frac{2,555}{92} = 27.772$$

The mean anticipated annual earnings are 27.8 thousand dollars, or \$27,800, 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.