**Social Studies 201**
**September 24, 2004**

## Mode

See text, section 5.2, pp. 163-171.

This section of the notes gives a formal definition of the mode, some examples of how the mode is determined, and possible interpretations of the mode as a measure of central tendency.

> **Definition**. The mode is the most common value of the variable or the value that occurs most frequently. This value is sometimes referred to as the modal value of the variable.

Two examples of how the concept of the mode is used in reporting statistical results are as follows.

- Activity limitations related to pain: the most common form of disability among working age adults.
  Source: Statistics Canada, *A profile of disability in Canada, 2001*, from http://www.statcan.ca/english/freepub/89-577-XIE/workage.htm

- Jewish most likely target of hate crimes in 12 major police forces in Canada.
  Source: Statistics Canada, *Pilot survey of hate crime 2001 and 2002*, from http://www.statcan.ca/Daily/English/040601/d040601a.htm

While the word "mode" is not used in either of the above quotes, the concept is implicit in the statement. In the first example, there are various causes of diability (not listed in the quote) such as limited sight or hearing, and mobility or agility problems. Of all these causes, pain is the single most common cause of disability among working age adults. In the second quote, others such as Muslims, South Asians, or gays and lesbians were also targets of hate crimes. Among all those targeted, more hate crime was reported against those of Jewish background than against any other group.

The concept associated with the mode is also used in ordinary language when discussing fashion, peak, or capacity. Fashion is often considered the most popular style, that is, a style more common than any other. Peak or

capacity use of roads or bridges indicates the use that occurs more than any lesser use.

The mode is determined differently, depending on whether the data are presented in ungrouped or grouped format. The following notes illustrate the different methods for the two formats.

### Mode for ungrouped data

Data from a sample or population are considered ungrouped when presented as a list of all the numbers or values of the variable being considered. In this case, the mode is the value (or values) of the variable occurring most frequently. In order to determine the mode, count the number of times each value occurs. The mode is the value with the greatest count or tally.

> **Example 3.1 – Ages of undergraduate students**. A sample of eleven University of Regina undergraduate students gives ages of 27, 19, 18, 18, 18, 29, 20, 39, 24, 18, and 19 years for these students. What is the mode of age?
>
> **Answer**. In order to make it easier to see which value occurs most frequently, place the values in order so the number of cases can more easily be counted. In increasing order, by age, the eleven values are: 18, 18, 18, 18, 19, 19, 20, 24, 27, 29, and 38. From this ordered list, the mode is 18. That is, there are four 18-year-old respondents, two 19-year-olds, and one of each of the other ages. Since the value 18 occurs more frequently than any of the other values, the mode is 18 years of age. Alternatively stated, the modal value of age for these eleven student is 18 years.

**Codes, categories, or numbers**. The values of the variable need not be numerical, but could merely be names, codes, or categories. This is demonstrated in the following example.

**Example 3.2 – Provincial political party preference**. Political preference at the provincial level, for seven Saskatchewan residents (hypothetical names), is given in Table 1.

Table 1: Political preference for seven Saskatchewan adults

| Name | Party preference |
|---|---|
| Linda | Saskatchewan |
| Jennifer | NDP |
| Susan | NDP |
| Kyle | Saskatchewan |
| Sandra | Green |
| Tom | NDP |
| Mitchell | Liberal |

**Questions**

1. What is the mode of political preference?
2. If Brad, a Saskatchewan Party supporter, joins these seven adults, what is the new mode?

**Answer**. First, be clear concerning what the variable is – in this example the variable is political party preference at the provincial level. While the variable does not have numerical values or codes, the variable takes on values like NDP, Saskatchewan, Green, and Liberal.

1. From Table 1, there are three NDP supporters, two Saskatchewan Party supporters, and one each of Green and Liberal Party supporters. More individuals in this sample support the NDP than support any other single party, so the mode of political party preference is NDP. Alternatively stated, the modal political preference in NDP.

2. If Brad is added to the group, then there are three NDP supporters, three Saskatchewan Party supporters, and one each of Green and Liberal Party supporters. As a result, there are two modes, NDP and Saskatchewan Party – each of these latter political parties is supported by three individuals, more than the number supporting any other party. The modal political preference in this enlarged sample is NDP and Saskatchewan Party.

**Conclusion**. When data are presented in ungrouped format, the mode is determined by counting the number of cases taking on each value of the variable. The mode is then the value occurring most frequently.

### Mode for grouped data

Where the data have been grouped into categories or intervals, data are presented as tables, frequency distributions, diagrams, or histograms. In these situations, the mode is the value of the variable with the greatest frequency of occurrence. In the case of a histogram, the mode is the value of the variable where the histogram reaches its peak.

**Data presented in intervals**. When a variable is numerical, values of the variable are often collected together and grouped into intervals, as in Examples 3.6 and 3.7 that follow. When data are grouped in this manner, the mode can be reported as either the interval containing the mode, or the midpoint of the interval containing the mode.

**Equal interval widths**. If the intervals for a numerical variable are of equal interval width, then the mode is the interval, or midpoint of the interval, occuring most frequently.

**Unequal interval widths**. In the case of intervals of unequal interval width, the mode is the interval, or midpoint of the interval, having the greatest density or relative frequency. See text, section 4.8.3, pp. 149-159 for a discussion of histograms and density.

**Example 3.7 – Distribution of household income in Saskatchewan**. The data in the first two columns of Table 2 come from Statistics Canada, 2000 General Social Survey, *Cycle 14: Access to and Use of Information Communication Technology*. These data describe the distribution of income for households surveyed.

Table 2: Frequency distribution of household income, Saskatchewan respondents

| Income in thousands of dollars | Number of households ($f$) | Interval width ($w$) | Density ($f/w$) |
|---|---|---|---|
| 0 - 20 | 184 | 20 | 9.2 |
| 20 - 30 | 113 | 10 | 11.3 |
| 30 - 40 | 112 | 10 | 11.2 |
| 40 - 60 | 204 | 20 | 10.2 |
| 60-100 | 173 | 40 | 4.3 |
| 100 plus | 79 | — | — |
| Total | 865 | | |

**Question**. Obtain the mode of household income for respondents in Table 2.

**Answer**. There are more respondents (204) in the forty to sixty thousand income category, so it initially appears that the mode is this category. However, this category is twice as wide as the two next lower income categories, so there should be an adjustment for this. That is, interval width differs across the different income categories.

To correct for this, in the third column, interval width is obtained by subtracting the lower end point of each interval from the upper end point. This third column gives the interval width in thousands of dollars. Divide the frequencies of occurrence ($f$) in the second column by the interval widths ($w$) in the third column to obtain the densities ($f/w$) listed in the last column of the table. From the density column, the interval with the greatest density

is 20-30, with a density of 11.3. While this density is only slightly above the density of 11.2 for the 30-40 interval, the mode is the 20-30 interval. As a result, the modal income of households in this sample is \$20,000 to \$30,000 or, alternatively stated, \$25,000, the midpoint of this interval.

Note that there are almost as many households in the \$30-40,000 interval as in the \$20-30,000 interval. When describing these data, a researcher might indicate this by stating that households are more concentrated in the \$20-40,000 income range than in any other interval of the same width. Also note that the density cannot be calculated for the \$100,000 plus income category, since there is no upper limit given for this interval. However, it is very unlikely the mode would be in this highest income interval, since the households in this high income category are relatively few in number and are spread across a wide range of high incomes.

**Recap and summary – characteristics of the mode**

The mode is a simple and straightforward measure that is relatively easy to determine. While it is often useful when first examining data, it is limited in what it indicates only the peak of a distribution. The mode does not indicate how the values of a variable are distributed across values other than where the peak or greatest density occurs.

- **Fashion**. One meaning of "mode" is fashion or what is most common or popular. That is, the mode refers to the the value or characteristic adopted by more people than any other value or characteristic. Where a sample or population has a distinctive fashion or characteristic, this modal characteristic is often worth reporting.

- **Not necessarily at the centre**. For some distributions, the mode may not be at the centre of a distribution, and may not really indicate the average, as the average is commonly understood. We often consider the average to be the centre of a distribution, but in the example of ages of eleven students (Example 3.1), the mode was the smallest value, age 18. The modal age was 18 years since many students enter the University aged 18, directly after completing high school. This produces a distribution with more students at this age than at any other age. While the mode may not indicate the centre in examples such as this, it is a useful measure to report when a distribution such as age of student has such a distinctive value.

- **Peak or capacity**. One use for the mode is when attempting to determine peak use or capacity utilization. For example, a power company would want to ensure that there is sufficient power generating capacity so there are no power blackouts, and that it can meet peak need for electical power. The company should plan its resources so it can meet this modal or peak use.

- **More than one mode**. A set of data, or a distribution, may have more than one mode. In the answer to Example 3.5, it was noted that if there had been one more male respondent with response 4, both 3 and 4 would be modes. This demonstrates the possibility that there could be two or more values that occur more frequently than other values. Where there are several modes in a distribution, it may be less

worthwhile reporting these modes, as compared with a distribution having a single mode. Where there are two or more modes, the modes cannot all be at the centre of the distribution and these most common values may be spread across many values of the sample or population.

- **Nominal scale**. One advantage of the mode is that it can be obtained for any data set. The median and mean require variables with an ordinal, interval, or ratio level of measurement, but the mode can be obtained for a variable with no more than a nominal scale of measurement.

- **Distinctiveness**. The mode is a measure that is useful in describing and analyzing the distribution of a variable when the distribution has a distinctive peak. When a distribution does not have a distinctive peak, so that there are several values, each of which occurs may times, the mode is a less useful measure.

- **Obtaining the mode**. The mode is the value of the variable that occurs most frequently. With ungrouped data, the mode is obtained by merely counting the number of times each value of the variable, with the mode being the value occurring most frequently. For data grouped into categories of equal interval width, the mode is the category, or midpoint of the category, that occurs most frequently. For data grouped into categories of unequal interval width, the mode is the category, or midpoint of the category, having the greatest density. Where a distribution is presented as a histogram, the mode is the value of the variable at the peak of the histogram.

## Median

See text, section 5.3, pp. 171-183.

### Introduction

This section of the notes gives a formal definition of the median, provides examples of how the median is determined, and suggests ways of interpreting the median as a measure of centrality. As with the mode, different methods are used to calculate the median depending on whether data are ungrouped or grouped.

### Defining the median

Roughly speaking, the median is the middle value, where values of a variable are ordered from low to high, or high to low. Unlike the mode, which can be obtained for any type of variable, the median requires data that are measured at least at the ordinal level of measurement. That is, it must be possible to order the values of the variable as less than or greater than – the median is the middle value of the ordered data set. Examples of ordinal level variables are attitudes or opinions (measured as strongly disagree to strongly agree, or on a 1-5 or 1-7 scale), order of finish in a contest or race, or social science scales such as degree of stress, IQ, and so on.

Data that are measured at the interval or ratio level, such as age or income, are also ordinal, in that each of age or income can be ordered as greater than, less than, or equal to other ages or incomes. Thus the median can be obtained for variables that are ordinal, interval, or ratio, but it cannot be obtained if the level of measurement is no more than nominal. (For a review of levels of measurement, see section 3.2 of the text).

> **Definition**. The median is the value of a variable such that one-half of the values are less than or equal to this value, and the other one-half of the values of the variable are greater than or equal to this value.

The median is thus the middle value of a distribution, dividing the cases equally into those less than or equal to the median and those greater than or equal to the median. A a measure of centrality, the median is very useful, since it indicates the value of the variable that splits a population in half and

is in the middle of the distribution. Researchers and analysts often consider the median to represent a "typical" value for a population.

**Using the median**

When examining the way incomes are distributed in a population, the median indicates the middle income and represents some sort of typical member of the population. For example, in the 2001 Census of Canada, Statistics Canada reported that the median Saskatchewan family income for the year 2000 was $49,300. While no single number can summarize incomes of all families in the province, this figure of $49,300 represents a sort of typical Saskatchewan family income, with one-half of families having income less than this and the other half having an income of $49,300 or more. By contrast, the median family income for Canada was $55,000, indicating that a "typical" Saskatchewan family had lower income than its counterpart across Canada as a whole.

**Source**: Statistics Canada, "Median Family Income in Constant (2000) Dollars for all Census Families, for Canada, Provinces, and Territories – 20% Sample Data. Obtained August 23, 2004 from web site:
http://www12.statcan.ca/english/census01/products/highlight/Income/Index.cfm?Lang=E

The following two quotes illustrate ways the median can be used.

> "An international comparison shows Canada's teenage pregnancy rate as middling among those of industrial nations."
>
> **Source**: *National Post*, October 20, 2000, p. A4.
>
> "Median net worth for all families rose about 10% from 1984 to 1999, but this net worth was not shared equally by all types of families. For instance, the median net worth of young couples with children fell 30% ... "
>
> **Source**: Statistics Canada, *The Daily*. From web site
> http://www.statcan.ca/Daily/English/020222/d020222a.htm

In the first example, the word "median" is not used, but the quote makes clear that about one-half of industrial countries have lower teen pregnancy rates while the other half have greater rates. While the median is not stated, the implication is that the Canadian teen pregnancy rate is at the median, or middle, level among industrial countries.

In the second example, the wealth of families increased in that the middle value of net worth increased about ten per cent over the fifteen years. Unfortunately, this has not benefited all families, so that a "typical" younger family with children had a thirty per cent lower wealth, or net worth.

**Obtaining the median**

While the definition of median is the same for all types of data, the method of obtaining the median differs depending whether the data are presented in ungrouped and grouped format. The method of obtaining the median for each of ungrouped and grouped data, along with examples, are presented in the following notes.

**Median for ungrouped data**

Beginning with a variable that has at least an ordinal level of measurement, but where there is a list of the values of the variable across all members of the sample or population, it is first necessary to place the values in order. That is, for ungrouped data, reorganize all the values in order from low to high, or high to low. Count the number of values in the data set and divide this by two, to determine the middle value. Then count the ordered values of the variable until the middle value is reached. This is the median value. That is, the median value of a variable is the middle value when the values are ordered.

**Middle value**. To determine the middle case, divide the total number of cases by two. If there are an odd number of cases in the data set, the median case is the case just after one-half of the total. That is, if there are fifteen cases, one-half of fifteen is seven and a half, so the middle value is the eighth case. If there are an even number of cases, the median is the two middle cases: the case representing one-half of the total number and the next case. In this situation, the median is either the two middle values or the average of these two values. For example, if there are sixteen cases, one-half of sixteen is eight – the middle cases are the eighth and ninth cases, or the average of these two cases.

**Example 3.8 – Ages of a sample of undergraduate students**. From a sample of eleven University of Regina undergraduates, the ages are 27, 19, 18, 18, 18, 29, 20, 39, 24, 18, and 19 years. (Sample drawn from the SSAE98 data set).

**Question**. Use these data to determine the following.

1. Obtain the median age of these eleven students.
2. If a 19 year old student joins these eleven, what is the median age of the twelve students?
3. If a 20 year old student joins the first eleven, what is the median age of the twelve students?
4. If a senior, aged 71 years, decides to return to university and joins the original eleven students, what is the median age of the twelve students? Suppose this returning student was not a senior, but an individual aged 51 years of age. Would this change the median?

**Answer**. The variable in this example is the age of students in years, a variable with a ratio level of measurement. As such, this variable is certainly ordinal, in that the ages can be ordered from low to high or high to low.

1. Since the median is the middle of an ordered set of numbers, begin by placing the numbers in order of their values. Ordered from the smallest to the largest age, the numbers representing the ages of the eleven students are 18, 18, 18, 18, 19, 19, 20, 24, 27, 29, and 38.

   Since there are eleven values of age, the middle value is at the sixth of these (one-half of eleven is 5.5, and the next case is the sixth). Counting from the smallest to the largest age, the first, second, third, and fourth students are 18 year olds, the fifth is a 19 year old, and the sixth is also a 19 year old. The median age for these students is thus 19 years. Note that there are five students at lower or equal age (18, 18, 18, 18, 19) and five students with greater ages (20, 24, 27, 29, and 38).

2. If a 19 year old is added to the original eleven, in order the ages are now: 18, 18, 18, 18, 19, 19, 19, 20, 24, 27, 29, and 38. With twelve students, the middle values are the sixth and seventh. That is, one-half of twelve is six, and the median divides the cases into the lowest six and highest six cases. Counting from the lowest to the highest ages, the sixth and seventh values are each 19, so the median age for these twelve students is 19 years.
3. If a 20 year old is added to the original eleven, the new ages ordered from low to high are: 18, 18, 18, 18, 19, 19, 20, 20, 24, 27, 29, and 38. There are again twelve values and the middle values are the sixth and seventh values. Counting from the lowest to the highest, the sixth and seventh values are 19 and 20. There are two ways of reporting the median in this case. The median can be reported as 19 and 20 years of age. Alternatively, the median can be reported as the simple average of these two, that is, 19 plus 20 divided by two, or 19.5. This is
$$\text{median} = \frac{19 + 20}{2} = 19.5$$
4. If a 71 year old is added to the original eleven, in order the ages are now: 18, 18, 18, 18, 19, 19, 20, 24, 27, 29, 38, and 71. There are again twelve values and the middle values are the sixth and seventh values. As with the previous answer, counting from the lowest to the highest, the sixth and seventh values are 19 and 20, so the median age is 19.5 years. This last example illustrates one property of the median – it is insensitive to values at the extremes of a set of data. In this case, replacing the 20 year old with a 71 year old did not change the median.

   If this student had been only 51, rather than 71 years of age, this would not change the median. In order, the ages would then be 18, 18, 18, 18, 19, 19, 20, 24, 27, 29, 38, and 51, so the middle value or median would again be 19.5 years of age.

**Ordered non-numerical categories**

In order to obtain the median, the values of the variable need not be numbers and could be names or categories, so long as these names or categories can be meaningfully ordered. In the following example, the median is obtained for a variable measuring respondents' extent of agreement or disagreement on an opinion question.

**Example 3.9 – "Help themselves"**. A sample of nine individuals from the SSAE98 data set gave the responses in Table 3. The responses represent views about the statement "The more money spent helping people, the less they will help themselves." Respondents could answer on a five-point scale: strongly disagree (SD), somewhat disagree (D), neutral (N), somewhat agree (A), and strongly agree (SA).

Table 3: View concerning "Help themselves"

| Identification Number | View |
|---|---|
| 1 | Somewhat disagree (D) |
| 2 | Strongly disagree (SD) |
| 3 | Neutral (N) |
| 4 | Somewhat disagree (D) |
| 5 | Strongly disagree (SD) |
| 6 | Somewhat agree (A) |
| 7 | Somewhat disagree (D) |
| 8 | Somewhat disagree (D) |
| 9 | Neutral (N) |

**Question**. Use the data in Table 3 to answer the following.

1. What is the median response? Explain how you obtained this.
2. If respondent identification number 2 leaves the group, explain why the median does not change.

3. If respondent identification number 2 leaves the group and two respondents who strongly agree with the statement join the group, explain how this changes the median.
4. If the new set of ten respondents in part 3 is joined by one more individual who strongly agrees with the statement, what is the identification number of the indvidual with the new median?

**Answer**. In this example, the variable is extent of agreement or disagreement with the statement. While responses such as strongly agree or neutral are not numerical, responses can be meaningfully ordered, so the median can be obtained.

1. To obtain the median, order the values from low to high, or high to low (it makes no difference which direction you go, you are attempting to find the middle value). Using the short form for the responses, in order from strongly disagree to strongly agree, the responses are

   SD, SD, D, D, D, D, N, N, A

   There are nine responses so the middle response is at $9/2 = 4.5$, or the fifth value. Counting from this ordered list, there are two SDs, the third value is D, the fourth is D, and the fifth is D. So the median is D, or somewhat disagree.
2. If the respondent with identification number 2 leaves the group, there are eight individuals remaining whose responses, in order, are

   SD, D, D, D, D, N, N, A

   Given the eight values, the median is the fourth and fifth values ($8/2 = 4$), since the values are split exactly in half at these values. Counting from the first SD, the fourth and fifth values are D and D, so the median is still D, or somewhat disagree.
3. If the eight individuals of part 2 are joined by two individuals who strongly agree, in order the responses are

SD, D, D, D, D, N, N, A, SA, SA

There are now ten values, with the middle values being the fifth ($10/2 = 5$) and sixth values. Counting from the greatest disagreement and proceeding toward agreement, there is one SD, and four Ds, so the fifth value is a D, followed by the sixth value of N. The median is thus D and N, that is, the median values are somewhat disagree and neutral. Given there are no numbers attached to these two responses, they cannot really be averaged. While it is awkward to report two values for the median, the middle of this list is really at these two values. Alternatively, it could be reported that the median is between somewhat disagree and neutral.

4. If one more strongly agree joins the group, the ordered responses are

   SD, D, D, D, D, N, N, A, SA, SA, SA

   There are now eleven responses, so the median is at the sixth value ($11/2 = 5.5$). Counting from the left, the sixth value is N, so the median is now a neutral response. There are two individuals, identification numbers 3 and 9 with a neutral response, so either or both of these individuals have the median response.

**Recap and summary**

Some of the issues to keep in mind when obtaining the median for ungrouped data are the following.

- To obtain the median, the variable must have at least an ordinal level of measurement. It is not possible to obtain the median for a variable measured at no more than the nominal level. For example, it would make no sense to have the median ethnicity or religious affiliation – variables that are no more than categorizations and have no meaningful order attached to them.
- For an ungrouped data set, begin by putting values in order from low to high or high to low. The median is the middle value.
- When ordering the list of value, make sure you include all cases. Count the number of cases before and after ordering to ensure you have not ignored or left out any cases. In the example of ages of students, the original list had eleven cases, so the ordered set must also have eleven values listed.
- To determine the median case, divide the total number of cases by two. If there are an odd number of cases in the data set, the median case is the case just after one-half of the total. If there are an even number of cases, the median is the two middle cases: the case representing one-half of the total number and the next case. In this situation, the median is either the two middle values or the average of these two values.
- The median is ordinarily a number but, as demonstrated in the example of a list of attitudes, it can be the name of the middle value.

The next section of these notes discusses how to obtain the median when data are grouped into categories or intervals.

**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 $\Sigma X$. 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 ($\Sigma X$) 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 4. 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 4: 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 $\Sigma X$ | 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 $\$52,000 \times 100 = \$5,200,000$.