Social Studies 201

October 13, 2004

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

Standard deviation and variance for grouped data

See text, section 5.9.2, pp. 237-249.

When working with data that are grouped into categories or intervals, the variance and standard deviation are again obtained using the deviations about the mean and the squared value of these. But in this case, the sum of squares of differences1 about the mean are weighted by the number of times each occurs. The definitions are as follows.

> **Definition**. A variable X that takes on values $X_1, X_2, X_3, ...X_k$ with respective frequencies $f_1, f_2, f_3, ...f_k$ has mean
>
> $$\bar{X} = \Sigma(fX)/n$$
>
> where $n = \Sigma f$. From Module 3, this is the formula for the mean using grouped data.
>
> The variance is
>
> $$\text{Variance} = s^2 = \frac{\Sigma f(X - \bar{X})^2}{n-1}$$
>
> and the standard deviation is the square root of the variance, that is,
>
> $$\text{Standard deviation} = s = \sqrt{s^2}.$$

Beginning with data that are grouped, again the mean is obtained first, then the difference of each value of X from the mean is calculated. These deviations about the mean are squared and then multiplied, or weighted, by the respective frequencies of occurrence (f). While the above method can be used, it results in awkward and time-consuming calculations. A more straightforward procedure is to use the following alternatie formula. See text pp. 242-244.

Alternative formula. An alternative formula for the variance and standard deviation for grouped data, giving exactly the same result, is

$$\text{Variance} = s^2 = \frac{1}{n-1}\left[\Sigma fX^2 - \frac{(\Sigma fX)^2}{n}\right].$$

The standard deviation is the square root of the variance

$$\text{Standard deviation} = s = \sqrt{s^2}.$$

The formula given in the defintion is computationally inefficient, so in this class we generally use the alternative formula for the variance. In the examples and exercises that follow, only the alternative formula is used. See text pp. 242-249 for proof of equivalence of formulae and an example.

Steps used in tabular format. As with ungrouped data, a tabular format is ordinarily used to obtain the variance and standard deviation. Employing the alternative formula, the procedure for calculating the variance and standard deviation is as follows:

- 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 fX with the products of the f (from second column) and the X (from first column).
- Sum the f entries in the second column to determine the sample size, that is, $n = \Sigma f$.
- Sum the products, fX, in the third column to obtain the column total ΣfX. Divide this sum by n to determine the mean of $\bar{X} = \Sigma fX/n$. To this step, the procedures are identical to those used in Module 3 for calculating the mean of grouped data.
- Create a fourth column with values of fX^2, that is the f multiplied by the square of the X. The square of the X values multiplied by the frequencies f are entered into the fourth column. Also, this is equivalent to multiplying the fX of the third column by another X (the value in the first column). That is, the fourth column is the entry

in the third column, multiplied by the entry in the first column. This produces the values of fX^2.

- Sum the values in the fourth column to obtain ΣfX^2.
- The sums of the fourth column (ΣfX^2) and the third column (ΣfX) are entered into the formula

$$s^2 = \frac{1}{n-1}\left[\Sigma fX^2 - \frac{(\Sigma fX)^2}{n}\right]$$

and this is the variance.

- The standard deviation s is the square root of the variance of the last step, that is,

$$s = \sqrt{s^2}$$

.

In terms of procedures, it is generally preferable to proceed row by row. Begin by entering all the X values in the first column and all the frequencies f in the second column. Then proceed row by row to obtain the entries in the third and fourth columns. That is, for the first row, calculate fX and enter it in the third column; then obtain fX^2 and enter it in the fourth column. Then go to the second row and do the same, and so on, until all the entries in each row have been obtained. Finally, sum all the entries in each column and enter the sums into the appropriate place in the formula for the variance and standard deviation.

The following example illustrates the use of the alternative formula.

Example 4.8 – Variation in alcohol consumption by income level. The data in Table 1 is adapted from Statistics Canada, *1991 General Social Survey - Cycle 6: Health.*

Table 1: Distribution of Saskatchewan adults by number of alcoholic drinks consumed per week and by personal income

Number of drinks per week	Number of respondents with personal income	
	$< \$20{,}000$	\$20,000 plus
None	178	91
1-5	76	77
6-10	28	28
11-15	12	21
16-48	8	18
Total	302	235

Use these data to compute the mean and standard deviation of the number of alcoholic drinks consumed per week for adults in each of the two income groups. Using these statistics and the data in Table 1, write a short note comparing the two distributions.

Answer. The variable in this question is number of alcoholic drinks consumed per week. Assuming that any drink is equal in alcoholic content to any other drink, this variable has a ratio scale. That is, the unit is one drink and a value of zero means no alcoholic consumption.

Low income group. The table for the calculations of the mean and standard deviation of the number of alcoholic drinks consumed per week for Saskatchewan adults at the lower income level is Table 2. This table is constructed for the alternative formula, that is, with columns for the values of f, X, fX, and fX^2.

The first step is to obtain the X values associated with each category into which the data are grouped – these are the midpoints of

Table 2: Calculations for measures of variation of number of alcoholic drinks consumed per week – income of less than $20,000

Number of drinks per week	X	f	fX	fX^2
None	0	178	0	0
1-5	3	76	228	684
6-10	8	28	224	1,792
11-15	13	12	156	2,028
16-48	32	8	256	8,192
Total		302	864	12,696

each interval (0, 3, 8, 13, and 32). The frequencies of occurrence for each interval, f, are copied from Table 1. The next column is the fX column and the final column the fX^2 column.

For the first row, multiply the f value of 178 by the X value of 0. This gives a value of $fX = 0$ and $fX^2 = 0$.

For the second row, $f = 76$ and $X = 3$, so the entry in the fX column is $76 \times 3 = 228$. Multiplying this fX by another X produces $228 \times 3 = 684$. This is the value entered into the final column. Note also that $fX^2 = f \times X \times X = 76 \times 3 \times 3 = 684$. However, it is more efficient to obtain this entry for the final column by multiplying the entry in the fX column by another X.

For the third row, $f = 28$ and $X = 8$, that is, there are 28 respondents who drink an average of 8 drinks per week. The entry in the fX column is $28 \times 8 = 224$. Multiplying this fX by another X produces $224 \times 8 = 1,792$ and this is the value entered into the fX^2 column.

For the fourth row, $f = 12$, $X = 13$, so $fX = 12 \times 13 = 156$. Multiplying this by another $X = 13$ gives $fX^2 = 156 \times 13 = 2,028$.

Finally, the last row has $f = 8$ and $X = 32$, so $fX = 8 \times 32 = 256$ and $fX^2 = fX \times X = 256 \times 32 = 8,192$.

The sum of the f column is $n = 302$, the sum of the entries in the fX column is $\Sigma fX = 864$, and the sum of the entries in the last column is $\Sigma fX^2 = 12,696$.

Using these values in the formulae for the mean and variance gives

$$\bar{X} = \frac{\Sigma fX}{n} = \frac{864}{302} = 2.861$$

or a mean of 2.9 drinks per week. The variance is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left[\Sigma fX^2 - \frac{(\Sigma fX)^2}{n}\right] \\ &= \frac{1}{301}\left[12,696 - \frac{864^2}{302}\right] \\ &= \frac{1}{301}\left[12,696 - \frac{746,496}{302}\right] \\ &= \frac{12,696 - 2,471.841}{301} \\ &= \frac{10,224.159}{301} \\ &= 33.967. \end{aligned}$$

The standard deviation is the square root of the variance, or

$$s = \sqrt{s^2} = \sqrt{33.967} = 5.828.$$

Rounded to the nearest tenth of a drink, the standard deviation for lower income individuals is 5.8 drinks per week.

Higher income group. For the higher income group,the tabular format for the calculations is provided in Table 3. The format and procedures are the same as for the lower income group. From Table 3, the sum of the f column is $n = 235$, the sum of the

Table 3: Calculations for measures of variation of number of alcoholic drinks consumed per week – income of $20,000 plus

Number of drinks per week	X	f	fX	fX^2
None	0	91	0	0
1-5	3	77	231	693
6-10	8	28	224	1,792
11-15	13	21	273	3,549
16-48	32	18	576	18,432
Total		235	1,304	24,466

entries in the fX column is $\Sigma fX = 1,304$ and the sum of the entries in the last column is $\Sigma fX^2 = 24,466$.

From the sums in Table 3, the mean is

$$\bar{X} = \frac{\Sigma fX}{n} = \frac{1,304}{235} = 5.549$$

or 5.5 drinks per week. The variance is

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left[\Sigma fX^2 - \frac{(\Sigma fX)^2}{n}\right] \\ &= \frac{1}{234}\left[24,466 - \frac{1,304^2}{235}\right] \\ &= \frac{1}{234}\left[24,466 - \frac{1,700,416}{235}\right] \\ &= \frac{24,466 - 7,235.813}{234} \\ &= \frac{17,230.187}{234} \\ &= 73.633 \end{aligned}$$

and the standard deviation is $s = \sqrt{s^2} = \sqrt{73.633} = 8.581$ or 8.6 drinks per week.

These statistics are summarized in Table 4. From these statistics, higher income individuals both consume more alcohol, and are more varied in their alcohol consumption pattern, than lower income individuals. Mean alcohol consumption for those with higher incomes (5.5 drinks per week) is almost double that for those with lower incomes (2.8 drinks per week). In addition, the standard deviation for higher income individuals is 8.6 drinks per week, over half again as great as for the lower income individuals ($s = 5.8$ drinks per week).

Table 4: Summary Statistics for Alcoholic Drinks Consumed

Measure	Low Income	High Income
Mean $= \bar{X}$	2.9	5.5
Variance$= s^2$	34.0	73.6
s.d $= s$	5.8	8.6