Social Studies 201

October 18, 2006

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 $n = \Sigma f$, 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.

Table 1: Frequency distribtions of Saskatchewan respondents, classified by number of alcoholic drinks consumed per week, low and high income

No. of alcoholic drinks drinks per week	No. of respondents with income of: < \$30,000	\$30,000 plus
None	370	188
1-4	214	185
5-9	94	106
10-19	54	74
20-39	30	29
Total	762	582

Alcohol consumption. The distributions of alcohol consumption for low and high income individauls in Saskatchewan is given in Table 1. From this table, the calculations for the mean and standard deviation are given in Table 2.

Table 2: Calculations for mean and standard deviation of alcohol consumption, low and high income individuals

	Low income			High income		
X	f	fX	fX^2	f	fX	fX^2
0	370	0.0	0.0	188	0.0	0.00
2.5	214	535.0	1,337.5	185	462.5	1,156.25
7.0	94	658.0	4,606.0	106	742.0	5,194.00
14.5	54	783.0	11,353.5	74	1,073.0	15,558.50
29.5	30	885.0	26,107.5	29	855.5	25,237.25
Total	762	2,861.0	43,404.5	582	3,133.0	47,146.00

For those with low income,

$$\begin{aligned}
\bar{X} &= \frac{2,861.0}{762} = 3.755 \\
s^2 &= \frac{1}{n-1}\left(\Sigma fX^2 - \frac{(\Sigma fX)^2}{n}\right) \\
&= \frac{1}{761}\left(43,404.50 - \frac{2,861.0^2}{762}\right) \\
&= \frac{1}{761}\left(43,404.50 - 10,741.89\right) \\
&= \frac{1}{761}\left(32,662.61\right) \\
&= 42.921 \\
s &= \sqrt{s^2} = \sqrt{42.921} = 6.551.
\end{aligned}$$

The mean number of alcoholic drinks consumed per week for those with low income is 3.8 drinks and the standard deviation is 6.6

drinks. The CRV is 175.6.

$$\text{CRV} = \frac{s}{\bar{X}} \times 100 = \frac{6.551}{3.755} \times 100 = 174.5$$

For those with higher incomes,

$$\begin{aligned}
\bar{X} &= \frac{3,133.0}{582} = 5.383 \\
s^2 &= \frac{1}{n-1}\left(\Sigma f X^2 - \frac{(\Sigma f X)^2}{n}\right) \\
&= \frac{1}{581}\left(47,146.00 - \frac{3,133.0^2}{582}\right) \\
&= \frac{1}{581}\left(47,146.00 - 16,865.45\right) \\
&= \frac{1}{581}\left(30,280.55\right) \\
&= 52.118 \\
s &= \sqrt{s^2} = \sqrt{52.118} = 7.219.
\end{aligned}$$

The mean number of alcoholic drinks consumed per week for those with high income is 5.4 drinks and the standard deviation is 7.2 drinks. The CRV is 134.1.

$$\text{CRV} = \frac{s}{\bar{X}} \times 100 = \frac{7.219}{5.383} \times 100 = 134.1$$

A summary of the statistics for the two groups is contained in Table 3.

From Table 3, the answer to this question is not entirely clear cut. In terms of actual amount of alcohol consumed per week, those in the higher income category have greater variation in that the variance and standard deviation for the high income group (52.1 and 7.1) exceed these same statistics for those of lower income (43.5 and 6.6). But the standard deviations are little different, so perhaps the CRV provides a better comparison. The CRV for those of lower incomes (174.5) is considerably greater than

Table 3: Summary of statistics for alcohol consumption of low and high income individuals

	Group	
Statistic	Low income	High income
Mean	3.8	5.4
Variance	43.5	52.1
Std. dev.	6.6	7.2
CRV	174.5	134.1

the CRV for those with higher incomes (134.1). This is because the mean is lower for those with low income. For those of lower income, the mean is lower and and the standard deviations for the two groups are similar, so this produces a larger CRV for low income and a smaller CRV for those with high incomes.

Last edited October 18, 2006.