2.9: Measures of Variance and Standard Deviation on Grouped Data

Learning Objectives

- Determine range, variance, and standard deviation of grouped discrete data
- Determine range, variance, and standard deviation of grouped relative frequency data

Section 2.9 Excel File (contains all of the data sets for this section)

Introduction to Measures of Spread on Grouped Data

As mentioned in the previous section, there are times when we may be given data in a summarized frequency distribution format instead of a collection of "raw" data. We now examine finding descriptive statistic measures of dispersion in grouped data: range, variance, and standard deviation. If the data is grouped over intervals, we can only estimate such measures since the grouping action has caused us to lose some data information. However, if the data is grouped into single-value classes, we can usually produce the same spread measures as if we had the raw data.

Range, Variance, and Standard Deviation of Grouped (non-interval) Data in a Frequency Table

Look at the frequency table from Text Exercise $2.8.1$ regarding quiz scores of twenty students:

Table 2.9.1: Grouped (non-interval) Frequency Distribution of Quiz 1 Data

Quiz Scores	Frequency
5	2
6	6
7	5
8	4
9	3

With such a table, we can easily find the range with no new ideas needed. The table shows the minimum data value is 5 and the maximum is 9. Because $9-5=4$, the range is a score difference of 4.

Next, assuming the data is population data, we examine the variance and standard deviation. Recall that variance is the average of all the various squared deviations of the individual data values from the mean of the data. In the previous section, we found the mean of this distribution to be $\mu=\frac{\sum(x_j\cdot f_j)}{\sum f_j}=\frac{140}{20}=7$. In a new column, we can construct the deviations from the mean and square those deviations as a first step. It becomes imperative that we not get tied up with all the messy numbers as we move through our work but keep our focus on what we are measuring:

Table 2.9.2: Computation of squared deviations from the mean

x_j	f_j	$(x_j-\mu)^2$
5	2	$(5-7)^2=4$
6	6	$(6-7)^2=1$
7	5	$(7-7)^2=0$
8	4	$(8-7)^2=1$
9	3	$(9-7)^2=4$
Totals:	$\sum f=20$	

Before we can average these squared deviation measures, we must remember that some squared deviations occur more frequently (as given by the frequency column) than others. For example, the frequency column shows that many more squared deviations are tied to the data values of 6 than those of the squared deviations associated with the data value of 9. We must "weight" these squared deviations by the frequency of their occurrence to account for the various twenty data values. In another new column, we form the products of $(x_j-\mu)^2\cdot f_j$ to accomplish this, and then proceed to "average" those weighted squared deviations:

Table 2.9.3: Computation of the variance

x_j	f_j	$(x_j-\mu)^2$	$(x_j-\mu)^2\cdot f_j$
5	2	4	$4\cdot 2=8$
6	6	1	$1\cdot 6=6$
7	5	0	$0\cdot 5=0$

x_j	f_j	$(x_j-\mu)^2$	$(x_j-\mu)^2 \cdot f_j$
8	4	1	$1 \cdot 4 = 4$
9	3	4	$4 \cdot 3 = 12$
Totals:	$\sum f_j = 20$		$\sum (x_j-\mu)^2 \cdot f_j = 30$
		Variance:	$\sigma^2 = \frac{\sum\left[(x_j-\mu)^2 \cdot f_j\right]}{\sum f_j} = \frac{30}{20} = 1.5$

As shown in the last two rows above, with our weighted squared deviations determined, we can find a meaningful average by summing our squared deviations and dividing by the number of data values involved. We have a variance measure of σ^2\(=1.5\) in the bottom right table cell. We can find the standard deviation by taking the square root of our variance: $\sigma = \sqrt{\sigma^2} = \sqrt{1.5} \approx 1.2247$. We note that our calculation work was for population data. If the table referenced sample data, we would have divided by $\sum(f_j) - 1$ instead on the last computation:

Variance and Standard Deviation from a Frequency Distribution

$$\text{Variance for sample data: } s^2 = \frac{\sum\left[(x_j-\bar{x})^2 \cdot f_j\right]}{\sum(f_j)-1}$$

$$\text{Standard Deviation for sample data: } s = \sqrt{\frac{\sum\left[(x_j-\bar{x})^2 \cdot f_j\right]}{\sum(f_j)-1}} = \sqrt{s^2}$$

$$\text{Variance for population data: } \sigma^2 = \frac{\sum\left[(x_j-\mu)^2 \cdot f_j\right]}{\sum f_j}$$

$$\text{Standard Deviation for population data: } \sigma = \sqrt{\frac{\sum\left[(x_j-\mu)^2 \cdot f_j\right]}{\sum f_j}} = \sqrt{\sigma^2}$$

As mentioned, working in a spreadsheet will often make the computation quicker and easier, especially when working with such columns of information.

? Text Exercise 2.9.1

The frequency table from Section 2.1 shows thirty student scores (discrete 10-point scale) for an assignment.

Table 2.9.4: Student quiz score data

Student Score	Frequency
3	1
4	1
5	3
6	5
7	5
8	7
9	5
10	3

Find the range, variance, and standard deviation of this grouped (non-interval) data. Assume the data is a sample from a larger population.

Answer

The table shows that the minimum data value is 3 and the maximum data value is 10. Because $10 - 3 = 7$, the range is a score difference of 7.

Next, we determine the variance by averaging the squared deviations from the mean, once weighted by the frequency counts. Since this is sample data, we also remember to divide by one less in the averaging step. In the previous section, we found the mean of this distribution to be $\bar{x} = \frac{\sum(x_j \cdot f_j)}{\sum f_j} = \frac{218}{30} \approx 7.2667$. We can construct the deviations from the mean and square those as the first step. Again, it becomes imperative that we not get distracted by the messy decimals we compute in our work. Keep focusing on what we measure in the computational work (squared deviations from the mean to be averaged) and calculate it accurately.

Table 2.9.5 Computation of square deviations from the mean

x_j	f_j	$(x_j-\mu)^2$
3	1	$(3-7.2667)^2 \approx 18.2044$
4	1	$(4-7.2667)^2 \approx 10.6711$
5	3	$(5-7.2667)^2 \approx 5.1378$
6	5	$(6-7.2667)^2 \approx 1.6044$
7	5	$(7-7.2667)^2 \approx 0.0711$
8	7	$(8-7.2667)^2 \approx 0.5378$
9	5	$(9-7.2667)^2 \approx 3.0044$
10	3	$(10-7.2667)^2 \approx 7.4711$
Totals:	$\sum f_j = 30$	

Next, we weight our various squared deviations by their frequency of occurrence, forming the products of $(x_j-\mu)^2 \cdot f_j$ to accomplish this:

Table 2.9.6 Computation of variance

x_j	f_j	$(x_j$
3	1	18.
4	1	10.
5	3	5.1
6	5	1.6
7	5	0.7
8	7	0.5
9	5	3.0
10	3	7.4
Totals:	$\sum f_j = 30$	
		Variance:

Again, our last row of the table shows the "averaging" of those weighted squared variations when working with sample data. If the data were given as population data, we would have divided by the sum of the frequencies instead of one less than that sum. This would have resulted in a slightly different value of $\sigma^2 \approx 3.1289$. When calculating, we must vigilantly remember the difference between sample and population variance. These data measures are distinct and continue to emphasize one, among other, reasons for knowing if the data is sample data or population data. As shown by the bottom right measure in the table, this grouped sample data has a variance measure of $s^2 \approx 3.2368$. The sample standard deviation measure is: $s = \sqrt{s^2} \approx \sqrt{3.2368} \approx 1.7991$. We can compare this result with the calculated variance for the ungrouped data set to see that we have produced the same measure.

Range, Variance, and Standard Deviation of Grouped (non-interval) Data in a Relative Frequency Table

What must we do to measure the variation of the data if, instead of a frequency distribution table, we have a relative frequency distribution of population data? The approach presented here only produces valid measures in population data since the relative frequency measures do not always disclose the sample size. Recall the sample size is essential for the computation of the sample variance as we must use $n-1$ in our averaging step. We start with the same example of twenty quiz scores.

Table 2.9.6: Grouped Relative Frequency Distribution of Quiz 1 Data

Quiz Scores	**Relative Frequency $P(x_j)$**

Quiz Scores	Relative Frequency $P(x_j)$
5	$\frac{2}{20} = 10\%$
6	$\frac{2}{20} = 30\%$
7	$\frac{5}{20} = 25\%$
8	$\frac{4}{20} = 20\%$
9	$\frac{3}{20} = 15\%$
Totals:	$\sum P(x_j) = 100\%$

Again, we can easily find the range--the table shows the minimum data value is 5 and the maximum data value is 9, leading us to a range of $9 - 5 = 4$.

Next, we use the same ideas as above to determine the variance and standard deviation of the data from this table. We need to find the average squared deviations of the data values from the mean. In Section 2.8, we found the mean of this relative frequency distribution to be $\mu = \sum (x_j \cdot P(x_j)) = 7$. We will add a column to our table to create the squared deviations from the mean.

Table $2.9.7$: Computation of the squared deviations from the mean

x_j	$P(x_j)$	$(x_j - \mu)^2$
5	$\frac{2}{20} = 10\%$	$(5-7)^2 = 4$
6	$\frac{2}{20} = 30\%$	$(6-7)^2 = 1$
7	$\frac{5}{20} = 25\%$	$(7-7)^2 = 0$
8	$\frac{4}{20} = 20\%$	$(8-7)^2 = 1$
9	$\frac{3}{20} = 15\%$	$(9-7)^2 = 4$
Totals:	$\sum P(x_j) = 100\%$	

Before we can average these squared deviation measures, we must weight these squared deviations by the relative frequency of occurrence to account for the fact that some data values, such as the 6, occur with different relative frequency than others, such as the 9. To do so, we form a new column for the products of $(x_j - \mu)^2 \cdot P(x_j)$ and then proceed to "average" those weighted squared deviations.

Table $2.9.8$: Computation of the variance

x_j	$P(x_j)$	$(x_j - \mu)^2$	$(x_j - \mu)^2 \cdot P(x_j)$
5	$\frac{2}{20} = 10\%$	4	$4 \cdot 0.10 = 0.40$
6	$\frac{2}{20} = 30\%$	1	$1 \cdot 0.30 = 0.30$
7	$\frac{5}{20} = 25\%$	0	$0 \cdot 0.25 = 0.00$
8	$\frac{4}{20} = 20\%$	1	$1 \cdot 0.20 = 0.20$
9	$\frac{3}{20} = 15\%$	4	$4 \cdot 0.15 = 0.60$
Totals:	$\sum P(x_j) = 100\%$		$\sum \left((x_j - \mu)^2 \cdot P(x_j)\right) = 1.50$
		Variance:	$\sigma^2 = \frac{\sum\left[(x_j-\mu)^2 \cdot P(x_j)\right]}{\sum P(x_j)} = \frac{1.50}{1} = 1.5$

As shown in the last two rows, with our weighted squared deviations determined, we can find the average by summing our squared deviations and then dividing by the total weighting of the relative frequency measures, which will always be the value $1.00 = 100\%$. We again get the same measure of variance, 1.5, as we did earlier when the data was in frequency table form.

Variance and Standard Deviation from a Relative Frequency Distribution

$$\text{Variance for population data: } \sigma^2 = \sum \left[(x_j - \mu)^2 \cdot P(x_j)\right]$$

$$\text{Standard Deviation for population data: } \sigma = \sqrt{\sum \left[(x_j - \mu)^2 \cdot P(x_j)\right]} = \sqrt{\sigma^2}$$

We should note that the computation work here is simpler when the grouped data is given in relative frequency rather than just frequency format. This is one reason we often look at data in relative frequency form.

? Text Exercise 2.9.2

We take the frequency table shown below from Section 2.1 regarding thirty student scores (discrete 10-point scale) for an assignment but with the distribution given in relative frequency format instead. This time, we assume the data is population data—our focus is only on these thirty students and not some larger group.

Table 2.9.9: Relative frequency distribution of student scores

x_j	$P(x_j)$
3	$\frac{1}{30} \approx 0.0333 = 3.33\%$
4	$\frac{1}{30} \approx 0.0333 = 3.33\%$
5	$\frac{3}{30} = 0.1000 = 10.00\%$
6	$0.1667 = 16.67\%$
7	$0.1667 = 16.67\%$
8	$0.2333 = 23.33\%$
9	$0.1667 = 16.67\%$
10	$0.1000 = 10.00\%$
Totals:	$\sum P(x_j) = 1.0000 = 100\%$

Find this grouped population data's range, variance, and standard deviation.

Answer

The given table clearly shows that the minimum data value is 3 and the maximum is 10, thus a range measure of 7.

For variance, we must first find the squares on the deviations from the mean, as shown in the added column below. Recall that, in Section 2.8, we found the mean of the relative frequency distribution as $\mu = \sum x_j \cdot P(x_j) \approx 7.2667$. We will use this value in our computation work.

Table 2.9.10 Computation of squared deviations from the mean

x_j	$P(x_j)$	$(x_j - \mu)^2$
3	$0.0333 = 3.33\%$	$(3 - 7.2667)^2 \approx 18.2044$
4	$0.0333 = 3.33\%$	$(4 - 7.2667)^2 \approx 10.6711$
5	$0.1000 = 10.00\%$	$(5 - 7.2667)^2 \approx 5.1378$
6	$0.1667 = 16.67\%$	$(6 - 7.2667)^2 \approx 1.6044$
7	$0.1667 = 16.67\%$	$(7 - 7.2667)^2 \approx 0.0711$
8	$0.2333 = 23.33\%$	$(8 - 7.2667)^2 \approx 0.5378$
9	$0.1667 = 16.67\%$	$(9 - 7.2667)^2 \approx 3.0044$
10	$0.1000 = 10.00\%$	$(10 - 7.2667)^2 \approx 7.4711$
Totals:	$\sum P(x_j) = 1.0000 = 100\%$	

Now we weight, through multiplication, our various squared deviations by their relative frequency of occurrence. This forms the products $(x_j - \mu)^2 \cdot P(x_j)$ in our next added column.

Table 2.9.11: Computation of the variance

x_j	$P(x_j)$	$(x_j - \mu)^2$	$(x_j - \mu)^2 \cdot P(x_j)$
3	$0.0333 = 3.33\%$	18.2044	$18.2044 \cdot 0.0333 \approx 0.6068$
4	$0.0333 = 3.33\%$	10.6711	$10.6711 \cdot 0.0333 \approx 0.3557$
5	$0.1000 = 10.00\%$	5.1378	$5.1378 \cdot 0.1000 \approx 0.5138$
6	$0.1667 = 16.67\%$	1.6044	$1.6044 \cdot 0.1667 \approx 0.2674$
7	$0.1667 = 16.67\%$	0.0711	$0.0711 \cdot 0.1667 \approx 0.0119$
8	$0.2333 = 23.33\%$	0.5378	$0.5378 \cdot 0.2333 \approx 0.1255$
9	$0.1667 = 16.67\%$	3.0044	$3.0044 \cdot 0.1667 \approx 0.5007$

10	$0.1000 = 10.00\%$	7.4711	$7.4711 \cdot 0.1000 \approx 0.7471$
Totals:	$\sum P(x_j) = 1.0000 = 100\%$		
		Variance:	$\sigma^2 = \sum \left[(x_j - \mu)^2 \cdot P(x_j) \right] \approx 3.1289$

Our last row illustrates the production of the variance as the average of the weighted squared deviations. The original table of data has a population variance of $\sigma^2 \approx 3.1289$and population standard deviation of $\sigma = \sqrt{\sigma^2} \approx \sqrt{3.1289} \approx 1.7689$.

We have computed the three desired measures of variation in the data and grouped them within a relative frequency table.

We can compute population variance and standard deviation even when given data in a relative frequency table. However, we cannot do so for sample data given only in a relative frequency table.

Section Summary

This section has demonstrated that we can often compute range, variance, and standard deviation even after the data has been grouped into a frequency table or a relative frequency table. We also remind ourselves that the formulas developed in this section came from the meaning of each measure. We do not memorize the formulas; we recall what each measure means and how and why we performed the computations. This section demonstrated how we can adjust our process to produce the same measures, but it also showed that, due to the "column" computation work, the use of a spreadsheet makes the process much easier.

The following optional section explores how we might estimate the range, variance, and standard deviation if our data has been grouped into interval classes. The ideas are similar, but we can only roughly estimate such measures because we have lost individual data representation (a loss of information about the data).
