2.9.1: Measures of Variance and Standard Deviation - Loss of Information - Optional Material

Learning Objectives

- Consider the loss of information with grouped data
- Discuss class approximations
- Develop methods to approximate the range, variance, and standard deviation from grouped data

Section 2.9.1 Excel File (contains all of the data sets for this section)

Dispersion Measures on Interval-Grouped Data with Loss Of Information

What if we have data grouped over intervals instead of discrete single value groups as previously in Section 2.9? As with our measures of central tendency, we have lost some information about the specific data values and are only able to roughly estimate our dispersion measures of range, variance, and standard deviation measures. We examine common approximation methods below; however, it should be known that the methods shown below are not unique as other similar but different choices can be made in what values are used from the class intervals.

The following table is again the frequency/relative frequency table based on data given by Florence Nightingale in her text *Notes on Nursing* (downloaded here). We will again assume that Ms. Nightingale collected the data in a way such that if, for example, someone was in their 29th year of age (such as 29.875years old), the data was reported as a 29. We will also take this information as a representation of population data in our following computation work.

Table 2.9.1.1: Grouped Frequency Distribution

Age Intervals (years)	Interval Notation (years)	Frequency	Relative Frequency
$20-30$	$[20, 30)$	$1,441$	$\frac{1,441}{25,466} \approx 0.0566 = 5.66\%$
$30-39$	$[30, 40)$	$2,477$	$\frac{2,477}{25,466} \approx 0.0973 = 9.73\%$
$40-49$	$[40, 50)$	$4,971$	$0.1952 = 19.52\%$
$50-59$	$[50, 60)$	$7,438$	$0.2921 = 29.21\%$
$60-69$	$[60, 70)$	$6,367$	$0.2500 = 25.00\%$
$70-79$	$[70, 80)$	$2,314$	$0.0909 = 9.09\%$
$80+$	$[80, \text{above})$	458	$0.0180 = 1.80\%$
Totals:		$25,466$	$1.0000 = 100\%$

We recall that we do not know, for example, how many 20 year old nurses there were in the data set, nor do we know how many 27 year old there were. We only know that there were $1,441$ nurses reporting ages of $20-29$. This clearly implies that we cannot know what the actual data values were in the original data set; we have lost specific information about the original data set.

We will similarly approximate our measures of dispersions based on this grouped data by use of the midpoint value of each interval as our best approximation single measure for all values within the interval. So, again we will assume that all $1,441$ people in their twenties are 25 years old, the midpoint of that interval. This is a drastic assumption in some sense, but with the loss of information on specific age measures in each interval, this is a reasonable way to approximate our dispersion measures just as with our central tendency measures. We will also again use 85 as our value for the last class interval of $[80, \text{above})$ even though the midpoint value could be larger if more was known about the actual data. This is also pointing to why it is not a general best practice when building frequency tables of data to use "and above" or "and below" within the last and first class interval descriptions; doing so provides even greater loss of key information about the data set.

We might choose to estimate the range measure using the largest and smallest midpoint values. That is, we estimate the range to be approximately $85-25=60$ years. We note that we consider this a very rough estimate and major decisions should not be based on this estimate. Likely the range measure is larger, but again due to the loss of information when the data was grouped, we can't know for sure. We also note that others might estimate the range from such grouped data differently (such as the highest class' upper limit value minus the lowest class' lower limit value.)

Now for the variance estimate. First, we also recall from the optional Section 2.8.1, that we computed a mean estimate value of $\mu \approx 54.2675$ years old. Variance, as the average of squared deviations from the mean, then leads to our producing the squared deviations column and weighting of those squared deviations by the relative frequency measures (we choose to use the relative frequency versus frequency approach in our work). Again, we use the midpoint of each interval as the data value to estimate deviation from the mean measures, and complete our work as in the non-interval grouped data approach of Section 2.9.

Table 2.9.1.2 Computation of variance

Age Intervals (years)	Midpoint (m_j) (years)	$P(m_j)$	$(m_j-\mu)^2 \cdot P(m_j)$
$20-29$	25	$0.0566=5.66\%$	$(25-54.2675)^2 \cdot 0.0566 \approx 48.4828$
$30-39$	35	$0.0973=9.73\%$	$(35-54.2675)^2 \cdot 0.0973 \approx 36.1213$
$40-49$	45	$0.1952=19.52\%$	$(45-54.2675)^2 \cdot 0.1952 \approx 16.7651$
$50-59$	55	$0.2921=29.21\%$	0.1567
$60-69$	65	$0.2500=25.00\%$	28.7966
$70-79$	75	$0.0909=9.09\%$	39.0721
$80+$	85	$0.0180=1.80\%$	17.0008
	Totals:	$1.0000=100\%$	$\sum\left[(m_j-\mu)^2 \cdot P(m_j)\right] \approx 186.3954$

So, we would estimate the variance of all this given population of non-domestic servant nurses in Great Britain to be about $\sigma^2 \approx 186.3954$ years^2, and hence the standard deviation to be about $\sigma \approx \sqrt{186.3954} \approx 13.6527$ years. So, with interval-grouped data, we can estimate the variance and standard deviation by the same overall process, described symbolically by the given formulas with the use of each interval's midpoint represented by m_j :

Variance from an Interval-Grouped Distribution

$$s^2 \approx \frac{\sum\left[(m_j-\bar{x})^2 \cdot f_j\right]}{\sum f_j-1} \text{ ; when working with frequency interval-grouped sample data}$$

$$\sigma^2 \approx \frac{\sum\left[(m_j-\mu)^2 \cdot f_j\right]}{\sum f_j} =$$

$$\sum(m_j \cdot P(x_j)) \text{ when working with frequency or relative frequency interval-grouped population data}$$

Standard Deviation from an Interval-Grouped Distribution

$$s \approx \sqrt{s^2} \text{ when working with interval-grouped sample data}$$

$$\sigma^2 \approx \sqrt{\sigma^2} \text{ when working with interval-grouped sample data}$$

? Text Exercise 2.9.1.1

A bakery has been keeping records on the shelf-life of its best selling cinnamon rolls package. The bakery has sent the following frequency table asking for the median and mean measures of the data. Assuming this is sample data, find reasonable estimates of the range, variance, and standard deviation values of the data.

Table 2.9.1.3 Grouped frequency distribution for shelf-life data

Shelf-life (days)	Frequency
$[3,8)$	3
$[8,13)$	19
$[13,18)$	43

Shelf-life (days)	Frequency
$[18, 23)$	21
$[23, 28)$	16
$[3, 8)$	3
$[28, 33)$	2
$[8, 13)$	19
$[13, 18)$	43

Answer

We again proceed by extending our table to include a column of midpoint values. Since this is sample data, we keep with frequency versus relative frequency measures in order to not lose sample size information.

Table 2.9.1.4 Preparatory computations using data from Table 2.9.1.3

Shelf-life (days)	Midpoint (m_j) (days)	Frequency f_j
$[3, 8)$	$\frac{3+8}{2} = 5.5$	3
$[8, 13)$	$\frac{8+13}{2} = 10.5$	19
$[13, 18)$	15.5	43
$[18, 23)$	20.5	21
$[23, 28)$	25.5	16
$[28, 33)$	30.5	2
	Totals:	104

First, we estimate the range to be $30.5 - 5.5 = 25$ days using our midpoint values. (As mentioned in the discussion above, one might instead choose to compute $33 - 3 = 30$ days for the range; this estimate would be considered a maximum amount the range might truly be.)

Next, we estimate the variance. In Section $2.8.1$,we estimated the mean of this data to be 17.1346days. So, to estimate the sample variance, we must form the weighted squared variations from the mean column, then sum those squared variations, and finally divide by one less than the sample size to form our "average" of the squared variations for sample variance purposes.

Table 2.9.1.5 Computation of variance

Shelf-life (days)	Midpoint (m_j) (days)	f_j	$(m_j - \mu)^2 \cdot f_j$
$[3, 8)$	5.5	3	$(5.5 - 17.1346)^2 \cdot 3 \approx 406.0918$
$[8, 13)$	10.5	19	$(10.5 - 17.1346)^2 \cdot 19 \approx 836.3404$
$[13, 18)$	15.5	43	$(15.5 - 17.1346)^2 \cdot 43 \approx 114.8924$
$[18, 23)$	20.5	21	237.8443
$[23, 28)$	25.5	16	1119.6787
$[28, 33)$	30.5	2	357.2678
	Totals:	104	$s^2 \approx \frac{\sum\left[(m_j-\mu)^2 \cdot f_j\right]}{\sum(f_j)-1} \approx \frac{3072.1154}{104-1} \approx 29.$

So, our estimate for the variance on the shelf-life of the packages of cinnamon rolls by this bakery would be about $s^2 \approx 29.8264$ days^2. And thus our standard deviation estimate would be $s \approx \sqrt{29.8264} \approx 5.4614$days. So the cinnamon roll packages roughly tend to last about 17.1 ± 5.5 days.

In summary, we have now seen how we can produce rough estimates for the dispersion measurements when given interval-grouped data. We note how we are really just extending previous ideas/computations. However, we also remind ourselves that the resulting values should be used with caution in the interpretation of the dispersion of the data, not as if the values were the actual true measures of the data.

2.9.1: Measures of Variance and Standard Deviation - Loss of Information - Optional Material is shared under a Public Domain license and was authored, remixed, and/or curated by The Math Department at Fort Hays State University.

- **3.3: Measures of Central Tendency** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.
- **1.10: Distributions** by David Lane is licensed Public Domain. Original source: https://onlinestatbook.com.