2.8.1: Measures of Median and Mean - Grouped Data Loss of Information - Optional Material

Learning Objectives

- Consider the loss of information with grouped data
- Discuss class approximations
- Develop methods to approximate the median and mean from grouped data

Section 2.8.1 Excel File (contains all of the data sets for this section)

Central Measures on Grouped Data with Loss Of Information

What if we have data grouped over intervals instead of discrete single value groups as previously? In this case, we have lost some information about the specific data values and are only able to roughly estimate the mean and median measures of the distribution. Below is a frequency/relative frequency table, Table 2.8.1.2 based on data given by Florence Nightingale in her text *Notes on Nursing* (downloaded here). The text listed the ages of a large sample of non-domestic servant nurses within Great Britain in 1851 in a grouped data interval format. We will assume that Ms. Nightingale collected the data in a way such that if, for example, someone was in their 29^{th} year of age (such as 29.875 years old), the data was reported as a 29 and not rounded up to 30...a common convention in reporting of ages for individuals. We have added the interval notation representation of the continuous variable of age per that convention to the table.

Table 2.8.1.1: 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:		$25,466$	$1.0000 = 100\%$

Notice we do not know how many 20 year old nurses there were in the data set, nor do we know how many 27 year old nurses there were. We only know that there were $1,441$ nurses reporting ages of $20-29$. This means 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 can, however, approximate descriptive statistics based on this grouped data. We will proceed as in the discrete case above, except we will use the midpoint value of each interval as our best approximation single measure for all values within the interval. For example, we will assume that all 1441 people in their twenties are exactly 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 measures. We will also use 85 as our value for the last class interval of $[80, \text{above})$ even though the midpoint value may be larger if more was known about the actual data. It is reasonable to believe that in 1851 most nurses above the age of 80 were likely closer to the 80 value than the 85 value; but this is an assumption we are making and must be disclosed.

These assumptions, across all the intervals, will only give us estimates of the actual true mean and median measures of center. So, for the median, we begin to accumulate our relative frequencies until we know where the 50^{th} percentile measure lies. Since $5.66\% + 9.73\% + 19.52\% = 34.91\%$, which is less than 50%, and $5.66\% + 9.73\% + 19.52\% + 29.21\% = 64.12\%$, which is

greater than 50%, we know the 50^{th} percentile location is within the interval $50-59$. Thus our estimate for the median would be 55 years old.

To estimate the arithmetic mean, we can use the midpoint of each interval as the data value associated with each of the relative frequency measures and complete our computation work as in the discrete case.

Table 2.8.1.2 Computation of mean using data from Table 2.8.1.1

Age Intervals (years)	Midpoint (m_j) (years)	$P(m_j)$	$m_j \cdot P(m_j)$
$20-29$	25	$0.0566 = 5.66\%$	$25 \cdot 0.0566 = 1.4150$
$30-39$	35	$0.0973 = 9.73\%$	$35 \cdot 0.0973 = 3.4055$
$40-49$	45	$0.1952 = 19.52\%$	$45 \cdot 0.1952 = 8.7840$
$50-59$	55	$0.2921 = 29.21\%$	16.0655
$60-69$	65	$0.2500 = 25.00\%$	16.2500
$70-79$	75	$0.0909 = 9.09\%$	6.8175
$80+$	85	$0.0180 = 1.80\%$	1.5300
	Totals:	$1.0000 = 100\%$	$\sum (m_j \cdot P(x_j)) \approx 54.2675$

So, we would estimate the mean age of all these sampled non-domestic servant nurses in Great Britain to be about 54.3 years old. In examining the relative frequency measures as tied to the age intervals, this value makes reasonable sense as the "balance point" of the distribution of the ages. So, in grouped data within intervals, we can estimate the mean by the same overall process, described symbolically by the given formula with the use of each interval's midpoint represented by m_j :

Mean from an Interval-Grouped Distribution

$$\bar{x} \approx \frac{\sum (m_j \cdot f_j)}{\sum f_j} = \sum (m_j \cdot P(m_j)) \text{ when working with interval grouped sample data}$$

$$\mu \approx \frac{\sum (m_j \cdot f_j)}{\sum f_j} = \sum (m_j \cdot P(m_j)) \text{ when working with interval grouped population data}$$

In summary, we have seen how we can still determine estimates for the median and mean measurement when given interval grouped data.

? Text Exercise 2.8.1.1

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. Find reasonable estimates of the mean and the median values of the data.

Table 2.8.1.3 Grouped frequency distribution for shelf-life data

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

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

Answer

We proceed by extending our table to include a column of midpoint values and to compute relative frequency measures. Do note we could also use straight frequency as a weighting measure, but choose to use the relative frequency approach instead.

Table 2.8.1.4 Preparatory computations using data from Table 2.8.1.3

Shelf-life (days)	Midpoint (m_j) (days)	Frequency	Relative Frequency (P \left(m_j \right) \)
$[3, 8)$	$\frac{3+8}{2} = 5.5$	3	$\frac{3}{104} \approx 0.0288$
$[8, 13)$	$\frac{8+13}{2} = 10.5$	19	$\frac{19}{104} \approx 0.1827$
$[13, 18)$	15.5	43	0.4135
$[18, 23)$	20.5	21	0.2019
$[23, 28)$	25.5	16	0.1538
$[28, 33)$	30.5	2	0.0192
	Totals:	104	1.0000

To estimate the median, we again focus on our relative frequency measures to get a "location". We notice that $2.88\% + 18.27\% = 21.15\%$, which is less than 50%, and $2.88\% + 18.27\% + 41.35\% = 62.50\%$, which is greater than 50%. The 50^{th} percentile location is within the interval $[13, 18)$. Thus our estimate for the median shelf-life of the packages of cinnamon rolls by this bakery would be 15.5 days.

Next, we weight each midpoint value by its corresponding relative frequency measure, before summing to produce our mean measure.

Table 2.8.1.5 Computation of mean shelf-life

Shelf-life (days)	Midpoint (m_j) (days)	$P(m_j)$	$m_j \cdot P(m_j)$
$[3, 8)$	5.5	0.0288	$5.5 \cdot 0.0288 \approx 0.1587$
$[8, 13)$	10.5	0.1827	$10.5 \cdot 0.1827 \approx 1.9183$
$[13, 18)$	15.5	0.4135	6.4087
$[18, 23)$	20.5	0.2019	4.1304
$[23, 28)$	25.5	0.1538	3.9231
$[28, 33)$	30.5	0.0192	0.5865
	Totals:	1.0000	17.1346

So, our estimate for the mean shelf-life of the packages of cinnamon rolls by this bakery would be about 17.1 days.

In summary, we have seen how we can determine estimates for the median and mean measurement when given interval-grouped data, but also heed the warning that these are just rough estimates and that we must not considered our results as the actual measures for the data that was originally collected.