Probability and Statistics

Topic 3 - Measure of Central Tendency and Dispersion from Grouped Data

Aamir Alaud Din, PhD

October 09, 2023

TABLE OF CONTENTS

1. RECAP
2. OBJECTIVES
3. THE WHY SECTION
4. MEAN OF GROUPED DATA
5. WEIGHTED MEAN
6. STANDARD DEVIATION FROM GROUPED DATA
7. SUMMARY

TABLE OF CONTENTS

1. RECAP
2. OBJECTIVES
3. THE WHY SECTION
4. MEAN OF GROUPED DATA
5. WEIGHTED MEAN
6. STANDARD DEVIATION FROM GROUPED DATA
7. SUMMARY

RECAP

- Measure of dispersion is the spread of the data about mean.
- Range, standard deviation, variance, and interquartile range are four popular the measures of dispersion.
- Range of a variable is the difference between the largest and smallest data value.
- Range is not resistant to extreme values in the data set.
- Population standard deviation is the square root of the sum of squared deviations about the population mean divided by the number of observations in the population.
- The conceptual formula to compute population standard deviation is $\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{N}}$.

RECAP

- The computational formula to compute population standard deviation is $\sigma = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N}}$.
- The conceptual and computational formulas give the same answer.
- The conceptual formula to compute sample standard deviation is $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$.
- The computational formula to compute sample standard deviation is $s = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}}$.
- The denominator in the square root of sample standard deviation is 1 less than the number of samples *n i.e.*, $n - 1$, known as the degrees of freedom, captures the variability in the population.

RECAP

- According to Empirical Rule to describe bell shaped (normal) data, 68%, 95%, and 99.7% of the data lie within 1^{st}, 2^{nd}, and 3^{rd} standard deviations about the mean, respectively.
- Chebyshev's Inequality is used to describe any data set.
- According to Chebyshev's Inequality, at least $\left(1 - \frac{1}{k^2}\right).100\%$ of the observations lie within k standard deviations of the mean, where k is any number greater than 1.
- Chebyshev's Inequality is applicable to skewed data.
- Student may write their own codes for Chebyshev's Inequality if not already given in the built-in statistical module of python (optional).

TABLE OF CONTENTS

1. RECAP
2. OBJECTIVES
3. THE WHY SECTION
4. MEAN OF GROUPED DATA
5. WEIGHTED MEAN
6. STANDARD DEVIATION FROM GROUPED DATA
7. SUMMARY

OBJECTIVES

After learning this topic and studying, you should be able to:

1. Approximate the mean of a variable from grouped data
2. Compute the weighted mean
3. Approximate the standard deviation of a variable from grouped data

TABLE OF CONTENTS

1. RECAP
2. OBJECTIVES
3. THE WHY SECTION
4. MEAN OF GROUPED DATA
5. WEIGHTED MEAN
6. STANDARD DEVIATION FROM GROUPED DATA
7. SUMMARY

THE WHY SECTION

- Grouped data is not a single number.
- It tells us how many and which data points belong to a specific group of numbers.
- For example, the numbers 8.3, 8.35, 8.95, and 8.99 are four data points in the complete data set which belong to the class 8 - 8.99.
- The reason of computing the mean and standard deviation are to find the single number which represents the complete data set and to know the spread of the data set about mean which we already studied in the previous topics (1 and 2).
- The formulas to find mean and standard deviation are little bit different from the ones we already studied, however, the concept is exactly the same.
- Moreover, the difference in formulas is just symbolic but the conceptually and numerically they are the same.

THE WHY SECTION

- The objective of this topic is to use these formulas, understand them in a way such that there appears no difference in these and already used formulas and compute the mean and standard deviation.
- We will also see that the formulas make the computations easier and simpler.
- We will also see that statistical measures are not deterministic.
- Often the only data available have already been summarized in frequency distributions (grouped data).
- Although we cannot find exact values of the mean or standard deviation without raw data, we can approximate these measures.

TABLE OF CONTENTS

1. RECAP
2. OBJECTIVES
3. THE WHY SECTION
4. MEAN OF GROUPED DATA
5. WEIGHTED MEAN
6. STANDARD DEVIATION FROM GROUPED DATA
7. SUMMARY

MEAN OF GROUPED DATA

- Because raw data cannot be retrieved from a frequency table, we assume that within each class the mean of the data values is equal to the class midpoint.
- We then multiply the class midpoint by the frequency.
- This product is expected to be close to the sum of the data that lie within the class.
- We repeat the process for each class and add the results.
- This sum approximates the sum of all the data.
- Let's look at the formulas of sample and population mean.
- We also dig out to know no difference in these and already used formulas.

MEAN OF GROUPED DATA

Formulas

Sample Mean

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum x_i f_i}{\sum f_i}$$

Population Mean

$$\mu = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum x_i f_i}{\sum f_i}$$

where where x_i is the midpoint or value of the ith class
f_i is the frequency of the ith class
n is the number of classes

MEAN OF GROUPED DATA

- In formulas for sample and population mean, $x_1 f_1$ *approximates* the sum of all the data values in the first class, $x_2 f_2$ *approximates* the sum of all the data values in the second class, and so on
- Notice that the formulas for the population mean and sample mean are essentially identical, just as they were for computing the mean from raw data.
- We already know how to define the number of classes and the classes themselves (frequency distributions and histograms).
- We just review how to find the middle point of each class.
- Consider two classes 8 - 8.99 and 9 - 9.99 and for the first class, the mid point is $(8 + 9)/2 = 8.5$ and so on.

MEAN OF GROUPED DATA

EXAMPLE 1

The frequency distribution in Table 13 represents the five-year rate of return of a random sample of 40 large-blended mutual funds. Approximate the mean fiveyear rate of return.

Table 13

Class (five-year rate of return)	Frequency
8–8.99	2
9–9.99	2
10–10.99	4
11–11.99	1
12–12.99	6
13–13.99	13
14–14.99	7
15–15.99	3
16–16.99	1
17–17.99	0
18–18.99	0
19–19.99	1

TABLE OF CONTENTS

1. RECAP
2. OBJECTIVES
3. THE WHY SECTION
4. MEAN OF GROUPED DATA
5. WEIGHTED MEAN
6. STANDARD DEVIATION FROM GROUPED DATA
7. SUMMARY

WEIGHTED MEAN

- When data values have different importance, or weight, associated with them, we compute the weighted mean.
- For example, your grade-point average is a weighted mean, with the weights equal to the number of credit hours in each course.
- The value of the variable is equal to the grade converted to a point value.

Weighted Mean

The weighted mean, $\bar{x}_w$, of a variable is found by multiplying each value of the variable by its corresponding weight, adding these products, and dividing this sum by the sum of the weights. It can be expressed using the formula

$$\bar{x}_w = \frac{x_1 w_1 + x_2 w_2 + \cdots + x_n w_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum x_i w_i}{\sum w_i}$$

WEIGHTED MEAN

Weighted Mean

where w_i is the weight of the ith observation
x_i is the value of the ith observation

EXAMPLE 2

Marissa just completed her first semester in college. She earned an A in her four-hour statistics course, a B in her three-hour sociology course, an A in her threehour psychology course, a C in her five-hour computer programming course, and an A in her one-hour drama course. Determine Marissa's grade-point average.

TABLE OF CONTENTS

1. RECAP
2. OBJECTIVES
3. THE WHY SECTION
4. MEAN OF GROUPED DATA
5. WEIGHTED MEAN
6. STANDARD DEVIATION FROM GROUPED DATA
7. SUMMARY

STANDARD DEVIATION FROM GROUPED DATA

- The procedure for approximating the standard deviation from grouped data is similar to that of finding the mean from grouped data.
- Again, because we do not have access to the original data, the standard deviation is approximate.

Sample and Population Standard Deviation

Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{(\sum f_i) - 1}}$$

STANDARD DEVIATION FROM GROUPED DATA

Sample and Population Standard Deviation

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2 f_i}{\sum f_i}}$$

where x_i is the midpoint or value of the ith class
f_i is the frequency of the ith class

Computational Formulas

$$s = \sqrt{\frac{\sum x_i^2 f_i - \frac{(x_i f_i)^2}{\sum f_i}}{(\sum f_i) - 1}} \quad \text{and} \quad \sigma = \sqrt{\frac{\sum x_i^2 f_i - \frac{(x_i f_i)^2}{\sum f_i}}{\sum f_i}}$$

STANDARD DEVIATION FROM GROUPED DATA

EXAMPLE 3

The data in Table 13 on page 176 represent the five-year rate of return of a random sample of 40 large-blended mutual funds. Approximate the standard deviation of the five-year rate of return.

Table 13

Class (five-year rate of return)	Frequency
8–8.99	2
9–9.99	2
10–10.99	4
11–11.99	1
12–12.99	6
13–13.99	13
14–14.99	7
15–15.99	3
16–16.99	1
17–17.99	0
18–18.99	0
19–19.99	1

TABLE OF CONTENTS

1. RECAP
2. OBJECTIVES
3. THE WHY SECTION
4. MEAN OF GROUPED DATA
5. WEIGHTED MEAN
6. STANDARD DEVIATION FROM GROUPED DATA
7. SUMMARY

SUMMARY

- The meaning of the measure of central tendency and measure of dispersion for grouped and ungrouped data are the same.
- To determine classes, there is no exact rule and different classes with varying class widths must be tested.
- The formulas for sample and population standard deviation for grouped and ungrouped data are symbolically different but exactly logically exactly same as for the ungrouped data.
- The estimated are approximations and not the deterministic values, after all, statistics is all about the logics.
- To estimate the weighted mean, the weights must be determined carefully after conduction of several logical meetings.


Thank You!