Weighted means and grouped data

Remember earlier that we learned to calculate the means of populations and samples as

Population mean: $$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Sample mean: $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

But what we didn’t say before is that both of these formulas for the mean give equal weight to every member of the population (or sample).

That’s not necessarily a bad thing, but it’s not always appropriate to give equal weighting to every member. Sometimes, calculating the mean will only make sense if we give different weights to different members.

Weighted means

Imagine that a cruise line wants to determine a mean number of pieces of luggage per passenger on a cruise ship, but the data they have is given by room. The table below shows the number of pieces of luggage per passenger in each state room in the sample, as well as the number of passengers in those same rooms.

Room	1	2	3	4	5
Bags per passenger	1.50	0.75	1.33	1.50	1.00
Passengers	2	4	3	8	4

If the cruise line wants to calculate mean bags per passenger, they can't simply take the mean of the values in the "Bags per passenger" row of the table. In other words, calculating the sample mean as

$$\bar{x} = \frac{1.50 + 0.75 + 1.33 + 1.50 + 1.00}{5} = \frac{6.08}{5} = 1.2160 \text{ bags per passenger}$$

gives an incorrect value for mean number of bags per passenger.

That's because the "Bags per passenger" values in the table don't hold equal weight. The 8 passengers in state room 4 each brought 1.5 bags, on average, and that should count more heavily toward the overall mean than the 1.5 bags brought by each passenger in state room 1.

The mean number of bags per passenger should instead be calculated as a weighted mean,

Weighted population mean: $$\mu = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

Weighted sample mean: $$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where x_i is the ith observation in the set, and w_i is its corresponding weight. So the mean number of bags per passenger in this sample is

$$\bar{x} = \frac{2(1.50) + 4(0.75) + 3(1.33) + 8(1.50) + 4(1.00)}{2 + 4 + 3 + 8 + 4}$$

$$\bar{x} = \frac{3 + 3 + 4 + 12 + 4}{2 + 4 + 3 + 8 + 4}$$

$$\bar{x} = \frac{26}{21}$$

$$\bar{x} \approx 1.2381$$

Notice how the actual mean is different than the mean we calculated earlier when we didn't accurately weight the data.

Grouped data

In the same way that the calculation of a sample mean or population mean is different when the data is unequally weighted, the calculation of mean, variance, and standard deviation will also be different for grouped data.

When we're given data as a frequency table where the data has already been grouped into buckets, or classes, then we'll use the weighted mean formula to estimate the mean of the data set. Keep in mind that we can't calculate an exact mean, because we don't have the raw data, only the frequency table.

We'll use the midpoint of each class M_i as the value of all the items in the class, and we'll use the frequency of each class f_i as the weight. Given a sample size n, or a population size N, the estimate of the mean for grouped data is

Grouped data population mean: $$\mu = \frac{\sum_{i=1}^{N} f_i M_i}{N}$$

Grouped data sample mean: $$\bar{x} = \frac{\sum_{i=1}^{n} f_i M_i}{n}$$

Once we've estimated the mean, we can use it to approximate the sample variance for the grouped data.

Grouped data population variance: $$\sigma^2 = \frac{\sum_{i=1}^{N} f_i(M_i - \mu)^2}{N}$$

Grouped data sample variance: $$s^2 = \frac{\sum_{i=1}^{n} f_i(M_i - \bar{x})^2}{n-1}$$

And then the approximation of the standard deviation of the grouped data set is simply the square root of the approximate variance.

Let's do an example where we calculate the mean, variance, and standard deviation for a sample of grouped data.

Example

The frequency distribution of the ages of a sample of children enrolled a youth soccer program are given in the table. Give an estimate for the mean age, then find the variance and standard deviation.

Age	Frequency
4 - 6	2
7 - 9	5
10 - 12	7
13 - 15	8
16 - 18	3

We'll need the midpoint of each class to calculate the mean of the grouped sample data, so let's add that column to the table.

Age	Midpoint	Frequency
4 - 6	5	2
7 - 9	8	5
10 - 12	11	7
13 - 15	14	8
16 - 18	17	3

Then the estimate of the sample mean is

$$\bar{x}=\frac{\sum_{i=1}^{n} f_iM_i}{n}$$

$$\bar{x}=\frac{(2)(5)+(5)(8)+(7)(11)+(8)(14)+(3)(17)}{2+5+7+8+3}$$

$$\bar{x}=\frac{10+40+77+112+51}{25}$$

$$\bar{x}=\frac{290}{25}$$

$$\bar{x}=11.6$$

We can use this mean to estimate the variance of the sample,

$$s^2=\frac{\sum_{i=1}^{n} f_i(M_i-\bar{x})^2}{n-1}$$

$$s^2=\frac{2(5-11.6)^2+5(8-11.6)^2+7(11-11.6)^2+8(14-11.6)^2+3(17-11.6)^2}{25-1}$$

$$s^2 = \frac{2(-6.6)^2 + 5(-3.6)^2 + 7(-0.6)^2 + 8(2.4)^2 + 3(5.4)^2}{24}$$

$$s^2 = \frac{2(43.56) + 5(12.96) + 7(0.36) + 8(5.76) + 3(29.16)}{24}$$

$$s^2 = \frac{87.12 + 64.8 + 2.52 + 46.08 + 87.48}{24}$$

$$s^2 = \frac{288}{24}$$

$$s^2 = 12$$

The standard deviation of the sample will be the square root of the variance.

$$s = \sqrt{12}$$

$$s = 2\sqrt{3}$$