Measure of Variation

The central tendency gives the central value of the data. However, the spread and variability of a dataset is equally important in statistics.

Range

Range is the difference between highest and the lowest value in the dataset. It is denoted by R.

R = higest value - lowest value

For Example:

A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown. Find the mean of each group.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

```
import statistics # importing the module
brand_A = [10, 60, 50, 30, 40, 20] # creating list for brand A
brand_B = [35, 45, 30, 35, 40, 25] # creating list for brand B

# calculating the mean
mean_A = statistics.mean(brand_A)
mean_B = statistics.mean(brand_B)

# displaying the result
print('The mean of Brand A is', mean_A, '\nThe mean of Brand B is', mean_B)

# calculating the range
range_A = max(brand_A) - min(brand_A)
range_B = max(brand_B) - min(brand_B)

# displaying the result
print('\nThe range of Brand A is', range_A, '\nThe range of Brand B is',
range_B)

```

```
The mean of Brand A is 35
The mean of Brand B is 35

The range of Brand A is 50
The range of Brand B is 20
```

The above example gives the importance of range. Even though the mean of both the brand is same, the range is significantly different. The outliers in the data can affect the range.

Question 1

The salaries for the staff of the XYZ Manufacturing Co. are shown here. Find the range.

Staff	Salary ($)
Owner	100,000
Manager	40,000
Sales representative	30,000
Workers	25,000
	15,000
	18,000

```
# Write your code here
# create a list
# calculate the range
# display the result
# Hint: Answer 85,000
```

Population Variance and Standard Deviation

Data variation is the difference of the data from its mean. The population variance is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 . The formula for the population variance is:

$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$

where,

X = individual value

μ = population mean

N = population size

The population standard deviation is the square root of the variance. The symbol for the population standard deviation is σ. Mathematically, population standard deviation is:

$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$

For Example:

Find the population variation and standard variation of the data:

10, 60, 50, 40, 20, 30

```
#Importing modules
import statistics

data1 = [10, 60, 50, 40, 20, 30] # data in list
data_mean = statistics.mean(data1) # calculating mean
substracted_value = [] # empty list

# substracting data with mean
for i in data1:
    x = i - data_mean
    substracted_value.append(x)

# empty list
squared = []

# squaring all the values
for j in substracted_value:
    y = j ** 2
    squared.append(y)

# summation of all the elements after squaring
summed = sum(squared)

# Calculating the population variance
pop_variance = summed/len(data1)

# displaying the value
print('The population variance is', round(pop_variance, 2)) # built-in 'round' funtion to display only the required digits after the dec
```

```
The population variance is 291.67
```

You might have noticed, this is the tidious work. So, Python has built-in function to calculate population variance. You can use this for further questions.

```
import statistics

data2 = [10, 60, 50, 40, 20, 30] # data set
variance = statistics.pvariance(data2) # calculating the population variance
print('The population variance is', round(variance,2))
```

```
6
7 std_dev = statistics.pstdev(data2) # calculating the standard deviation
8 print('\nThe standard deviation is', round(std_dev,2))
```

```
The population variance is 291.67

The standard deviation is 17.08
```

Question 2

Find the population variation and standard deviation of the data from Brand B: 25, 30, 35, 35, 40, 45

```
1 # import module
2 # calculate population variance
3 # calculate the standard deviation
4 # display the result
```

Sample Variance and Standard Deviation

From the above formula, the sample variance could be calculated as:

$s^2 = \frac{\Sigma(X-\bar{X})^2}{n}$

where,

$s^2: \quad sample\ variance$
$X: \quad individial\ value$
$\bar{X}: \quad sample\ mean$
$n: \quad sample\ size$

However, this fomula does not give best estimation of sample variance when the poplation is large and the sample is small (less than 30). Thus, the following formula is used for sample.

$s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$

and,

$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$

where,

$s^2: \quad sample\ variance$
$s: \quad sample\ standard\ deviation$
$X: \quad individial\ value$
$\bar{X}: \quad sample\ mean$
$n: \quad sample\ size$

The above formula looks more tedious and time consuming. There is one shortcut formula for which does not require to calculate sample mean.

Sample Variance

$s^2 = \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}$

Sample Standard Deviation

$s = \sqrt{\frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}}$

Note: ΣX^2 and $(\Sigma X)^2$ are not same. The former is the sum after squaring individual values, but later is the sum of individual values before squaring.

You might be thinking, how can I incorporate these all formula in Python to calculate sample variance and standard deviation? For this reason, Python has built-in function.
For example:
The number of public school teacher strikes in Pennsylvania for a random sample of school years in shown. Find the sample variance and

sample standard deviation.
9, 10, 14, 7, 8, 3

```
import statistics # importing the module
data3 = [9, 10, 14, 7, 8, 3] # creating data list

# using built-in funtion
variance = statistics.variance(data3)
deviation = statistics.stdev(data3)

# displaying the result
print(f"The sample variance is {round(variance,2)} and the sample standard deviation is {round(deviation,2)}.")
```

```
The sample variance is 13.1 and the sample standard deviation is 3.62.
```

Question 3

The average weekly unemployment benefits (in dollars) for a random selection of states are listed below. Calculate the sample variance and standard deviation for the data.
239, 214, 327, 416, 321, 289, 209, 356, 190, 272, 252, 272, 310, 276, 251

```
# import module
# calculate population variance
# calculate the standard deviation
# display the result
```

Coefficient of Variation

Two samples can be compared with variance and standard deviation. But if one has to compare difference between two different units, coefficient of variation is needed.

Coefficient of Variation is the ratio of standard deviation and its mean. It is expressed in percentage.

For samples,

$CVar = \frac{s}{\bar{X}} \, . \, 100$

For populations,

$CVar \, = \, \frac{\sigma}{\mu} \, . \, 100$

For example:

The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is 5225 USD, and the standard deviation is 773 USD. Compare the variations of the two.

```
# given data
mean_sales = 87
sales_std_dev = 5
comm_mean = 5225
comm_std_dev = 773

# calculating coefficient of variation
sales_CVar = (sales_std_dev / mean_sales) * 100
comm_CVar = (comm_std_dev / comm_mean) * 100

# displaying the result
print(f"The coefficient of variance for the sales data is {sales_CVar:.2f}%\n")
print(f"The coefficient of variance for the commissions data is {comm_CVar:.2f}%\n")
print(f"Since, coefficient of variation of commissions {comm_CVar:.2f}% is larger than the sales {sales_CVar:.2f}%, thus, the commisions
```

```
The coefficient of variance for the sales data is 5.75%

The coefficient of variance for the commissions data is 14.79%

Since, coefficient of variation of commissions 14.79% is larger than the sales 5.75%, thus, the commisions are more variable than the sa
```

Question 4

The mean for the number of pages of a sample of women's fitness magazines is 132, with a variance of 32; the mean for the number of advertisements of a sample of women's fitness magazines is 182, with a variance of 62. Compare the variations.

```
# provide suitable variable for the given data
# calculate the coefficient of variations
# display your results
```