Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 80%. We keep your highest score.

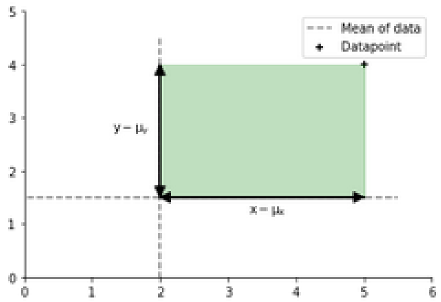
1. In this question we will look at a two-dimensional dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ with N samples. Each sample $\mathbf{x}_i$ in the dataset is a two-dimensional vector with coordinates x, y, i.e., the first component of the vector is denoted by x and the other one by y.

1 / 1 point

The covariance between two *scalar* random variables is

$$\mathrm{cov}[x, y] = E[(\mathrm{x} - \mu_{\mathrm{x}})(\mathrm{y} - \mu_{\mathrm{y}})] \approx \frac{1}{N}\sum_{i=1}^N (\mathrm{x} - \mu_{\mathrm{x}})(\mathrm{y} - \mu_{\mathrm{y}}).$$

In the formula for covariance, we can think of each individual multiplication as the calculation of an area, a rectangle with sides $\mathrm{x} - \mu_{\mathrm{x}}$ and $\mathrm{y} - \mu_{\mathrm{y}}$.



For this datapoint, an increase in x from the mean is linked to an increase in y. Where $\mathrm{x} - \mu_{\mathrm{x}}$ and $\mathrm{y} - \mu_{\mathrm{y}}$ have the same sign, the contribution to the covariance is positive and in green, while if the signs are opposite it will be negative and in red. In other words, green means that x and y are positively correlated, while red means they're negatively correlated.

The total sum of areas, divided by the number of points n, will be the value of the covariance.

Run the code once to see this, then uncomment the line that will show the rectangles and run again.

```
1   # RUN THE CODE ONCE, THEN UNCOMMENT LINE 29 TO VISUALISE COVARIANCE
2
3   fig, ax = plt.subplots()
4
5   #Choose an array by deleting the # in front of the word "data" below.
6   #To switch, put the # back and delete another one
7
8   #Random:
9   data = np.array([[1,2],[5,4],[-2,-3],[4,-2],[2,3],[8,-9]])
10
11  #Straight line:
12  #data = np.array([[1,1],[-3,-3],[2,2],[7,7]])
13
14  #Q1: square
15  data = np.array([[0,0],[4,4],[0,4],[4,0]])
16
17  #Feel free to input your own array or modify the ones above!
18
19  # First calculate the mean with NumPy function np.mean().
20  # The first argument is the dataset and "axis" specifies the direction
21  # Variance in 1D can be calculated similarly with np.var()
22  mean_data = np.mean(data, axis=0)
23  create_plot(data) #which also adds 1d variances
24
25  area=0
26  mean = mean_data
27
28  for i in range(len(data)):
29  #     show_rectangle(mean, data[i])
30      # and a calculation that adds (or subtracts)
31      # the value of the area to our value of the covariance:
32      area += calculate_area(mean, data[i])
33
34  plt.show()
```

Run

Reset

The dashed lines meet at the mean of the dataset. The blue lines represent the magnitude of the variance of the x (horizontal) and y (vertical) components of the dataset.

If red and green balance out, the covariance will be 0. Otherwise the sign of the covariance will give a direction in which the points appear to correlate.

What is $\mathrm{cov}(x, y)$ for the dataset in the array labelled "Q1: square"? Is it what you would expect from the plot?

0.0

Correct

Correct! Since the points are evenly distributed around the mean they balance out and there is no way to determine a direction of correlation.

2. The covariance matrix is given by

1 / 1 point

$$\begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) \\ \text{cov}(y,x) & \text{cov}(y,y) \end{bmatrix} = \begin{bmatrix} \text{var}(x) & \text{cov}(x,y) \\ \text{cov}(y,x) & \text{var}(y) \end{bmatrix}$$

Compute the covariance matrix for the following dataset

$$\mathcal{D} = \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 5 \\ 4 \end{bmatrix} \right\}$$

Here, every column vector represents a data point.

Do the exercise using pen and paper. You can check if your answer makes sense with this codeblock.

```
data = np.array([[1,2],[5,4]])

mean_data = np.mean(data, axis=0)
create_plot(data)

area=0
mean = mean_data

for i in range(len(data)):
    show_rectangle(mean, data[i])
    area += calculate_area(mean, data[i])

plt.show()
```

Run

Reset

- ○ $\begin{bmatrix} 2 & 2 \\ 4 & 1 \end{bmatrix}$
- ○ $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$
- ◉ $\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$

Correct

3. Consider a data set $\mathcal{D}$ with covariance matrix $\begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$.

1 / 1 point

What is the covariance matrix if we multiply every vector in $\mathcal{D}$ by 2?

Run the codeblock below to observe what happens to the example in Question 2 (this will NOT give you the answer to this question but might aid intuition).

```
data = np.array([[1,2],[5,4]])
#data *= 2
#Uncomment the line above to multiply by 2 and run again

mean_data = np.mean(data, axis=0)
create_plot(data)

area=0
mean = mean_data

for i in range(len(data)):
    show_rectangle(mean, data[i])
    area += calculate_area(mean, data[i])

plt.show()
```

Run

Reset

- ◉ $\begin{bmatrix} 12 & 8 \\ 8 & 16 \end{bmatrix}$
- ○ $\begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$
- ○ $\begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$
- ○ $\begin{bmatrix} 16 & 8 \\ 8 & 12 \end{bmatrix}$

Correct

Yes, every element in the covariance matrix is multiplied by 4.

4. Consider the data set $\mathcal{D} = \left\{ \begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}7\\4\end{bmatrix} \right\}$ with covariance matrix $\begin{bmatrix}9 & 3\\3 & 1\end{bmatrix}$.

1 / 1 point

Compute the new covariance matrix when we add $\begin{bmatrix}2\\2\end{bmatrix}$ to each element in $\mathcal{D}$.

Run the codeblock below to observe what happens to the example in Question 2 when a vector is added to every point (this will NOT give you the answer but might aid intuition).

```
data = np.array([[1,2],[5,4]])
#data += [2,2]
#Uncomment line above after first run to add [2,2], then run again

mean_data = np.mean(data, axis=0)
create_plot(data)

area=0
mean = mean_data

for i in range(len(data)):
    show_rectangle(mean, data[i])
    area += calculate_area(mean, data[i])

plt.show()
```

Run

Reset

- $\begin{bmatrix}11 & 5\\5 & 3\end{bmatrix}$
- $\begin{bmatrix}9 & 3\\3 & 1\end{bmatrix}$
- $\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$

Correct
Well done. The covariance will not change.

5. We are looking at a data set $\mathcal{D}$ where every element in $\mathcal{D}$ consists of an x and y coordinate. The data covariance matrix is given by

1 / 1 point

$$\begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$$

Which of the following statements is correct?

- x and y are positively correlated, i.e., when x increases then y increases on average, and vice versa.
- x and y are negatively correlated, i.e., when x increases then y decreases on average, and vice versa.
- x and y are uncorrelated, i.e., when x increases then y does not change on average (and vice versa).

Correct
Well done!