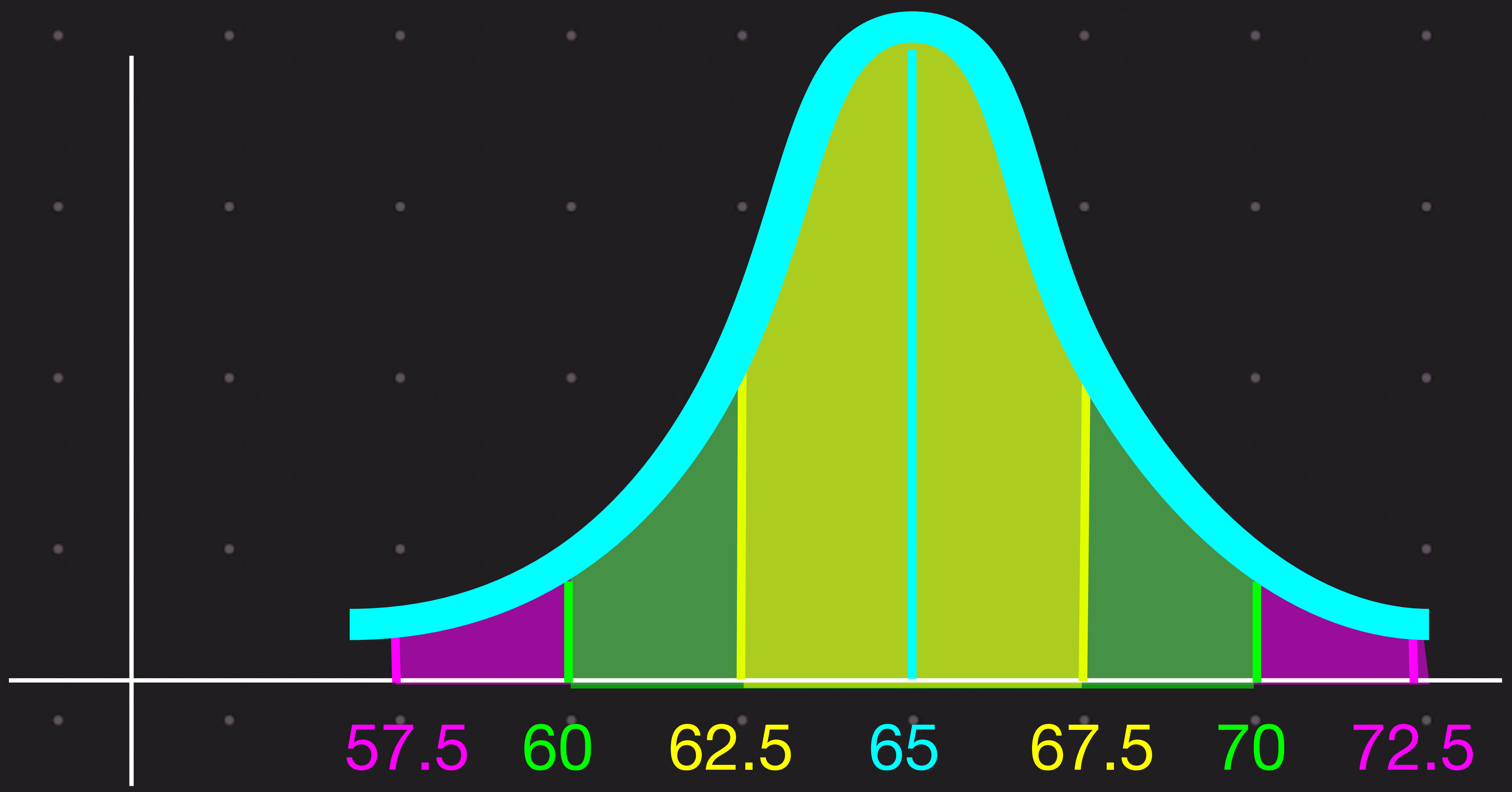
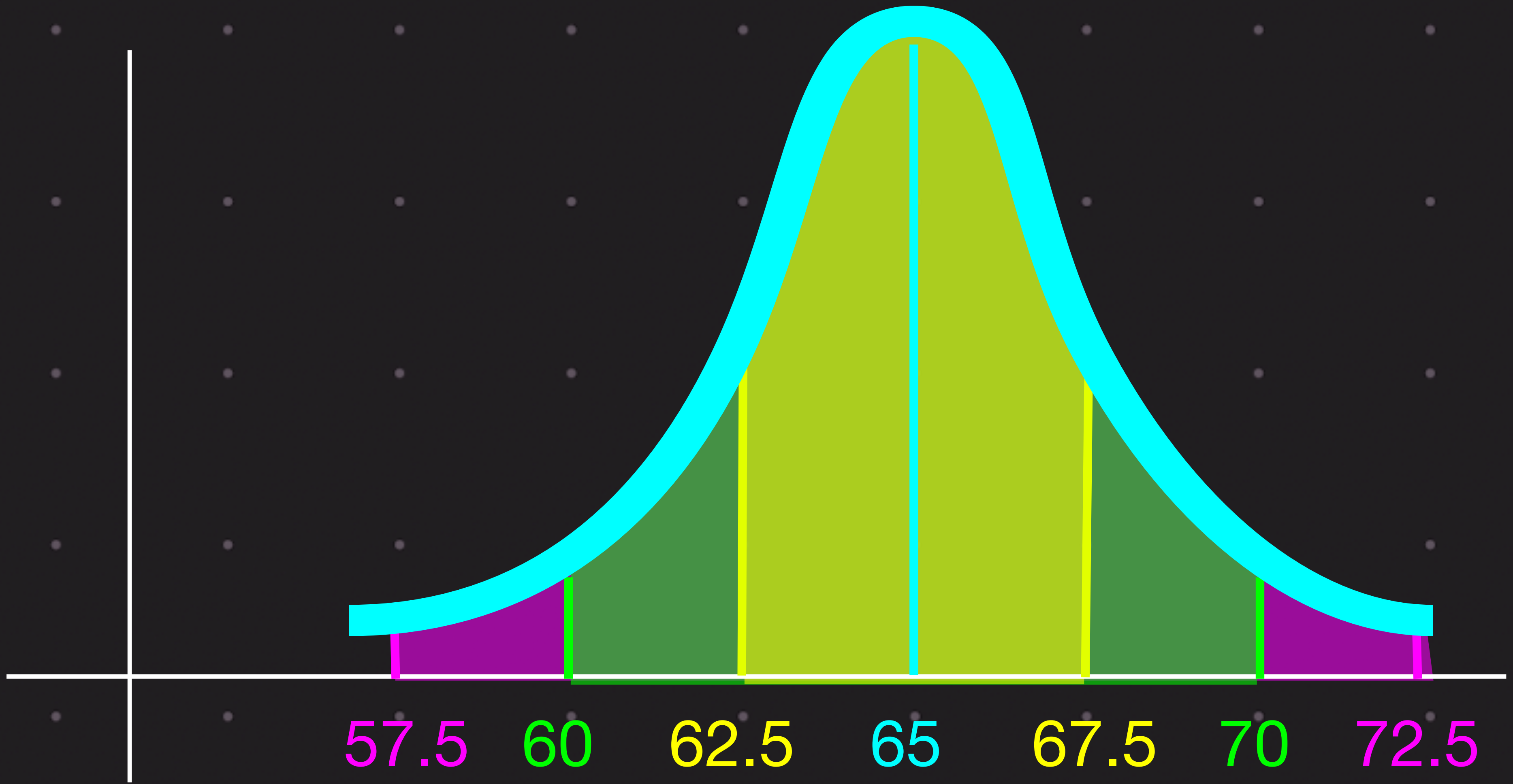
Recap: Heights example. Mean is 65, std dev is 2.5

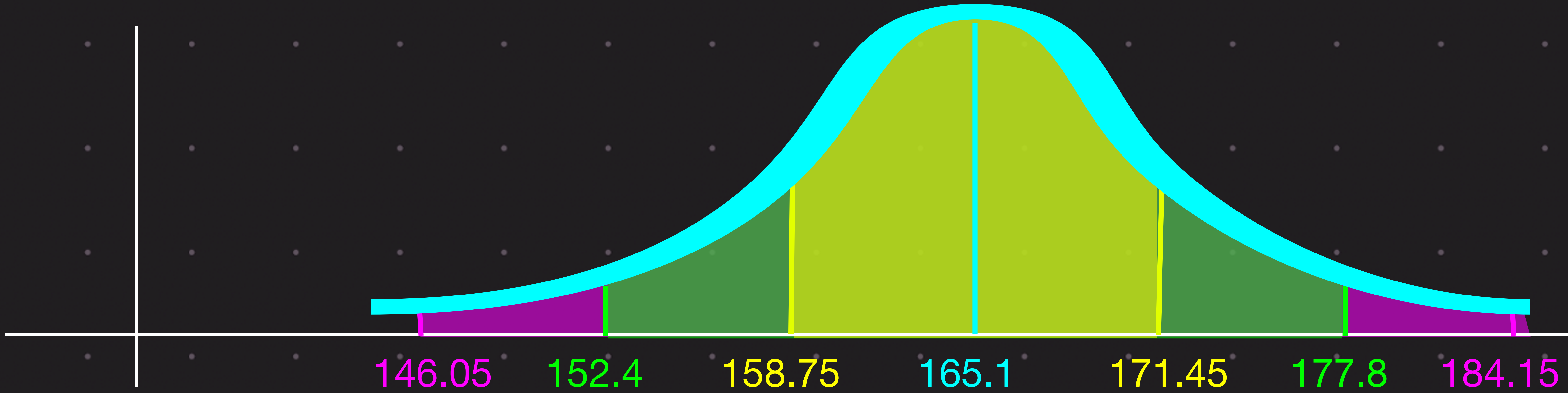
How would this look like if we look at heights in centimetres?



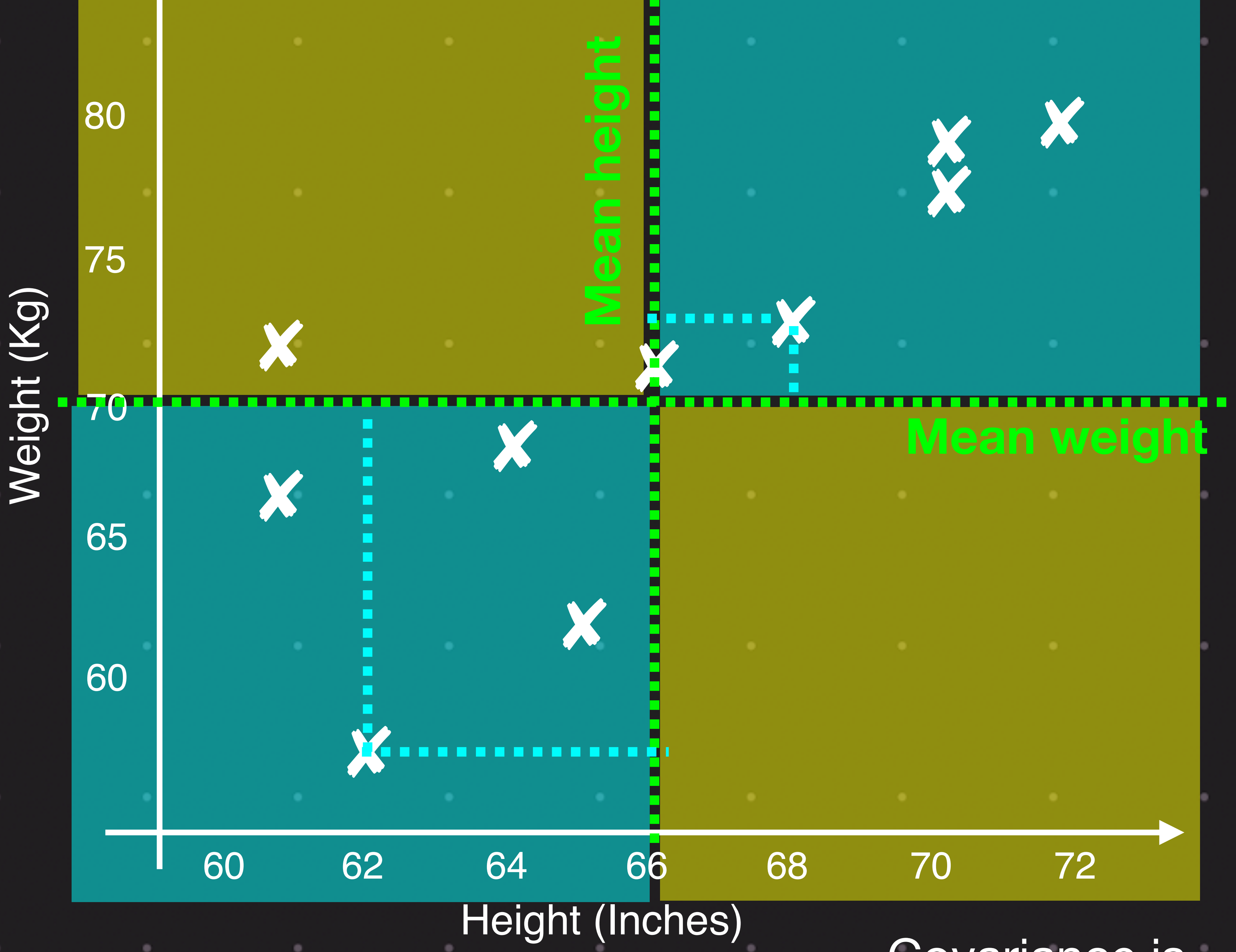
Recap: Heights example. Mean is 65, std dev is 2.5

How would this look like if we look at heights in centimetres?





Height (inches)	Weight (kg)
68	72
62	58
64	67
61	72
70	79
66	61
61	68
65	64
71	80
72	79
$\bar{h} = 66$	$\bar{w} = 70$



Positive correlation

- Top right
- Bottom left

Negative correlation

- Top left
- Bottom right

$$\text{cov}(h, w) = \frac{1}{n} \sum_i (h_i - \bar{h})(w_i - \bar{w})$$

$$(68 - 66)(72 - 70) = 2 * 2 = 4$$
$$(62 - 66)(58 - 70) = (-4) * (-12) = 48$$
$$(64 - 66)(67 - 70) = (-2) * (-3) = 6$$
$$(61 - 66)(72 - 70) = (-1)(2) = -2$$
$$\vdots$$
$$(72 - 66)(80 - 70) = (6)(10) = 60$$

Covariance is the average of all these numbers

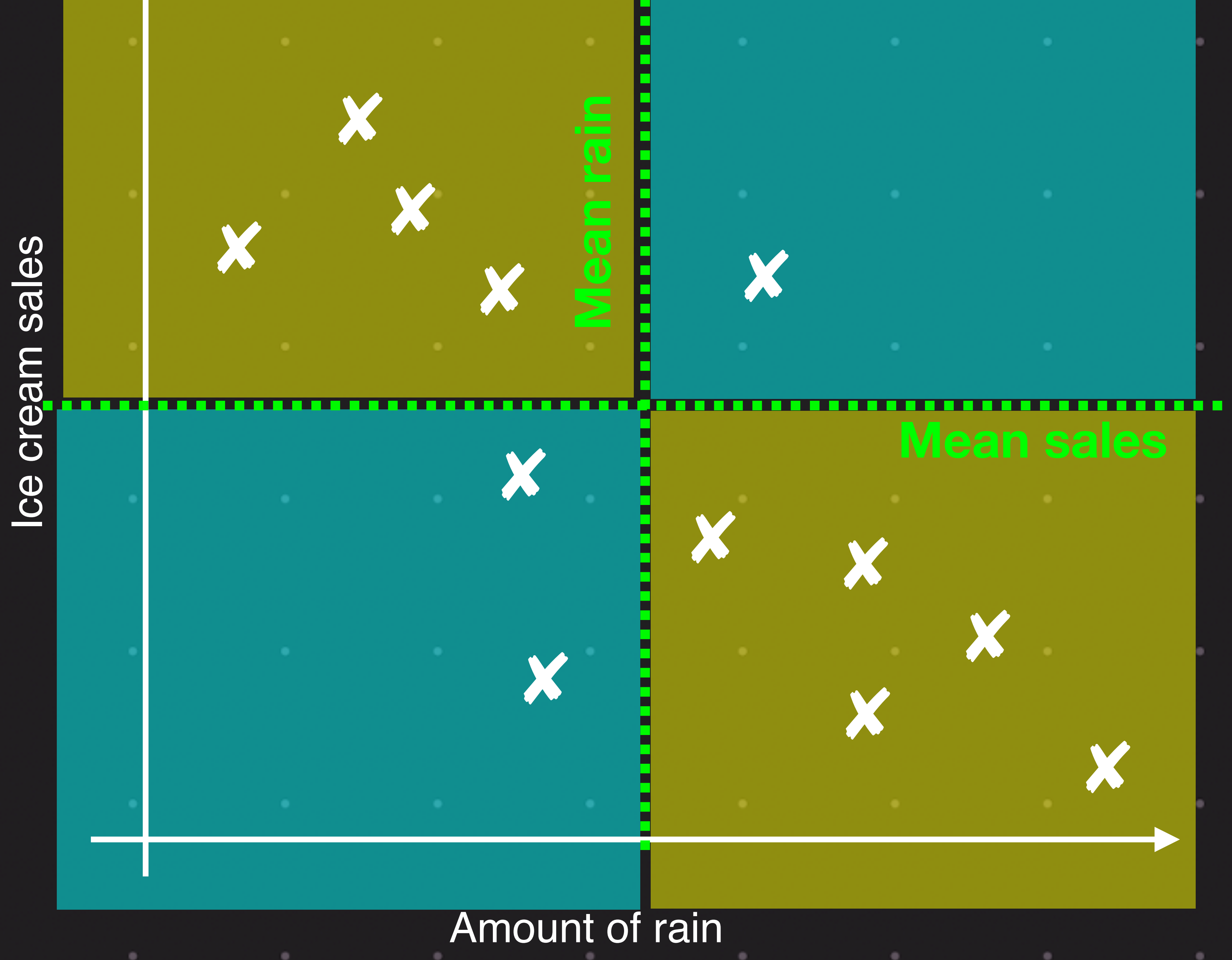
$$\frac{1}{10}(4 + 48 + 6 - 2 + \cdots + 60)$$

Which has more influence? Positive or negative

Positive has more influence

We say that these two features are positively correlated

Ice cream Vs Rain



Positive correlation

- Top right
- Bottom left

Negative correlation

- Top left
- Bottom right

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Which has more influence? Positive or negative

Negative has more influence

We say that these two features are positively correlated

Height Vs Rain

Positive correlation

- Top right
- Bottom left

Negative correlation

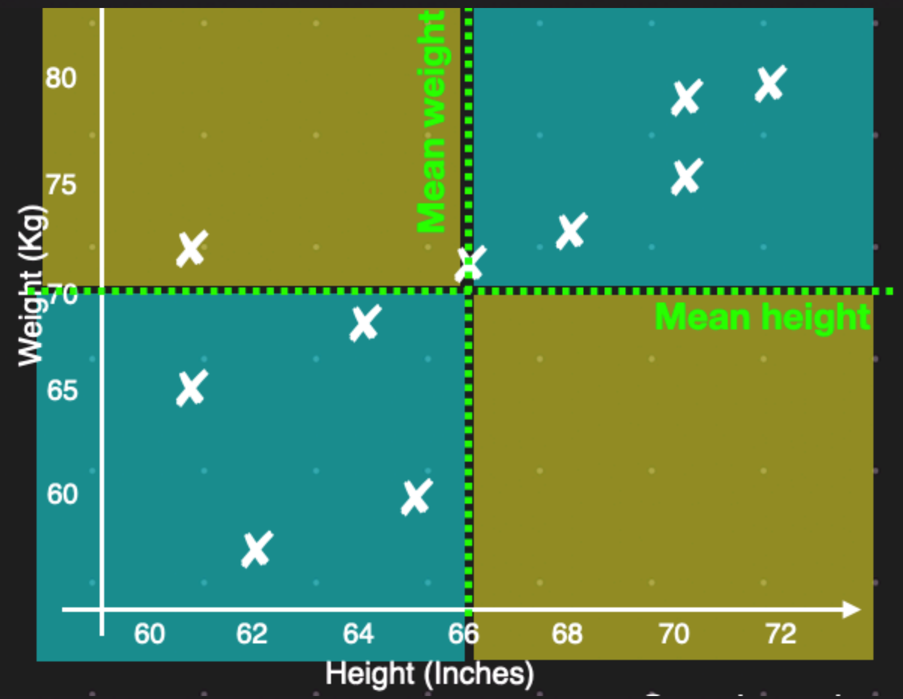
- Top left
- Bottom right

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Which has more influence? Positive or negative

Both have (approximately) equal influence

We say that these two features are uncorrelated



Suppose we express height in centimetres and weight in pounds

Simply stretching the axis should not have much influence on how we quantify correlation

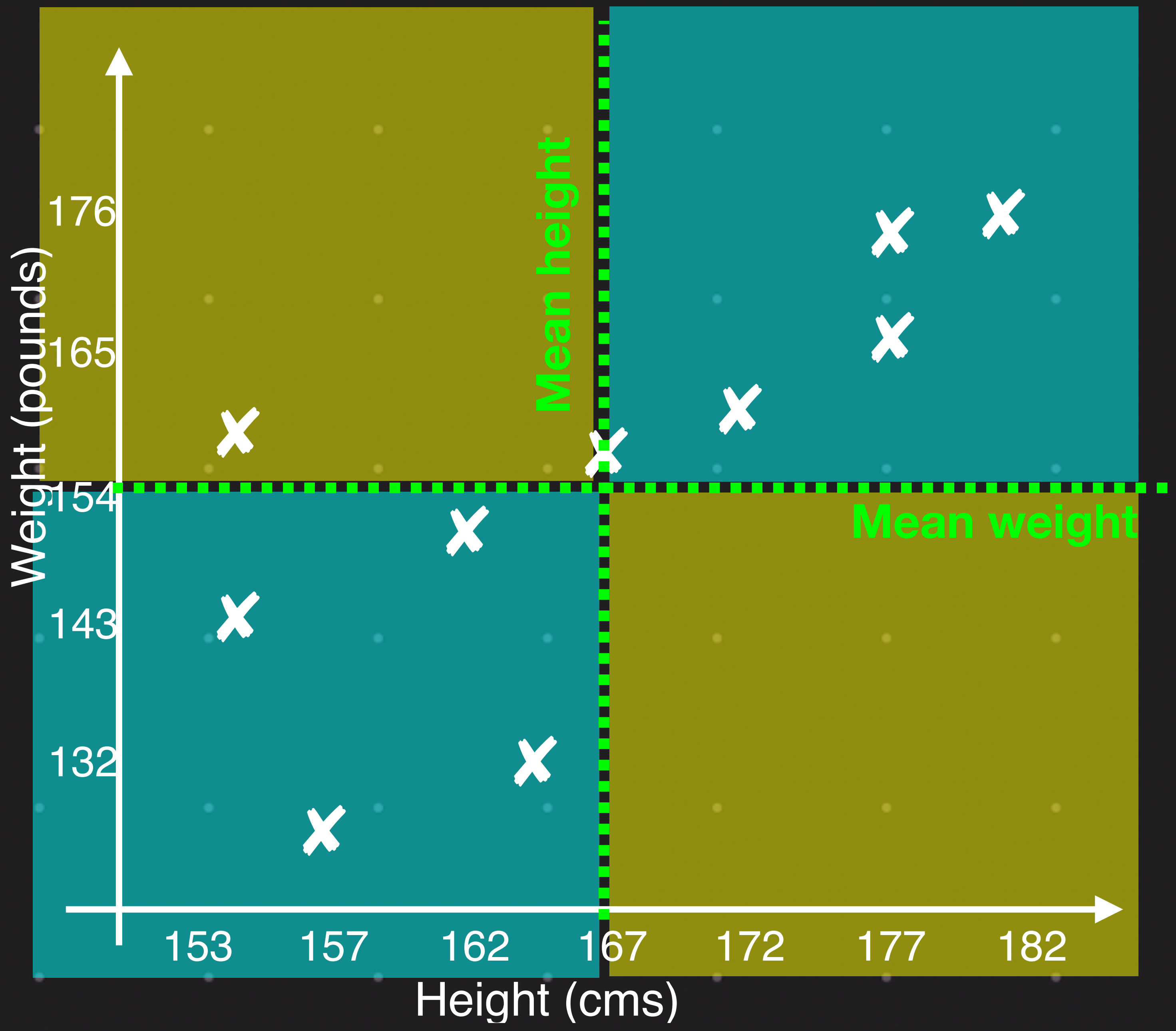
The definition of "correlation" does a standardisation of "covariance"

Positive correlation

- Top right
- Bottom left

Negative correlation

- Top left
- Bottom right



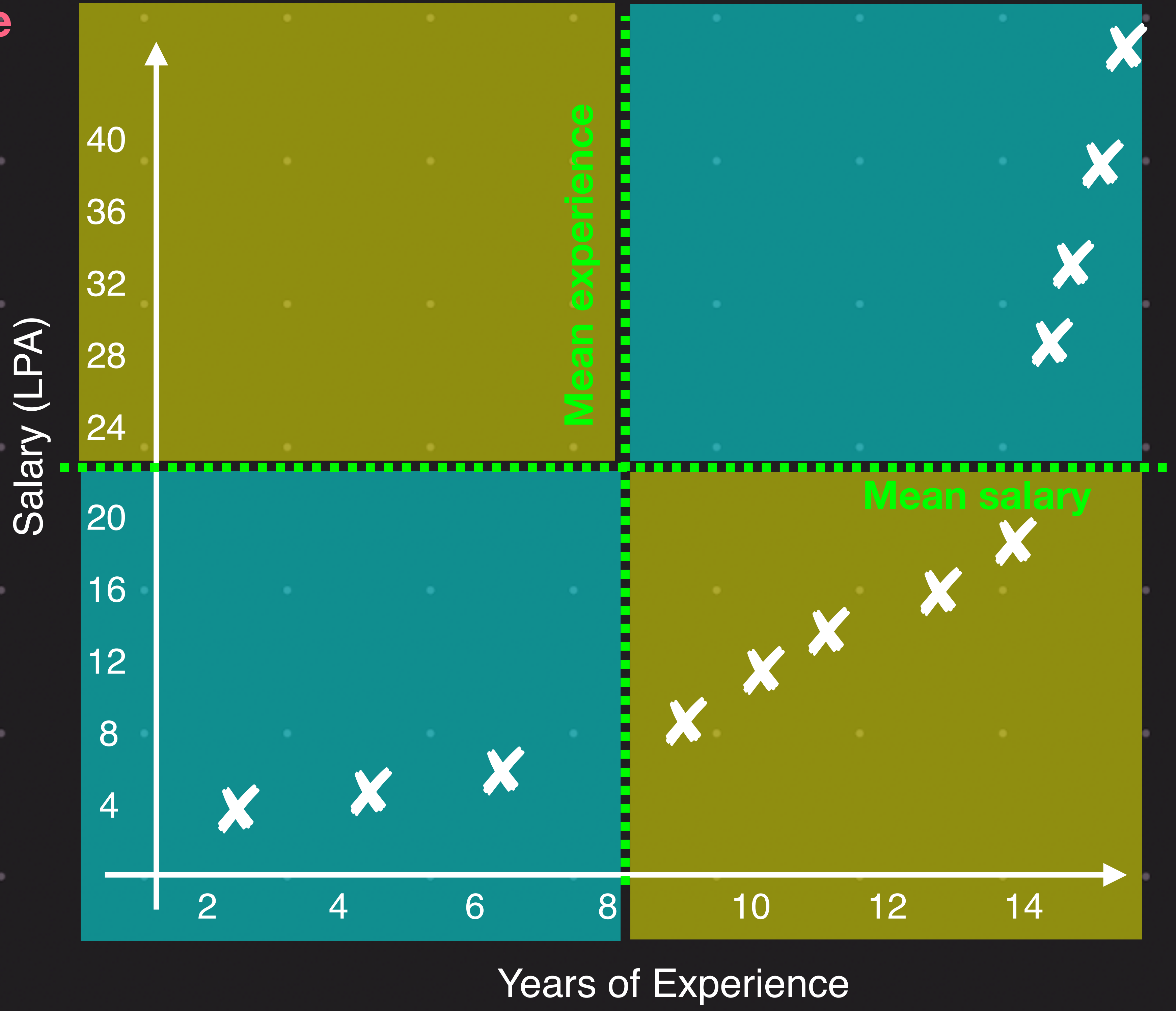
If we apply the formula of correlation, we get the same number whether we use the inch/ Kg axis or cms/pounds axis

$$\text{cov}(x, y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

$$-1 \leq \rho_{xy} \leq 1$$

Salary Vs Experience



Positive correlation
- Top right
- Bottom left

Negative correlation
- Top left
- Bottom right

$$\text{cov}(x, y) = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

$$-1 \leq \rho_{xy} \leq 1$$

Strange phenomenon: Even though we know that the two features are related, the correlation turns out to be very low

Spearman to the rescue!!!

Rank along both the x and y-axis, then take the correlations of the ranks

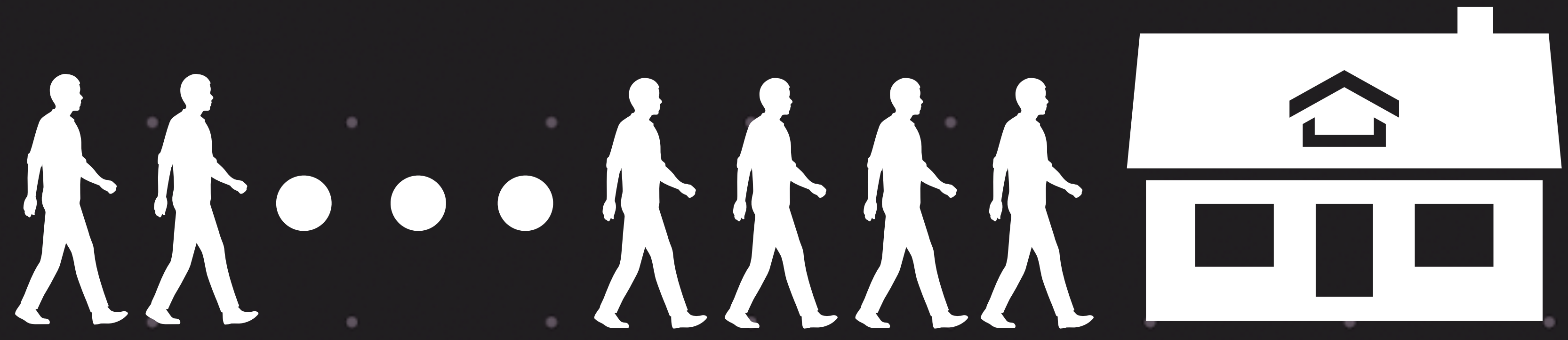
The average number of customers entering a store is 2000 per month

A marketing company is hired to improve this number

The next month, number of customers was seen to be 2128

With 95% confidence, is this improvement statistically significant?

2000 per month on average



What should the null and alternate hypothesis be?

$H_0 : \mu = 2000 \quad H_a : \mu > 2000$

What is the test statistic?

N: Number of people entering the store in a month

Distribution of the test statistic N? Poisson with rate 2000 per month

Right, left tailed, or two-tailed? Right tailed

What is the p-value?

$P[N \geq 2128 \,|\, H_0$ is true$] = 1 - P[N \leq 2127 \,|\, H_0$ is true$]$ = `1 - poisson.cdf(2127, mu=2000)` $= 0.002$

What is α? $\alpha = 0.05$

Is p-value $< \alpha$? Yes

We reject the null hypothesis We say that the marketing worked

Recommender System

When a customer buys a T-shirt, a recommender algorithm also suggests a few related items

The recommender system in production (legacy) that has a success rate of 10%

You and your team have developed a new deep learning algorithm for recommendation

It is tested before deploying. Of the next 500 customers, 72 bought items recommended by the new model.

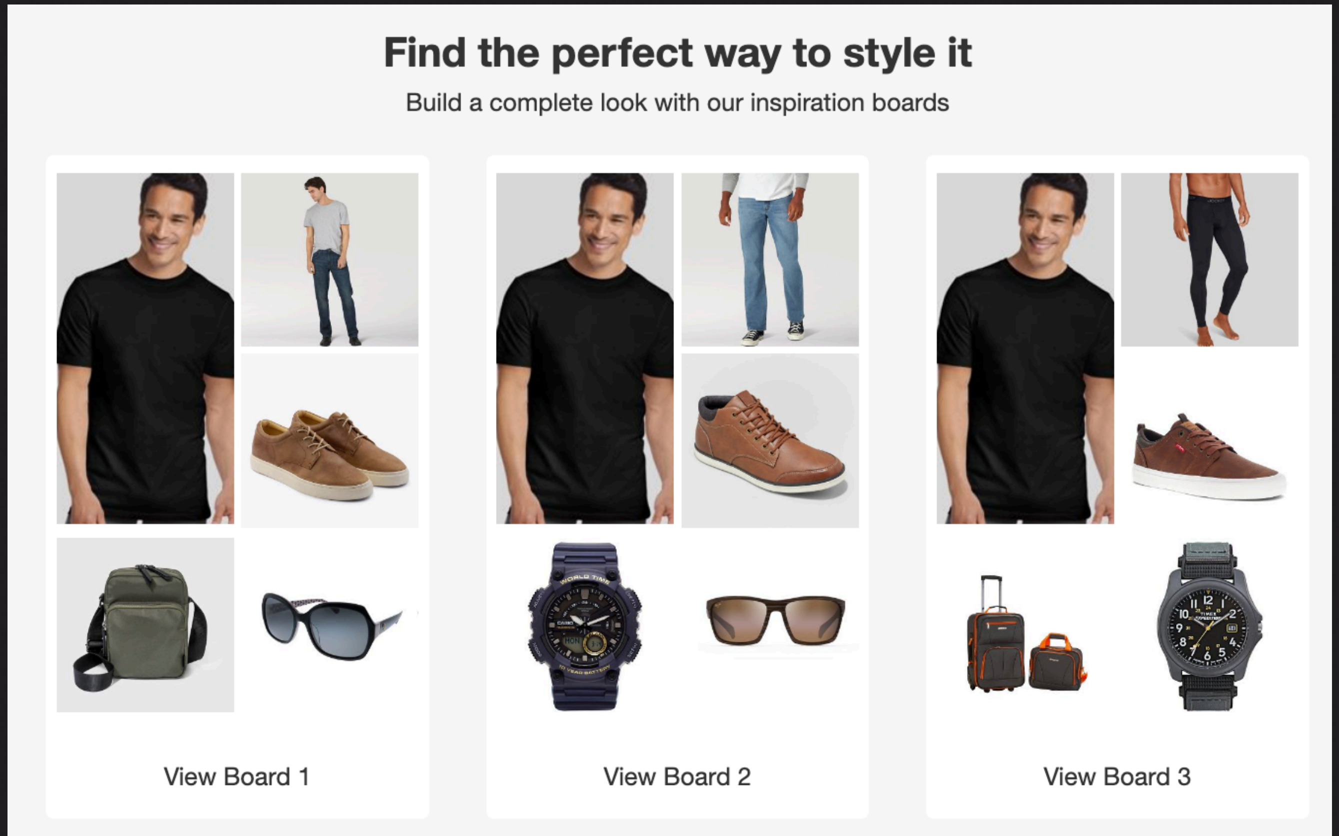
Is the improvement brought by the new model is statistically significant at 95% confidence?

Null and alternate hypothesis?

$H_0 : p = 0.1 \qquad H_a : p > 0.1$

H_0 assumes new model has same performance

This means that the $72/500 = 0.14$ of the new model is just fluke



What is the test statistic?

X : Number of people who bought the recommended items

Distribution of the test statistic X? Binom(n=500, p=0.1)

Right, left tailed, or two-tailed? Right tailed

What is the p-value?

$P[X \geq 72 \,|\, H_0 \text{ is true}] = 1 - P[X \leq 71 \,|\, H_0 \text{ is true}] =$ `1 - binom.cdf(71, n=500, p=0.1)` $= 0.001$

What is α? $\alpha = 0.05$

Is p-value $< \alpha$? Yes We reject the null hypothesis We say that the new model is better

Pearson Correlation

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i} (x_i - \bar{x})(y_i - \bar{y})$$

$$\rho_{hw} = \frac{\text{cov}(h, w)}{\sigma_h \sigma_w}$$

Spearman Correlation

Pearson correlation of rank(X) and rank(Y)