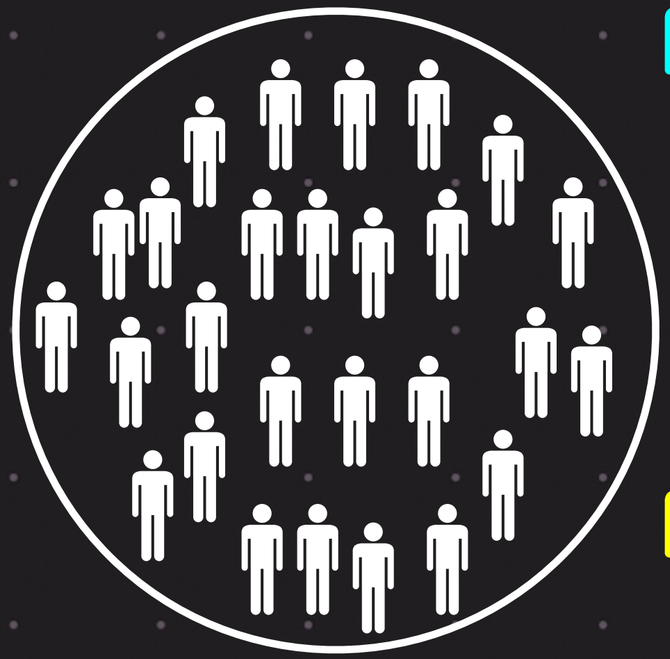
SDE-2 Salary



- Average of these three numbers is 35 L
- What is the unknown number? 34 L

If we know the sample mean of 3 numbers, then knowing 2 numbers is enough to know everything

- Average of these four numbers is 37 L
- What is the unknown number? 39 L

If we know the sample mean of 4 numbers, then knowing 3 numbers is enough to know everything

If we know the sample mean of n numbers, then knowing $n-1$ numbers is enough to know everything

Degree of freedom is said to be $n-1$

DF = $n-1$

Height and Weight

	Height (inches)	Weight (kg)
	73	85
	68	73
	74	96
	71	82
	62	70
Average	71	81.2

- We know the average height and weight of 5 people

  We want to fill the table

- How many minimum numbers in the table should we know?

  We need minimum 8 numbers DF = 8

  The number 8 comes as (5-1) + (5-1)

- In general, DF = n1 + n2 - 2

Sachin - Centuries and winning

Sachin has scored 46 centuries in 360 matches.

Of these 360 matches, India has won 184.

We want to construct the contingency table with centuries and win

Century \ Win	False	True	
False	160	154	314
True	16	30	46
	176	184	360

- We know these 5 numbers from data
  We want to fill the contingency table
- If we know this one number, can we fill the table with the other three? Yes

  One number is all we need!

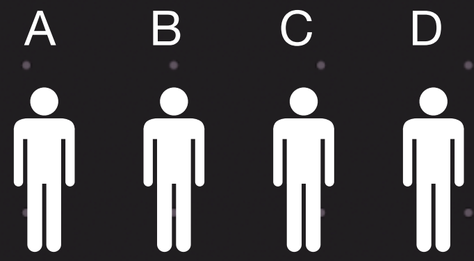
  DF = 1

Regional support for politicians

	A	B	C	D	Total
X	90	60	104	95	349
Y	30	50	51	20	151
Z	30	40	45	35	150
Total	150	150	200	150	650

- We know the total numbers from data
  We want to fill the contingency table
- How many minimum numbers in the table should we know?
  If we know these 6 numbers, can we fill the table? Yes
  DF = 6
- In general, DF = (#rows - 1) (#columns - 1)

4 politicians



3 cities

Degrees of Freedom

- If we know the sample mean of n numbers, then knowing $n - 1$ numbers is enough to know everything

  DF = $n - 1$

- If we know the sample means of two sets of numbers n_1 and n_2 numbers, then knowing $n_1 + n_2 - 2$ numbers is enough to know everything

  DF = $n_1 + n_2 - 2$

- In a contingency table, if we know the row sums and column sums, then

  DF = (#rows - 1) (#columns - 1)

Chi-Square Test

(A favourite word used by product managers)

- Suppose we have a lot of features in a machine learning model x_1, x_2, x_3, x_4
- We may have very big equation in these features

$$y = ax_1^2 + bx_2 + \cdots +$$

- Often you will be asked to do chi-squared test to remove variables that are not significant
- "This feature (say x_3) is not relevant, we have done chi-squared test. Let us remove this feature"

Going forward, the model will only use x_1, x_2, x_4

Chi-Square Test Coin toss 50 times

Let us set up the null and alternate hypothesis

H_0 : Fair coin H_a : Biased coin

We shall use a new test statistic called

χ^2 Test statistic ("chi-squared")

$$\chi^2 = \frac{(28-25)^2}{25} + \frac{(22-25)^2}{25} = 0.72$$

If the coin is fair, should this number be large or small?

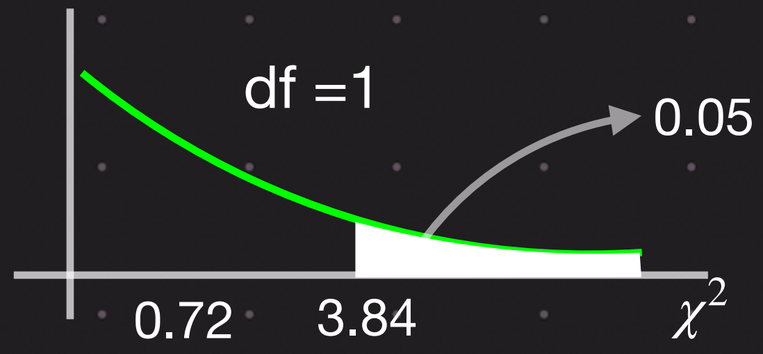
Small

	Heads	Tails
Expected	25	25
Actual	28	22

Knowing one number, we know the full table DF = 1

DF = (#rows - 1) (#cols - 1) $= (2-1)(2-1) = 1$

Let us see the distribution of the χ^2 test statistic with df = 1



$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

Critical region for 95% confidence

```
from scipy.stats import chi2
cr = chi2.ppf(q=0.95, df=1)
```

cr = 3.84

```
from scipy.stats import chisquare
chi_stat, p_value = chisquare(
    [28, 22], [25, 25]
)

chi_stat = 0.72
p_value = 0.396
```

Fail to reject H_0 since observed χ^2 0.72 is less than 3.84

p-value > 0.05

Chi-Square Test Coin toss 50 times

Let us set up the null and alternate hypothesis

H_0 : Fair coin H_a : Biased coin

We shall use a new test statistic called χ^2 Test statistic ("chi-squared")

$$\chi^2 = \frac{(45-25)^2}{25} + \frac{(5-25)^2}{25} = 32$$

If the coin is fair, should this number be large or small?

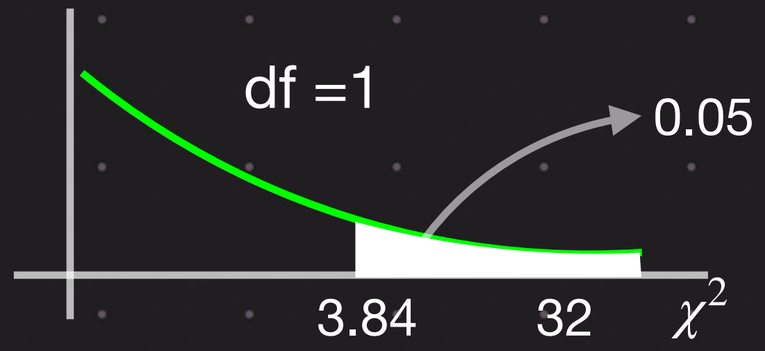
Small

	Heads	Tails
Expected	25	25
Actual	45	5

Knowing one number, we know the full table DF = 1

DF = (#rows - 1) (#cols - 1) $= (2-1)(2-1) = 1$

Let us see the distribution of the χ^2 test statistic with df = 1



$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

Critical region for 95% confidence

```
from scipy.stats import chi2
cr = chi2.ppf(q=0.95, df=1)
```

cr = 3.84

```
from scipy.stats import chisquare
chi_stat, p_value = chisquare(
    [45, 5], [25, 25]
)

chi_stat = 32
p_value = 1.54e-08
```

Reject H_0 since observed χ^2 32 is greater than 3.84

p-value < 0.05

Chi-Square Test

Dice, 36 times

	1	2	3	4	5	6
Expected	6	6	6	6	6	6
Actual	2	4	8	9	3	10

H_0 : Fair dice H_a : Biased dice

Test statistic

$$\chi^2 = \frac{(2-6)^2}{6} + \frac{(4-6)^2}{6} \cdots + \frac{(10-6)^2}{6} = 9.66$$

Critical region for 90% confidence

```
from scipy.stats import chi2
cr = chi2.ppf(q=0.90, df=5)
cr = 9.24
```

Reject H_0 since observed χ^2 9.66 is greater than 9.24

Degrees of freedom

DF = (#rows - 1) (#cols - 1)

DF = (2 - 1)(6 - 1) = 5

$\alpha = 0.1$

```
from scipy.stats import chisquare
chi_stat, p_value = chisquare(
    [2, 4, 8, 9, 3, 10],
    [6, 6, 6, 6, 6, 6]
)

chi_stat = 9.66
p_value = 0.0852
```

p-value < 0.1

Online Vs Offline shopping Does gender effect this?

Observed

	Male	Female		
Offline	527	72	599	66%
Online	206	102	308	34%
	733	174	907	

Expected

	Male	Female	
Offline	484	115	599
Online	249	59	308
	733	174	907

All these are observed values

To compute χ^2 test statistic, what do we need? The expected values

What percent people prefer offline? 66%

Among 733 males, how many are expected to prefer offline? 733 * 0.66 = 484

Among 174 females, how many are expected to prefer offline? 174 * 0.66 = 115

What percent people prefer online? 34%

Among 733 males, how many are expected to prefer online? 733 * 0.34 = 249

Among 174 females, how many are expected to prefer online? 174 * 0.34 = 59

Online Vs Offline shopping Does gender effect this?

Observed

	Male	Female		
Offline	527	72	599	66%
Online	206	102	308	34%
	733	174	907	

Expected

	Male	Female	
Offline	484	115	599
Online	249	59	308
	733	174	907

DF = (2-1) * (2-1) = 1

$$\chi^2 = \frac{(527-484)^2}{484} + \frac{(72-115)^2}{115} + \frac{(206-249)^2}{249} + \frac{(102-59)^2}{59} = 59$$

Critical region for 90% confidence

```
from scipy.stats import chi2
chi2.ppf(q=0.9, df=1)
```

cr = 2.7

Reject H_0 since χ^2 is greater than 2.7

$\alpha = 0.1$

```
from scipy.stats import chi2_contingency
observed = [
    [527, 72],
    [206, 102]
]
chi_stat, p_value, df, exp_freq = chi2_contingency(observed)
chi_stat = 57.04
p_value = 4e-14
```

p-value < 0.1

Assumptions of Chi2 test

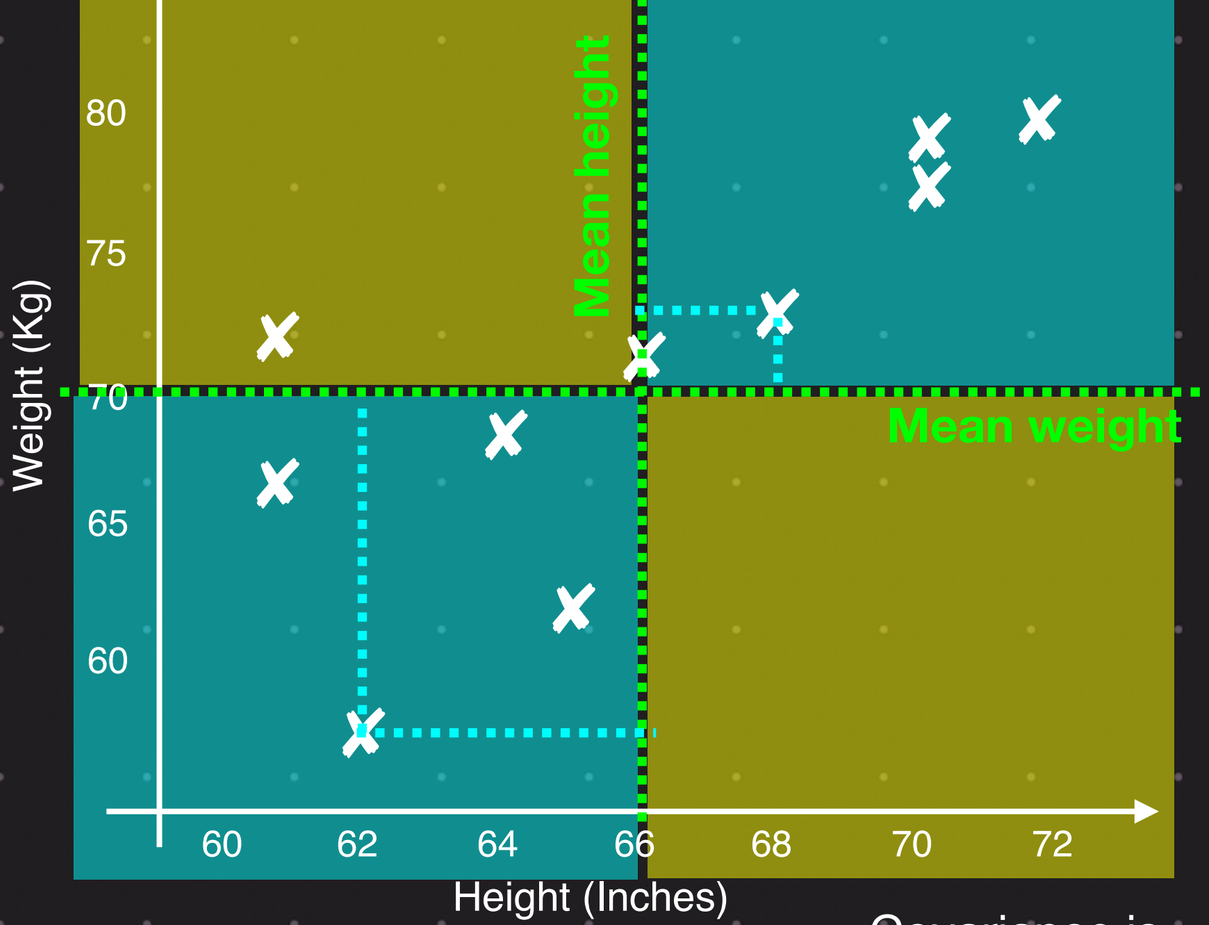
Variables are categorical

Observations are independent

Each cell is mutually exclusive

Expected value in each cell is greater than 5 (at least in 80% of cells)

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

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

$$(68 - 66)(72 - 70) = 2 * 2 = 4$$
$$(62 - 66)(58 - 70) = (-4) * (-12) = 48$$
$$(64 - 66)(67 - 70) = (-2) * (-3) = 6$$
$$(61 - 66)(72 - 70) = (-5)(2) = -10$$
$$\vdots$$
$$(72 - 66)(80 - 70) = (6)(10) = 60$$

Covariance is the average of all these numbers

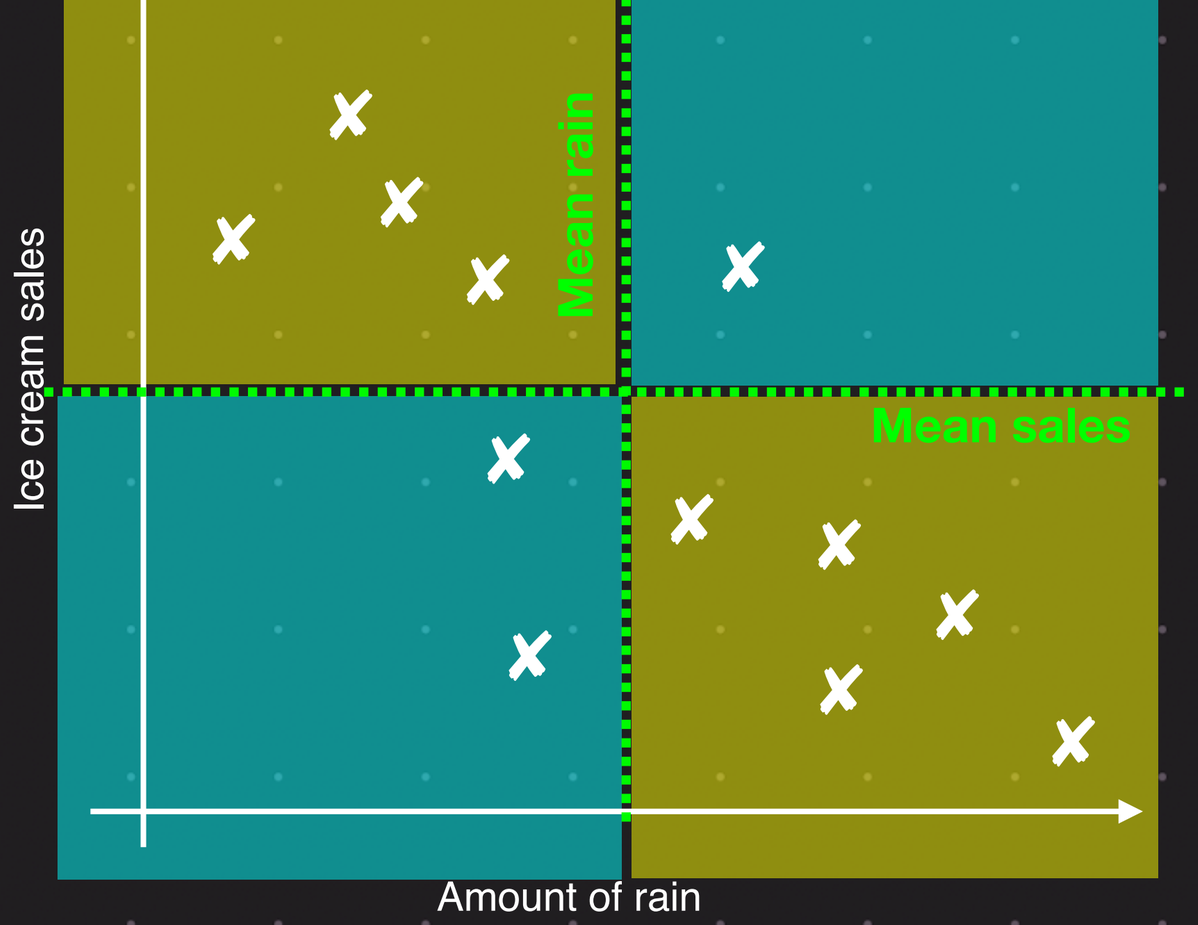
$$\frac{1}{10}(4 + 48 + 6 - 2 + \cdots + 60)$$

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

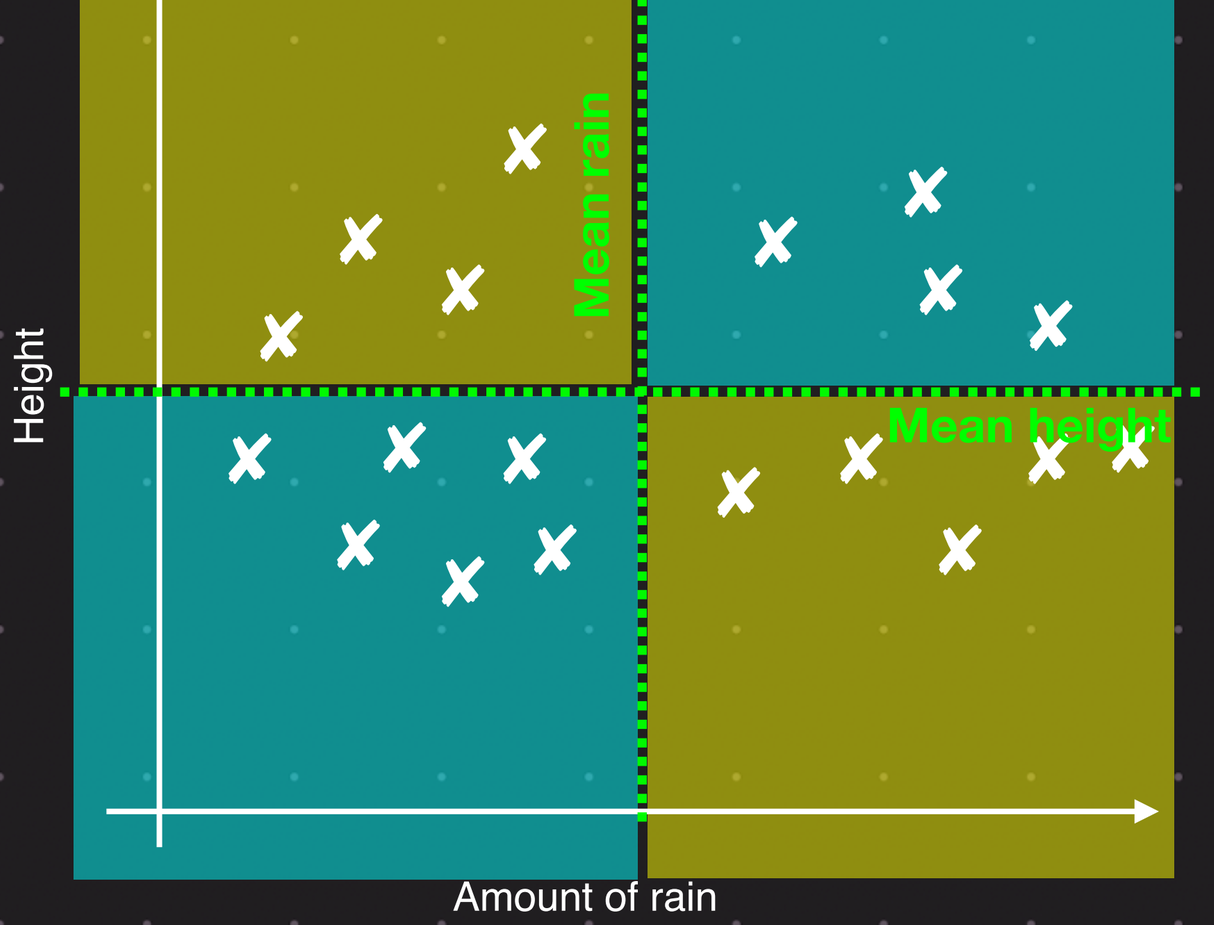
$$\text{cov}(x, y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

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

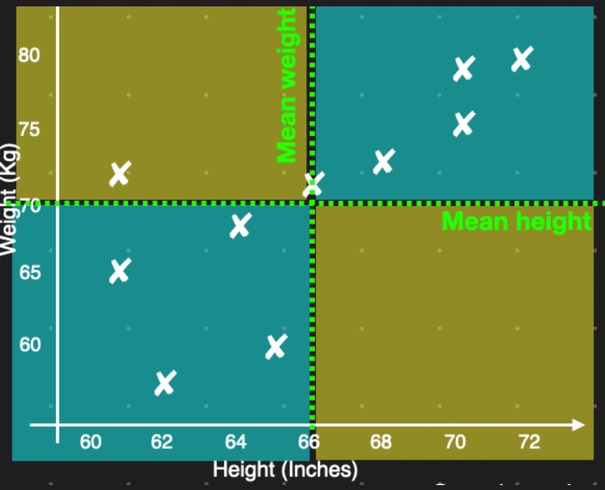
Negative correlation
- Top left
- Bottom right

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

Which has more influence? Positive or negative

Both have (approximately) equal influence

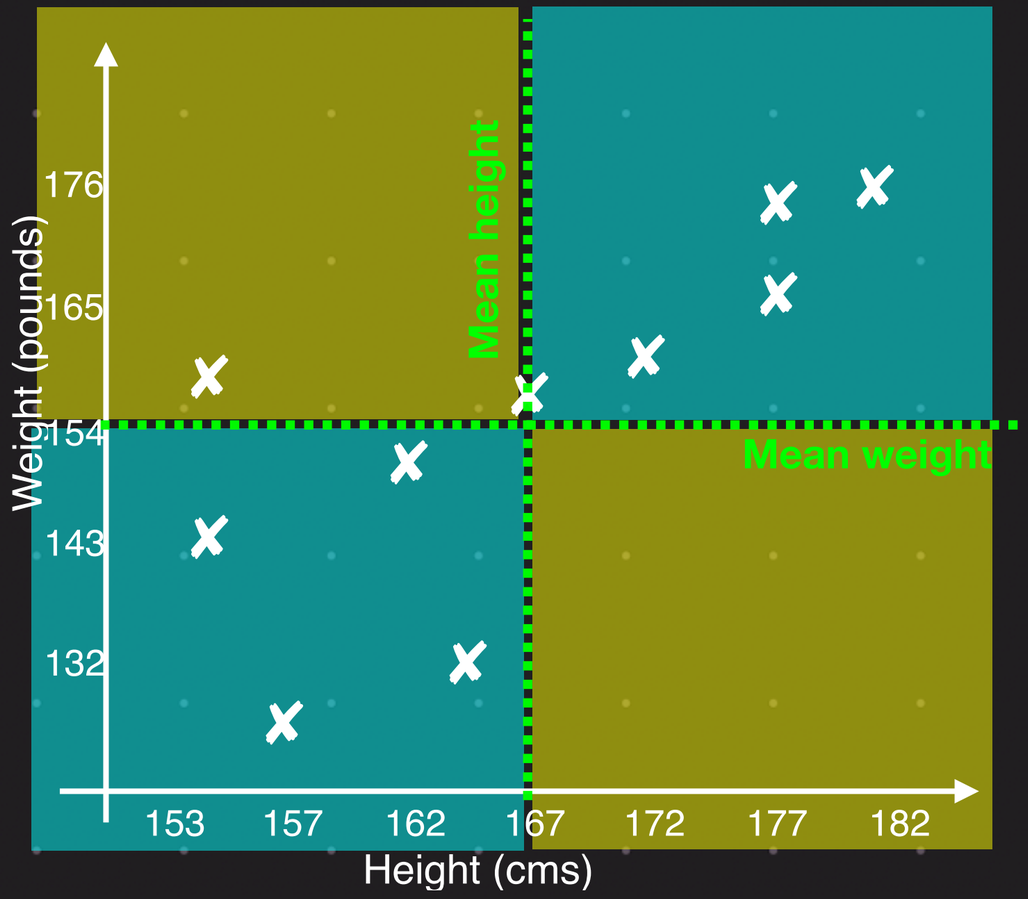
We say that these two features are uncorrelated



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”



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

Positive correlation

- Top right
- Bottom left

Negative correlation

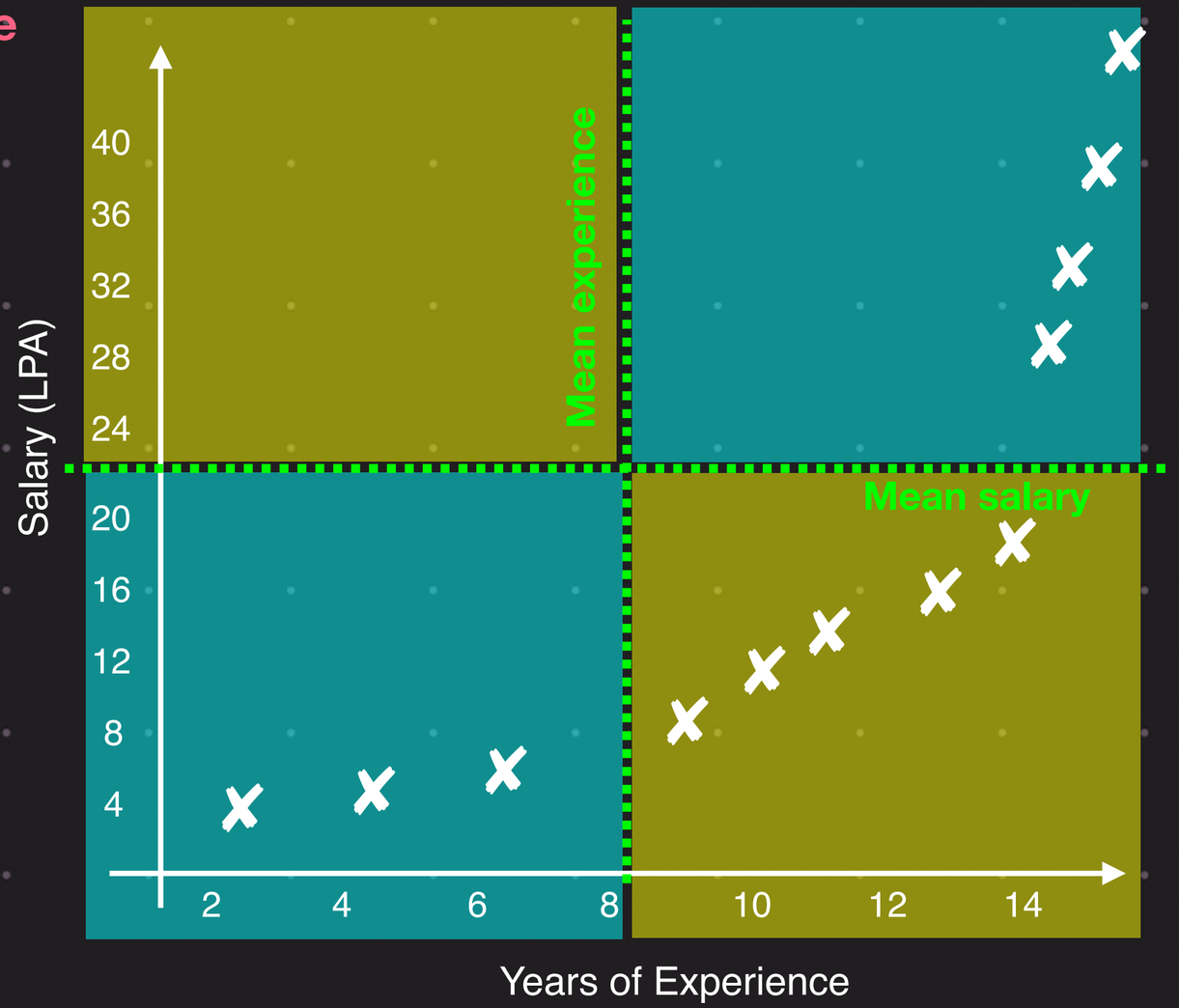
- Top left
- Bottom right

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

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

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

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!!!

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)