LOGISTIC REGRESSION

LOGISTIC REGRESSION EQUATION:

The underlying algorithm of Maximum Likelihood Estimation (MLE) determines the regression coefficient for the model that accurately predicts the probability of the binary dependent variable. The algorithm stops when the convergence criterion is met or maximum number of iterations are reached. Since the probability of any event lies between 0 and 1 (or 0% to 100%), when we plot the probability of dependent variable by independent factors, it will demonstrate an ‘S’ shape curve.

LOGIT TRANSFORMATION

- Logit Transformation is defined as follows-

Logit = Log (p/1-p) = log (probability of event happening/probability of event not happening) = log (Odds)

Logistic Regression is part of a larger class of algorithms known as GLM (Generlized Linear Model)

GENERALIZED LINEAR MODEL (GLM)

- Logistic Regression is part of a larger class of algorithms known as
- Generalized Linear Model (GLM).
- The fundamental equation of generalized linear model is:

$$g(E(y)) = \alpha + \beta x1 + \gamma x2$$

CASE-STUDY DATA

We are provided a sample of 1000 customers.

We need to predict the probability whether a **customer of a Particular Age** will buy (y) a particular magazine or

not.

As we’ve a categorical outcome variable, we’ll use logistic regression.

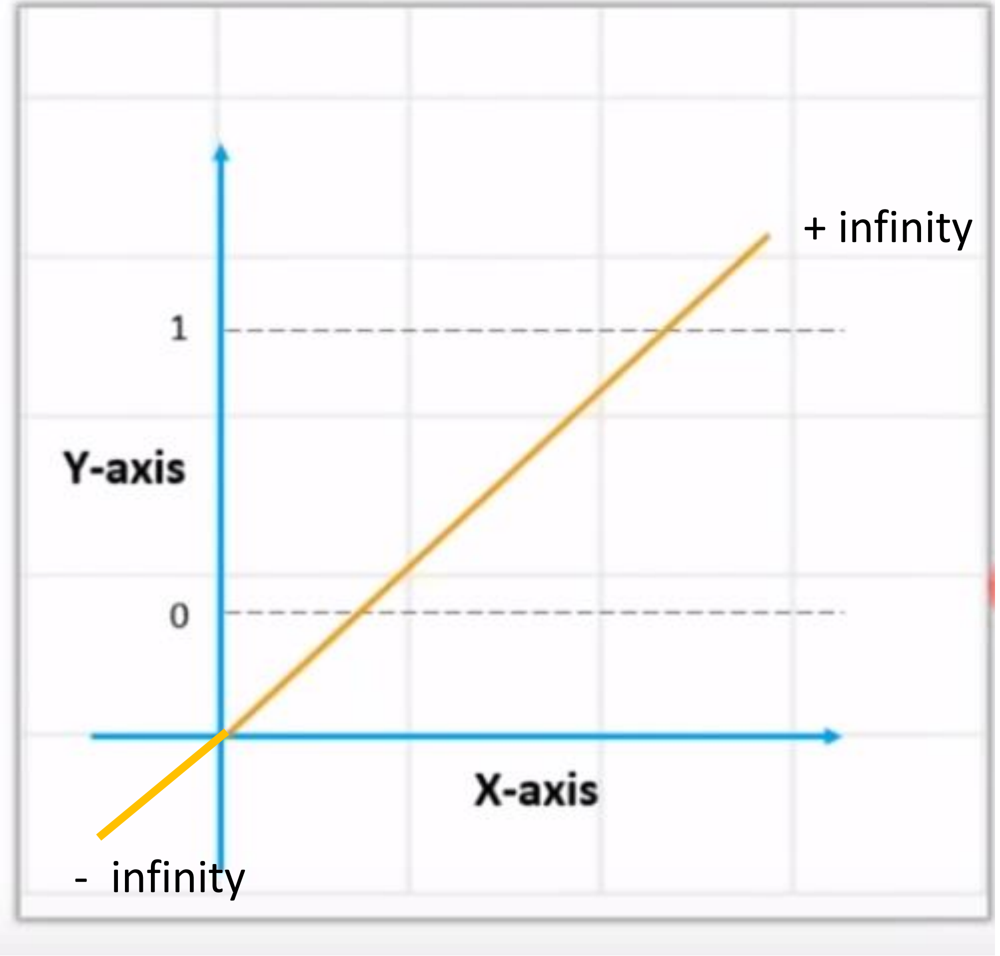
LINEAR TO LOGISTIC – (A)

- To start with logistic regression, first write the simple linear regression equation with dependent variable enclosed in a link function:

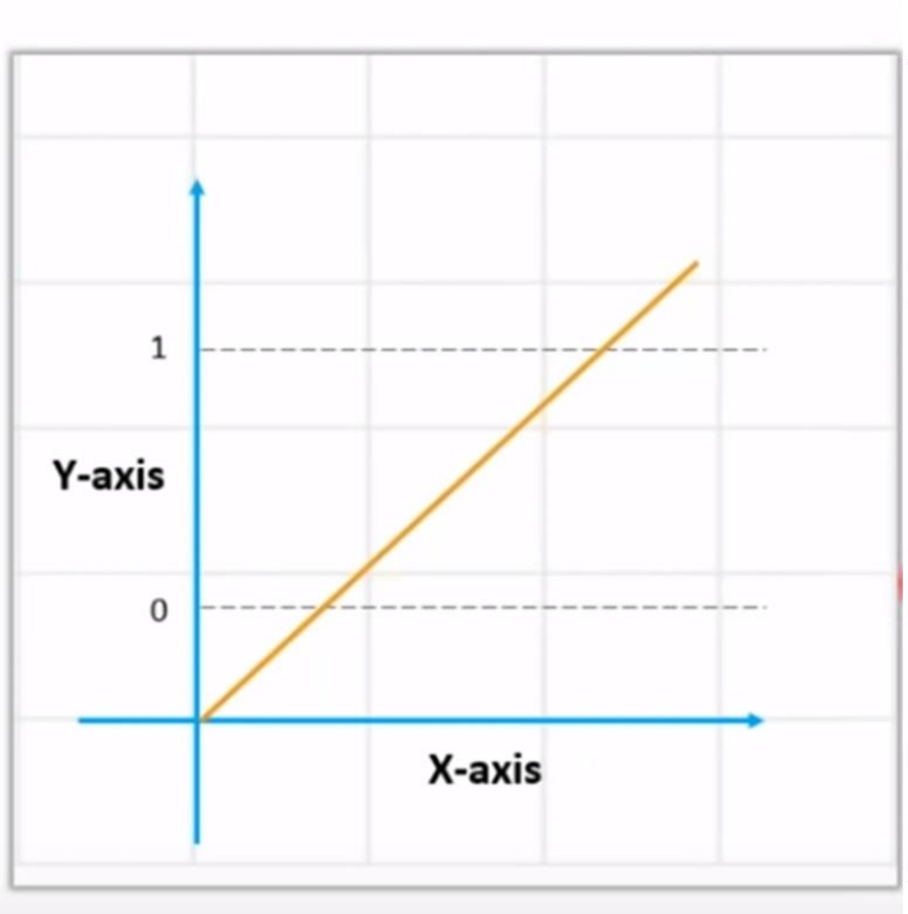
$$g(y) = \beta o + \beta(\text{Age}) \text{——— (a)}$$

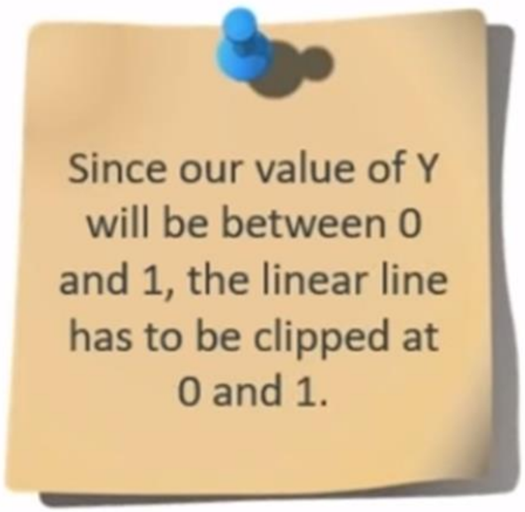
For understanding, consider '*Age*' as independent variable.

LINEAR REGRESSION

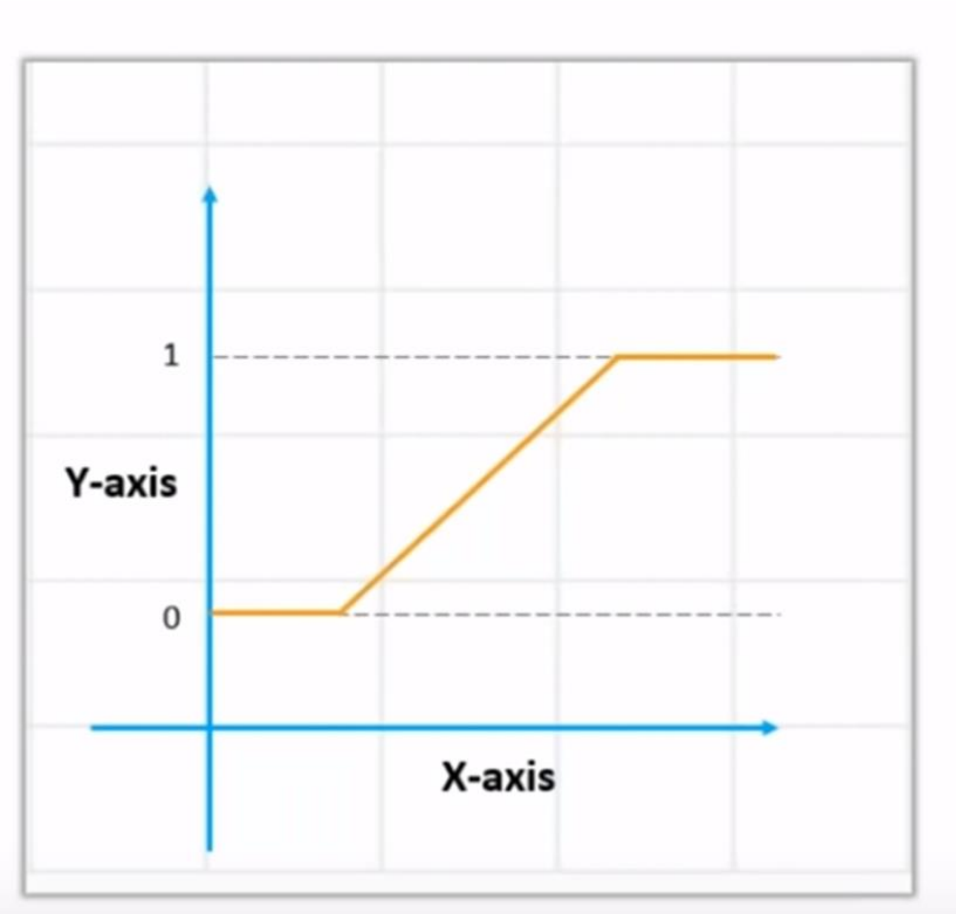


LINEAR REGRESSION EQUATION: Y = B0 + B1X1 + B2X2 …. + BNXN

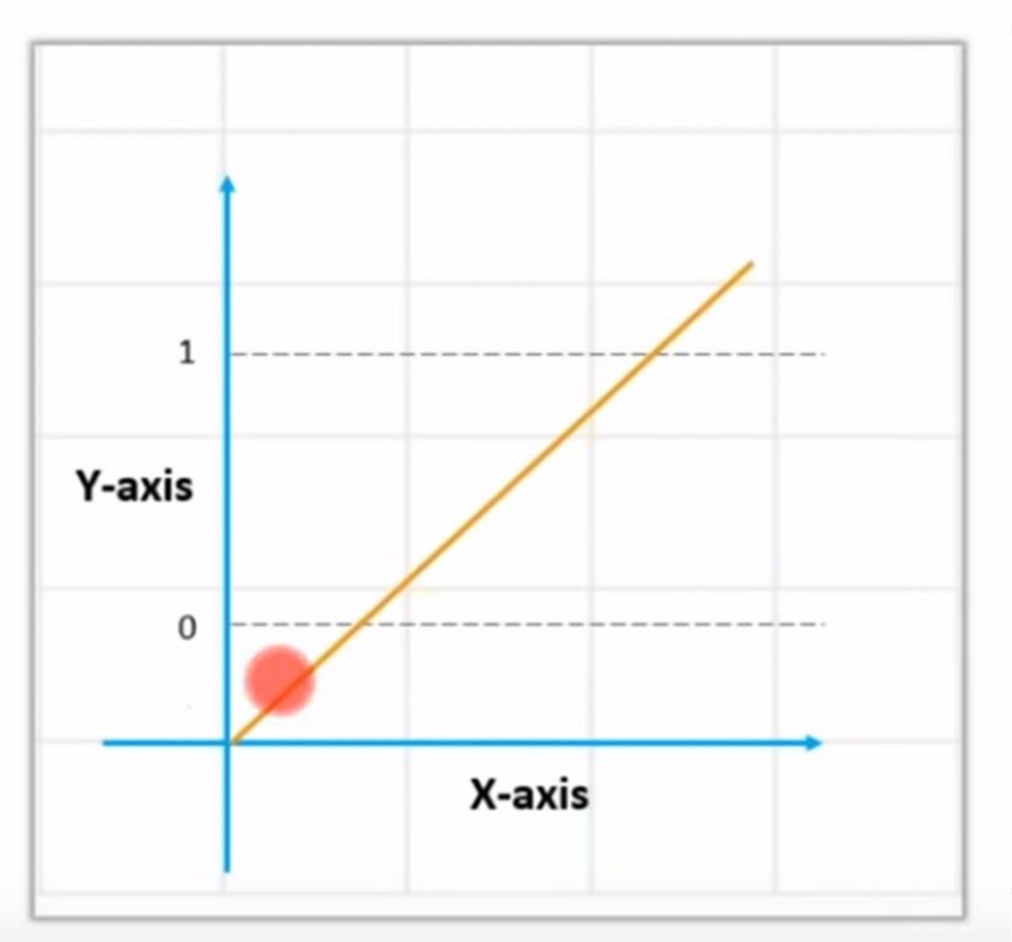


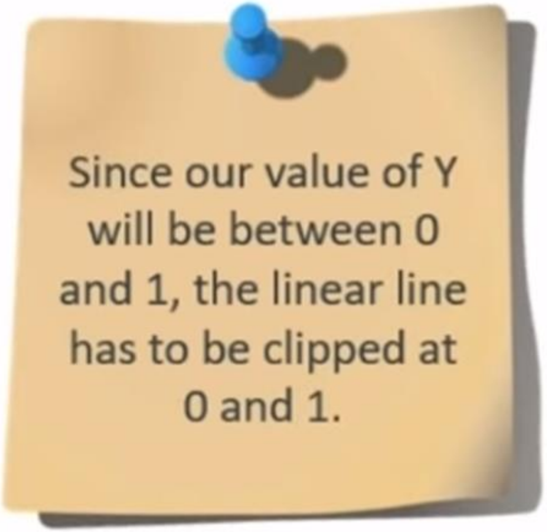


HOW TO GET THE VALUE OF 0 AND 1



VALUE OF Y – BETWEEN 0 AND 1





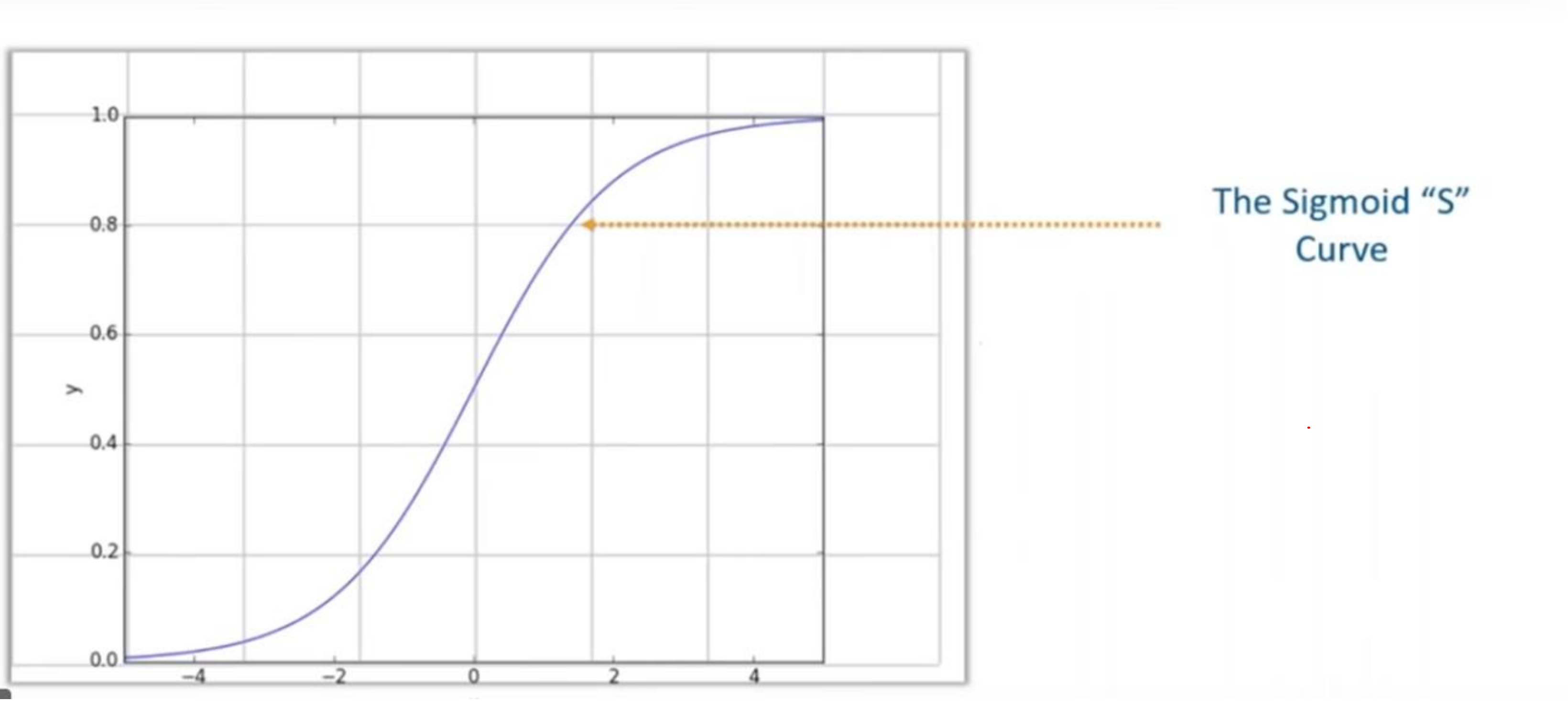
HOW TO GET THE VALUE OF 0 AND 1? USE SIGMOID

- We Apply sigmoid function on the linear regression equation to get the S-curve so that it lies between 0 and 1

Sigmoid function: $p = 1 / 1 + e^{-y}$

- A sigmoid function is a mathematical function/equation having a characteristic "S"-shaped curve or sigmoid curve.

SIGMOID – S-CURVE



Convert Linear to Logistics

- Linear regression equation: $\mathbf{y = \beta 0 + \beta 1X1 + \beta 2X2 \ldots. + \beta nXn}$
- Sigmoid function: $\boldsymbol{p} \mathbf{= 1 / 1 + e^{-y}}$

  $\mathbf{e^{-y}}$ y is replaced
- Logistic Regression equation: $\boldsymbol{p} \mathbf{= 1 / 1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 \ldots. + \beta nXn)}}$

LOGISTIC REGRESSION FORMULA

Logistic Regression

Putting z value to sigmoid function

Linear Regression Equation

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + .. + \beta_k x_k$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-z}}$$

$$p = \frac{e^z}{e^z + 1}$$

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k} + 1}$$

Looks like very hard to solve it, so let's try to transform it into some easy to solve equation with the help of Odds ratio.

Replace p in odd ratio and solve

$$\text{Odds Ratio } S = \frac{Probability\ of\ Success}{Probability\ of\ Failure} = \frac{p}{1-p}$$

$$S = \frac{\dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k} + 1}}{1 - \dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k} + 1}}$$

$$S = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k}$$

Take log each side and solve

$$\mathbf{Ln(S)} = \boldsymbol{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k}$$

Transformed into Linear Regression known as log of Odds

Logistic Regression

Putting z value to sigmoid function

Linear Regression Equation

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + .. + \beta_k x_k$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-z}}$$

$$p = \frac{e^z}{e^z + 1}$$

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k} + 1}$$

Replace p in odd ratio and solve

$$\text{Odds Ratio } S = \frac{\textit{Probability of Success}}{\textit{Probability of Failure}} = \frac{p}{1-p}$$

$$S = \frac{\dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k} + 1}}{1 - \dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k} + 1}}$$

$$S = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}$$

Take log each side and solve

$$\mathrm{Ln}(S) = \boldsymbol{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}$$

Transformed into Linear Regression known as log of Odds

Logistic Regression

Putting z value to sigmoid function

Linear Regression Equation

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-z}}$$

$$p = \frac{e^z}{e^z + 1}$$

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k} + 1}$$

Replace p in odd ratio and solve

$$\text{Odds Ratio } S = \frac{\textit{Probability of Success}}{\textit{Probability of Failure}} = \frac{p}{1-p}$$

$$S = \frac{\dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k} + 1}}{1 - \dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k} + 1}}$$

$$S = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}$$

Take log each side and solve

$$\mathbf{Ln(S)} = \boldsymbol{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}$$

Transformed into Linear Regression known as log of Odds

$$S = \cfrac{\cfrac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\dots+\beta_k x_k}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\dots+\beta_k x_k}+1}}{1-\cfrac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\dots+\beta_k x_k}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\dots+\beta_k x_k}+1}}$$

Adding Fractions with Unlike Denominators

find common denominator

$$\frac{3}{5}+\frac{3}{2}$$

$$\frac{2}{2}\times\frac{3}{5}+\frac{3}{2}\times\frac{5}{5}$$

$$\frac{6}{10}+\frac{15}{10}$$

$$\frac{21}{10}$$

$$S = \frac{\frac{e}{e+1}}{1-\frac{e}{e+1}} \Rightarrow \frac{e}{e+1} \div \left(1-\frac{e}{e+1}\right)$$

$$\Rightarrow \quad '' \quad \div \left(\frac{1}{1}-\frac{e}{e+1}\right)$$

$$=$$

$$\div \left(\frac{1}{1}\times e+1 - \frac{e}{e+1}\times 1\right)$$

$$\div \frac{e+1-e}{(e+1)} \Rightarrow \frac{1}{e+1}$$

$$\Rightarrow \frac{e}{(e+1)} \times \frac{(e+1)}{1}$$

$$\Rightarrow e$$

$$S = \frac{\dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k} + 1}}{1 - \dfrac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k} + 1}}$$

Logistic Regression

Putting z value to sigmoid function

Linear Regression Equation

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-z}}$$

$$p = \frac{e^z}{e^z + 1}$$

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k} + 1}$$

Looks like very hard to solve it, so let's try to transform it into some easy to solve equation with the help of Odds ratio.

Odds Ratio $S = \frac{p}{1-p}$

Replace p and solve

$$S = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}$$

Take log each side and solve

$$\mathrm{Ln}(S) = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k}_{\text{Transformed into Linear Regression}}$$

Notes:

- The log of Odds is called Logit and transformed model is linear in β_s
- So solving the logistic regression problem essentially reduces to finding the β_s that minimizes the error.
- Now suppose with one predictor we got the Linear Regression eq. ln(s) = -20.40782+.42592*x. And now we want to classify for given x = 50 then:
- Ln(s) = -20.40782+.42592*50 = 0.89 => s = $e^{0.89}$ = 2.435
- $s = \frac{p}{1-p}$ => $p = \frac{s}{s+1} \Rightarrow p = 2.435/(1+2.435) = .709$
- So using a probability of 0.50 as a cut-off between predicting the two classes 1 or 0, this member would be classified as class 1 with a probability of 70%

Final Notes:

1. The log of Odds is called Logit and transformed model is linear in β_s
2. So solving the logistic regression problem essentially reduces to finding the β_s that minimizes the error.
3. Now suppose with one predictor we got the Linear Regression eq.
   - ln(s) = -20.40782+.42592*x.
4. And now we want to predict for given x = 50 then put x = 50 in above eq:
5. Ln(s) = -20.40782+.42592*50 = 0.89 => s = $e^{0.89}$ = 2.435

(S)This is odds ratio value

6. s $= \frac{p}{1-p}$ => $p = \frac{s}{s+1} \Rightarrow p =$ 2.435/(1+2.435) = .709
7. So using a probability of 0.50 as a cut-off between predicting the two classes 1 or 0, this member would be classified as class 1 with a probability of 70%

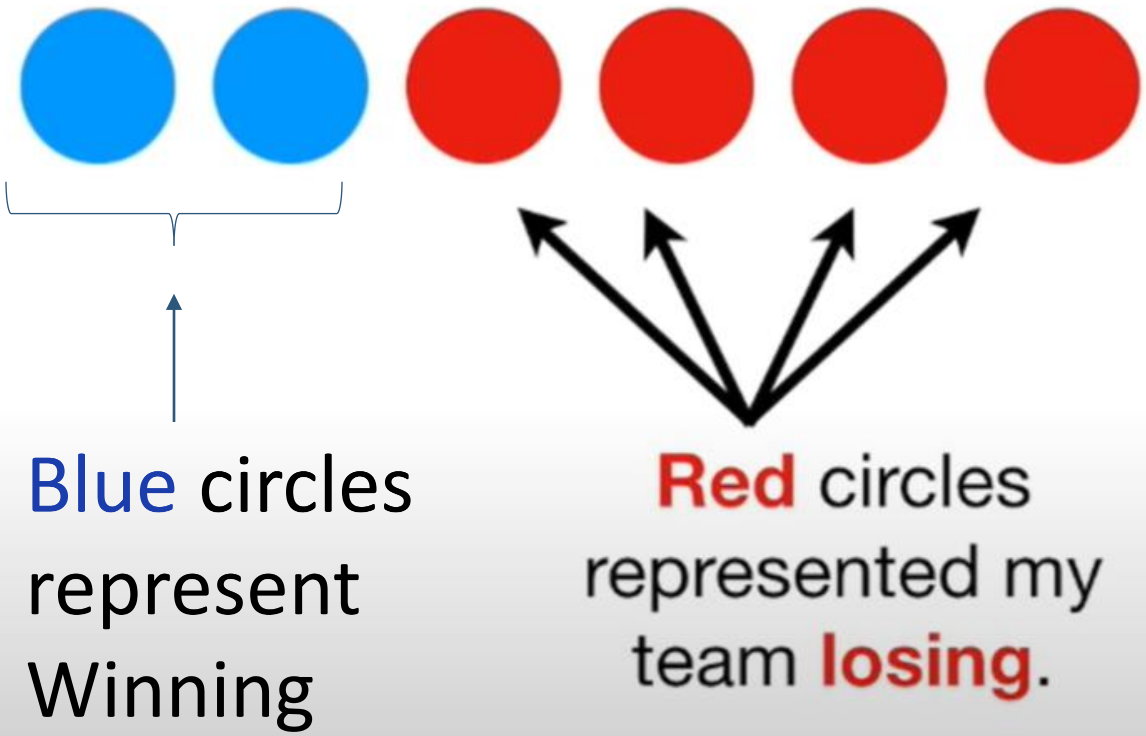
Odds Ratio

We want to find the probability (P) of occurrence, from the Odds ratio. So we put the value of S in the equation = p/1-p

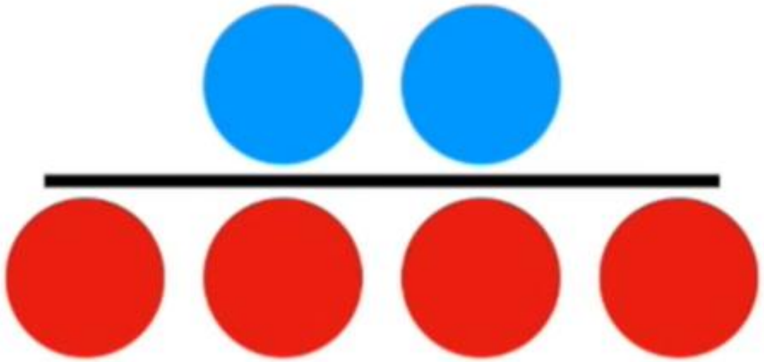
DIFFERENCE BETWEEN ODDS VS LOG(ODDS)

What are odds? =

$$\frac{\text{…the ratio of something happening (i.e. my team winning)…}}{\text{…to something not happening (i.e. my team not winning).}}$$

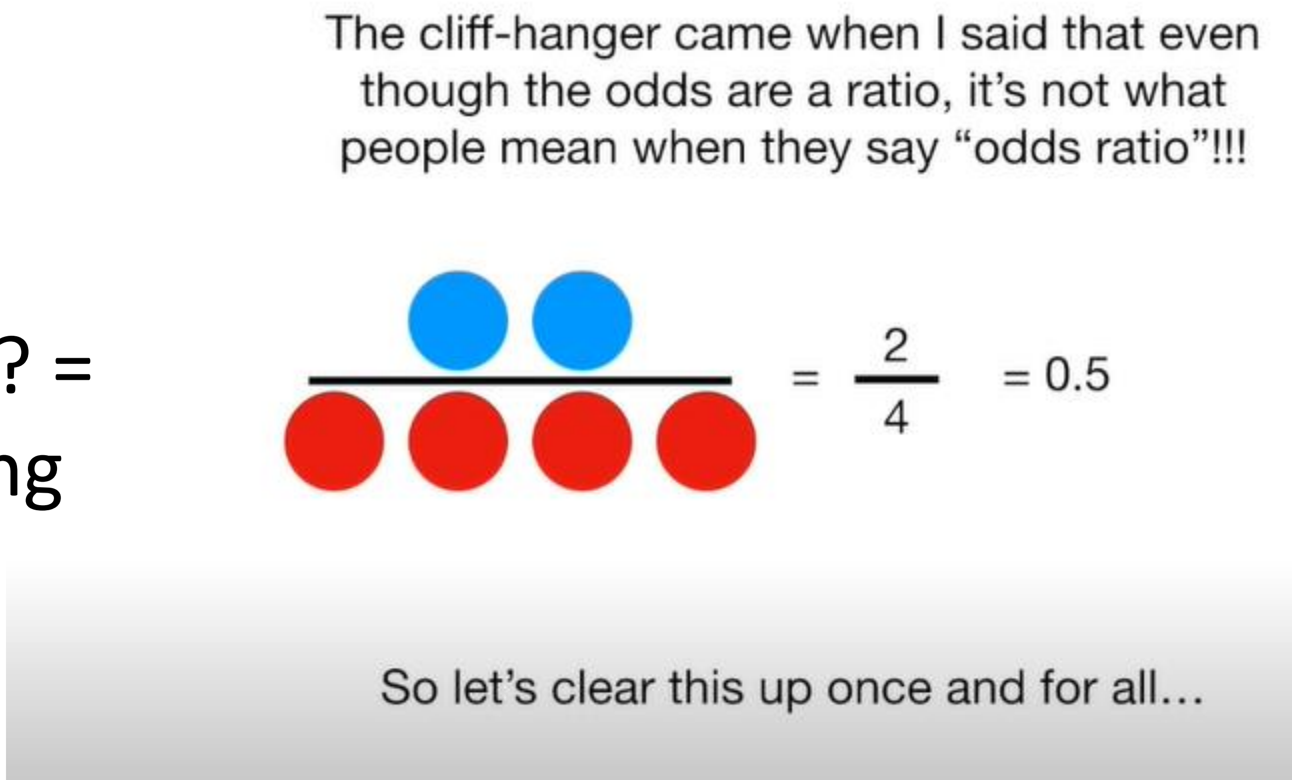

Blue circles represent Winning
Red circles represented my team losing.

What are odds? =
Odds of Winning



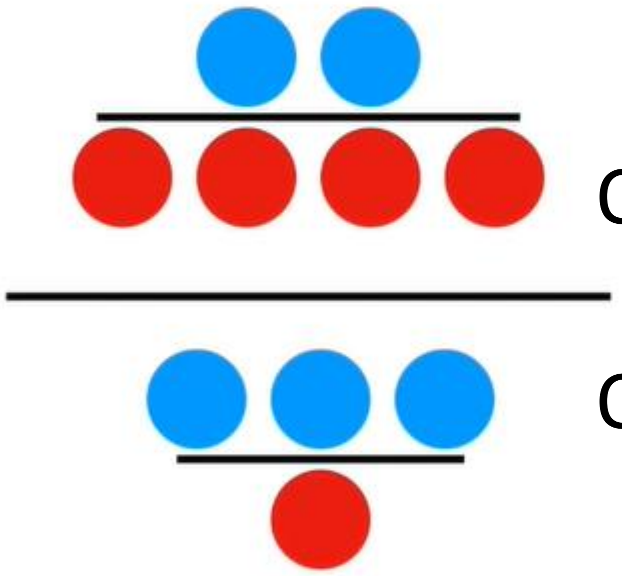
$= \frac{2}{4}$

What are odds? =
Odds of Winning



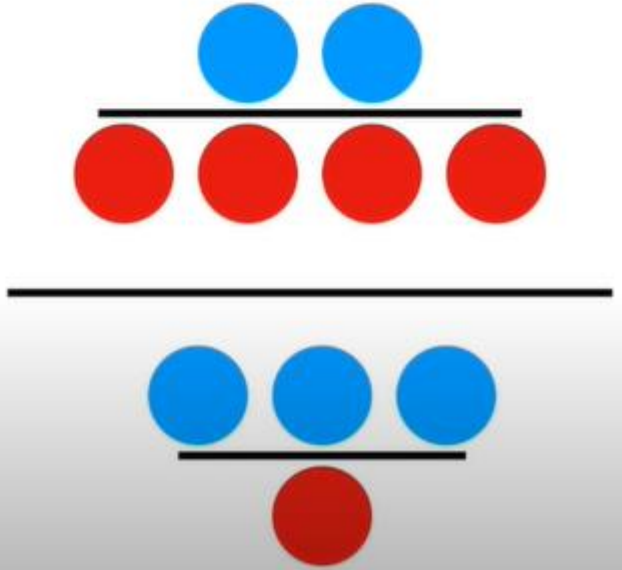
When people say “odds ratio”, they are talking about a “**ratio of odds**”.

What are Odds Ratio? =



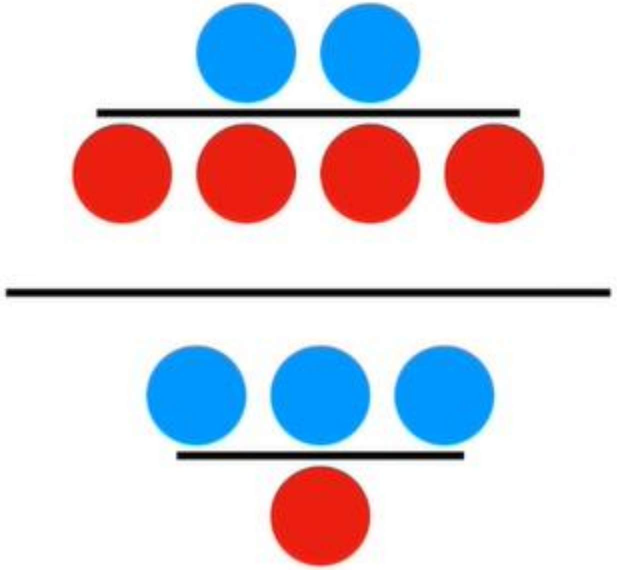
$$\frac{\text{Odds of Winning for Example 1}}{\text{Odds of Winning for Example 2}}$$

Here example 1 and Example 2 are, two different Games, and we are just using the Ratio of the Odds for each example



Doing the math gives us…

$$= \frac{2/4}{3/1}$$

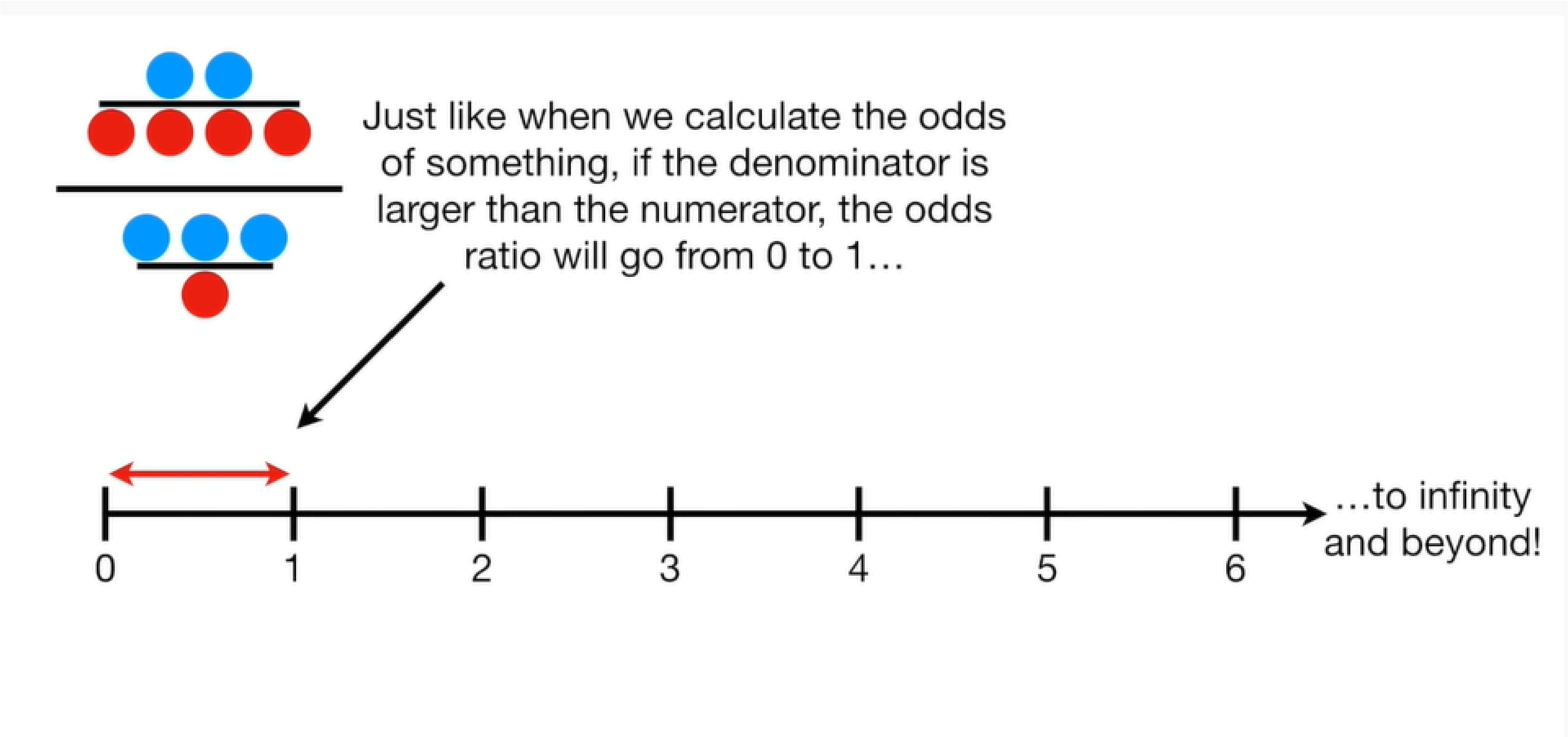


Doing the math gives us…

$$= \frac{2/4}{3/1} = 0.17$$

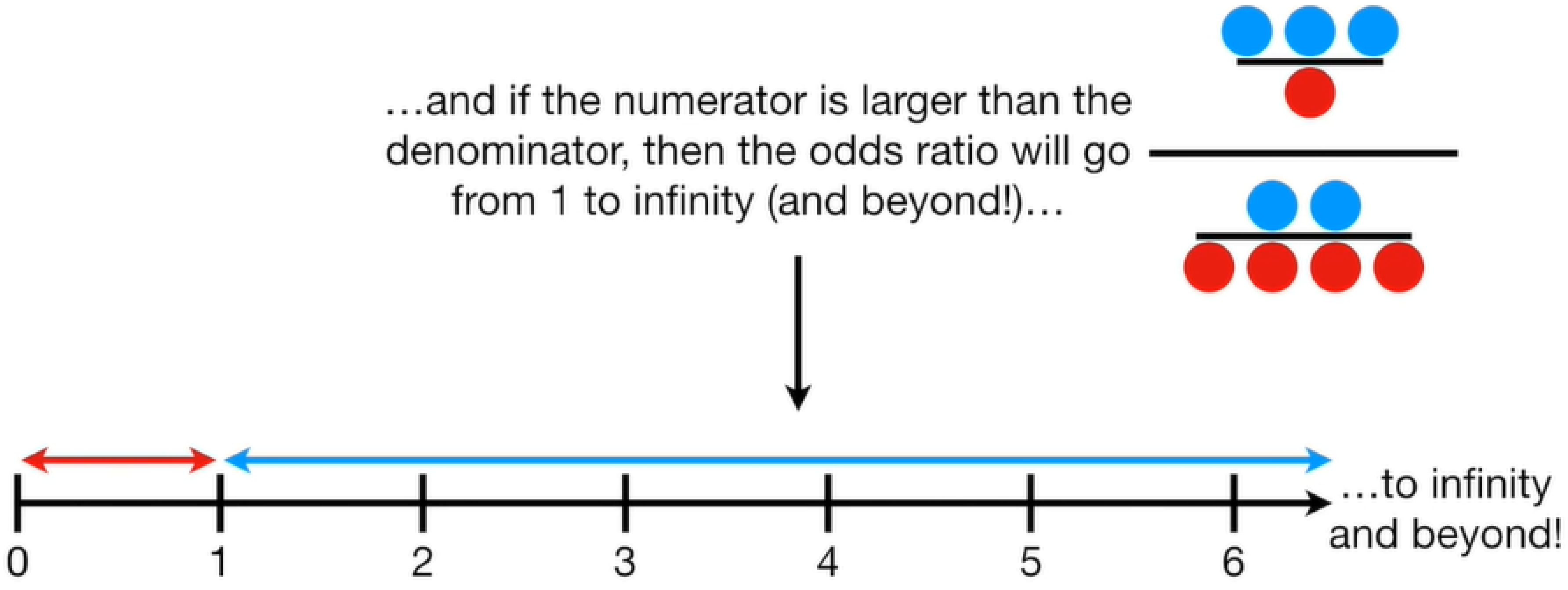
Odds of Winning for Example 1

Odds of Winning for Example 2



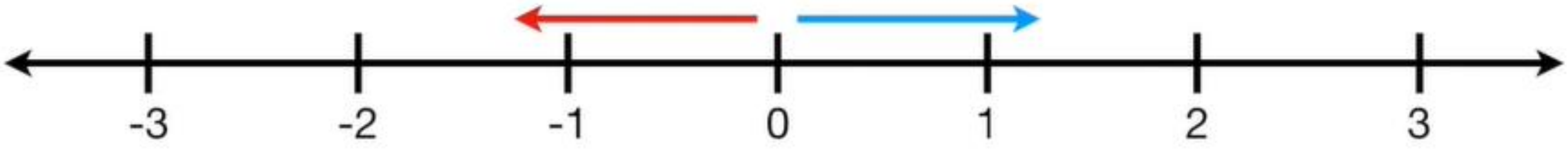
Odds of Winning for Example 2

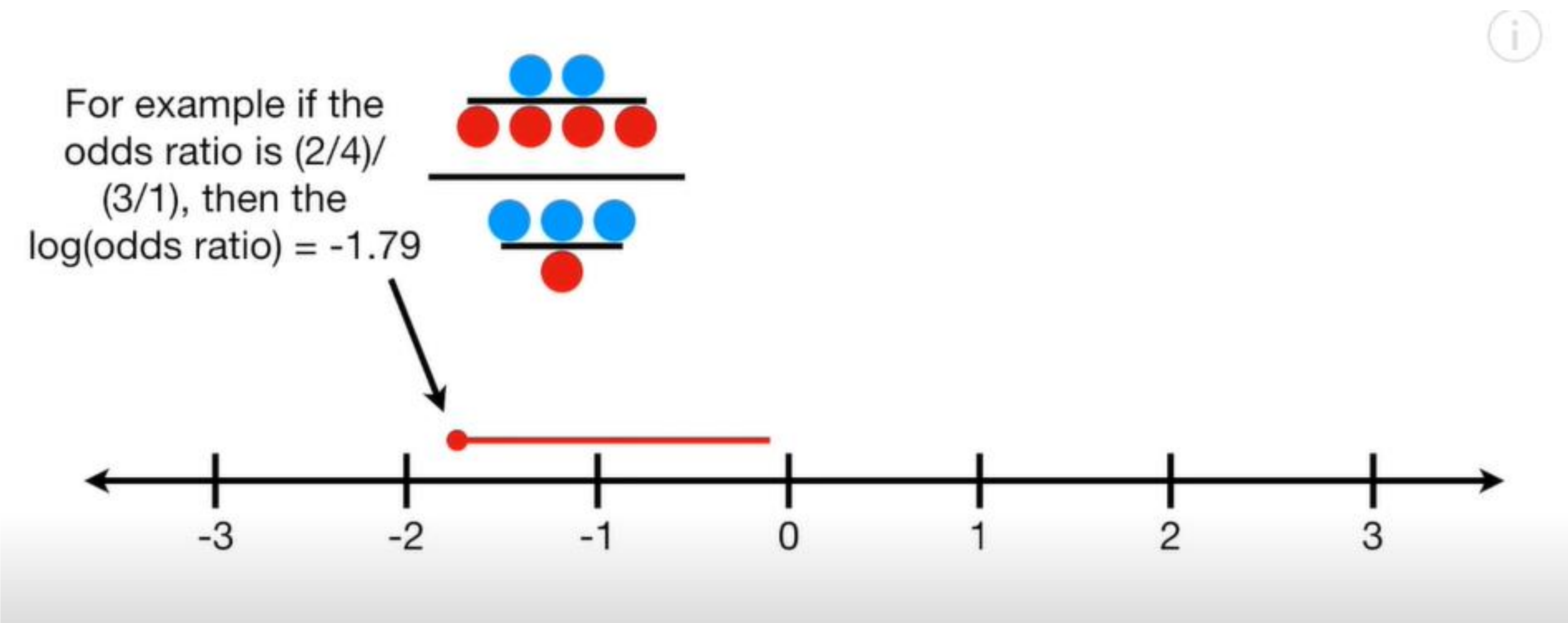
Odds of Winning for Example 1

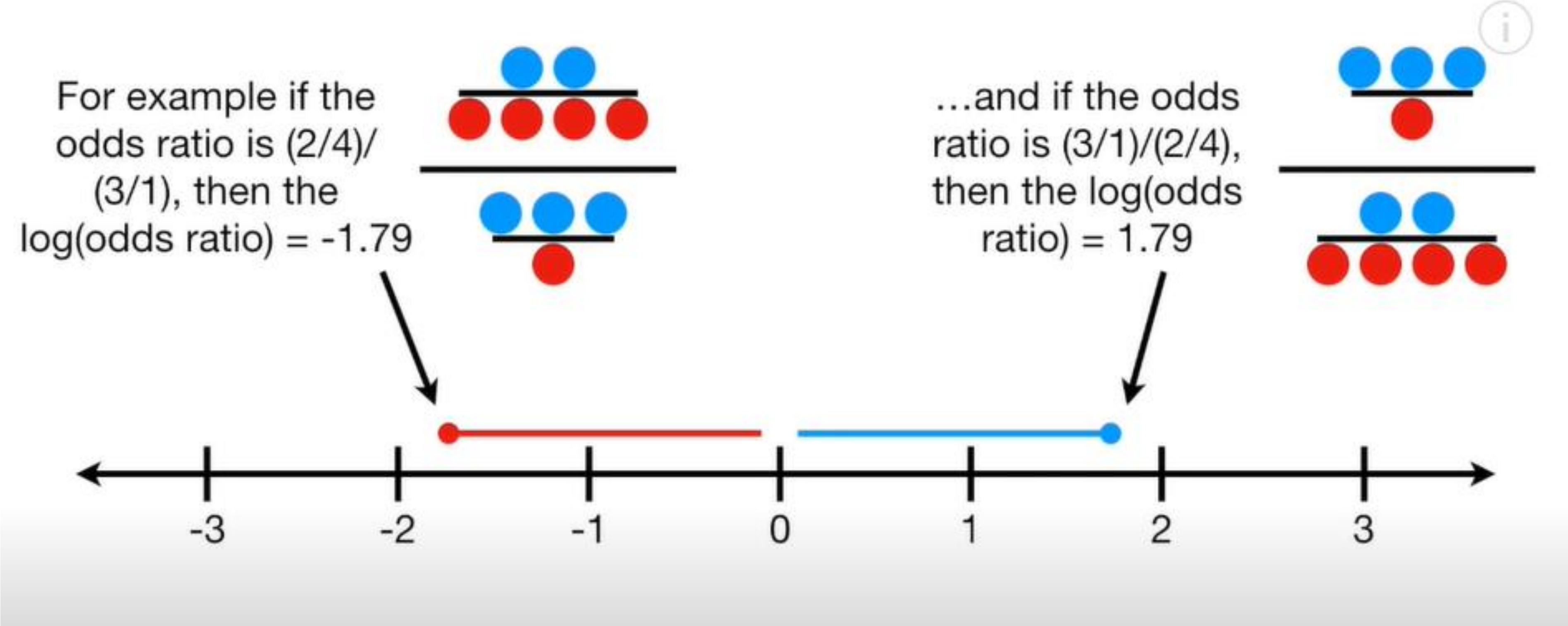


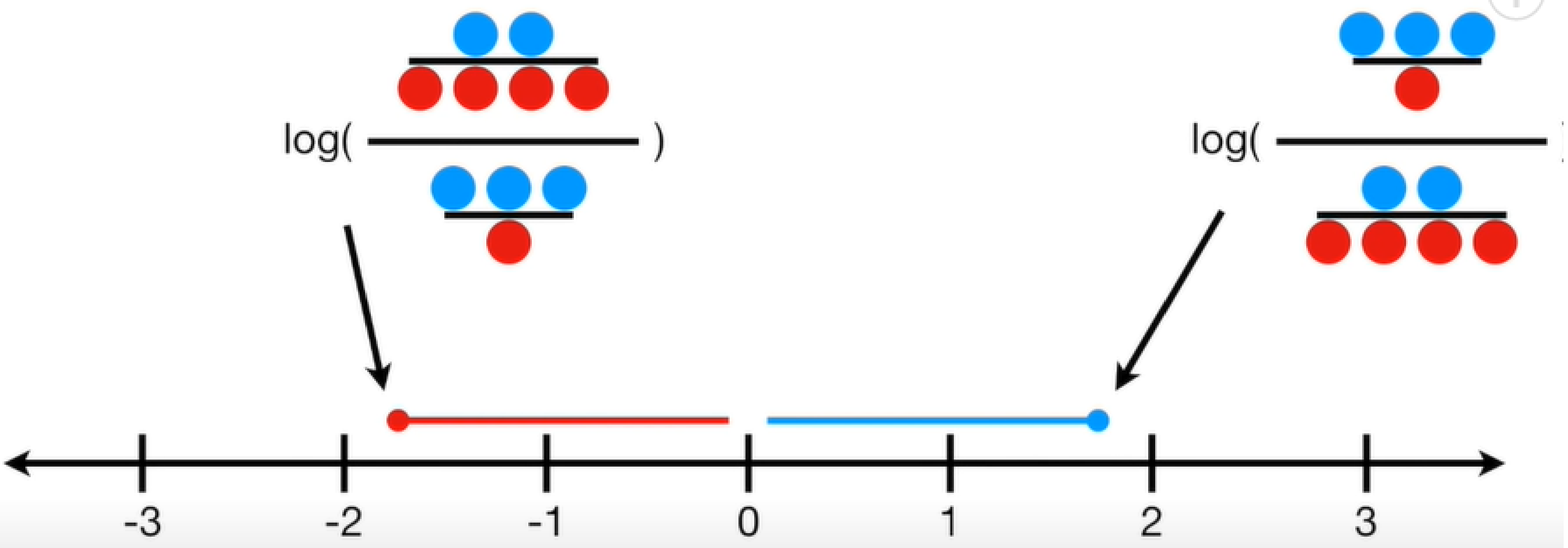
$$\log\left(\frac{\text{Odds of Winning for Example 1}}{\text{Odds of Winning for Example 2}}\right)$$

…and, just like the odds, taking the log of the odds ratio (i.e. log(odds ratio)) makes things nice and symmetrical.




For example if the odds ratio is (2/4)/(3/1), then the log(odds ratio) = -1.79
-3
-2
-1
0
1
2
3


For example if the odds ratio is (2/4)/(3/1), then the log(odds ratio) = -1.79
...and if the odds ratio is (3/1)/(2/4), then the log(odds ratio) = 1.79
-3
-2
-1
0
1
2
3



Great! Now that we've got that cleared up, what can we do with odds ratios?

Example of how to use Odds Ratio
Before doing the example let us understand the
Confusion Matrix

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	**Negative Predictive Value** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$

Predicted Values \ Actual Values	1	0
1	TP	FP
0	FN	TN

Actual Values

1 0

Predicted Values

1

0

Find the relation between Mutated Genes and Persons having Cancer ?
Here we use Odds Ratio to find the relationship

What can we do with Odds Ratio?

Here's an example of the "odds ratio" in action!

Has the mutated gene	Has Cancer: Yes	Has Cancer: No
Yes	23	117
No	6	210

Here's an example of the
"odds ratio" in action!

Find the relation between Mutated Genes and Cancer

Total :356

		Has Cancer: Yes	Has Cancer: No	
Has the mutated gene	Yes	23	117	140 have mutated gene
	No	6	210	216 do not have mutated gene
		29 have cancer	327 Do not have cancer	356

356

Here's an example of the
"odds ratio" in action!

Total :356

		Has Cancer: Yes	Has Cancer: No
Has the mutated gene	Yes	23	117
	No	6	210

140 have mutated gene

216 do not have mutated gene

356

29 have cancer

327 Do not have cancer

356

Here's an example of the "odds ratio" in action!

Total :356

Has the mutated gene	Has Cancer: Yes	Has Cancer: No	
Yes	23	117	140 have mutated gene
No	6	210	216 do not have mutated gene
	29 have cancer	327 Do not have cancer	

356

356

Here’s an example of the
“odds ratio” in action!

Total :356

Has the mutated gene	Has Cancer: Yes	Has Cancer: No
Yes	23	117
No	6	210

29 have cancer

327 Do not have cancer

356

Given person has mutated gene, the odds that they have cancer are
23/117

Given person has mutated gene, the odds that they have cancer are
6/210

Here's an example of the
"odds ratio" in action!

		Has Cancer: Yes	Has Cancer: No
Has the mutated gene	Yes	23	117
	No	6	210

29 have cancer

327 Do not have cancer

$$\frac{23/117}{6/210} = 6.88$$

Odds Ratio is : 6.88

If person has mutated gene then the odds are 6.88 times greater they will have cancer

Here's an example of the
"odds ratio" in action!

		Has Cancer: Yes	Has Cancer: No
Has the mutated gene	Yes	23	117
	No	6	210

29 have cancer

327 Do not have cancer

23/117

6/210 =
0.2/0.03=6.88

We can use the odds ratio to find the relation ship between mutated gene and cancer. If there is mutated gene is the odds higher that person will have cancer.

Odds Ratio is :
23/117//6/210 = 0.2/0.03=6.88
If person has mutated gene then the odds are 6.88 times greater they will have cancere
Log(odds)Ratio
Log (6.88) =1.93

Here's an example of the
"odds ratio" in action!

Total :356

		Has Cancer Yes	Has Cancer No
Has the mutated gene	Yes	23	117
	No	6	210

140 have mutated gene

216 do not have mutated gene

29 have cancer

327 Do not have cancer

Odds ratio
and the log(odds ratio) is
Like R-square.
It tells us the relationship
Between the mutated
gene
And cancer. Large values
mutated genes is a good
predictor of cancer. Small
values the mutated
Genes is not a Good
Predictor of cancer.

We can use the odds ratio to find the relation ship between mutated gene and cancer. If there is mutated gene is the odds higher that person will have cancer.

Given that a person has a mutated gene , that odds that they have cancer are:
23/117
Given a person does not have a mutated gene, the odds that they have cancer:
6/210
Odds Ratio is :
23/117//6/210 = 0.2/0.03=6.88
Log(odds)Ratio
Log (6.88) =1.93

What can we do with Odds Ratio?

Here's an example of the "odds ratio" in action!

		Has Cancer: Yes	Has Cancer: No
Has the mutated gene	Yes	23	117
	No	6	210

Total :356

140 have mutated gene

216 do not have mutated gene

29 have cancer

327 Do not have cancer

Given that a person has a mutated gene , that odds that they have cancer are: 23/117

What can we do with Odds Ratio?

Here's an example of the "odds ratio" in action!

		Has Cancer: Yes	Has Cancer: No
Has the mutated gene	Yes	23	117
	No	6	210

29 have cancer

327 Do not have cancer

Total :356

140 have mutated gene

216 do not have mutated gene

We can use the odds ratio to find the relation ship between mutated gene and cancer

What can we do with Odds Ratio?

Here's an example of the "odds ratio" in action!

		Has Cancer: Yes	Has Cancer: No
Has the mutated gene	Yes	23	117
	No	6	210

Total :356

140 have mutated gene

216 do not have mutated gene

29 have cancer

327 Do not have cancer

We can use the odds ratio to find the relation ship between mutated gene and cancer

What can we do with Odds Ratio?

Here's an example of the "odds ratio" in action!

		Has Cancer	
		Yes	No
Has the mutated gene	Yes	23	117
	No	6	210

29 have cancer

327 Do not have cancer

Total :356

140 have mutated gene

216 do not have mutated gene

We can use the odds ratio to find the relation ship between mutated gene and cancer