DATA MINING 2
Odds and Log Odds

Riccardo Guidotti

a.a. 2021/2022

Contains edited slides from StatQuest

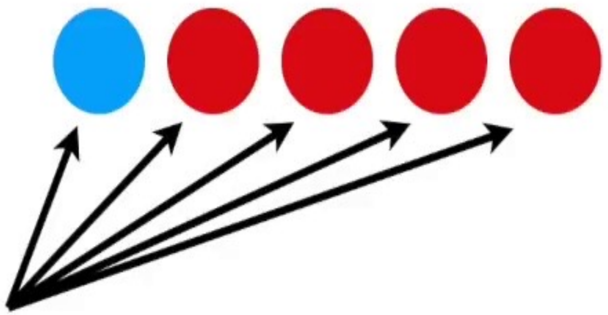
UNIVERSITÀ DI PISA

Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

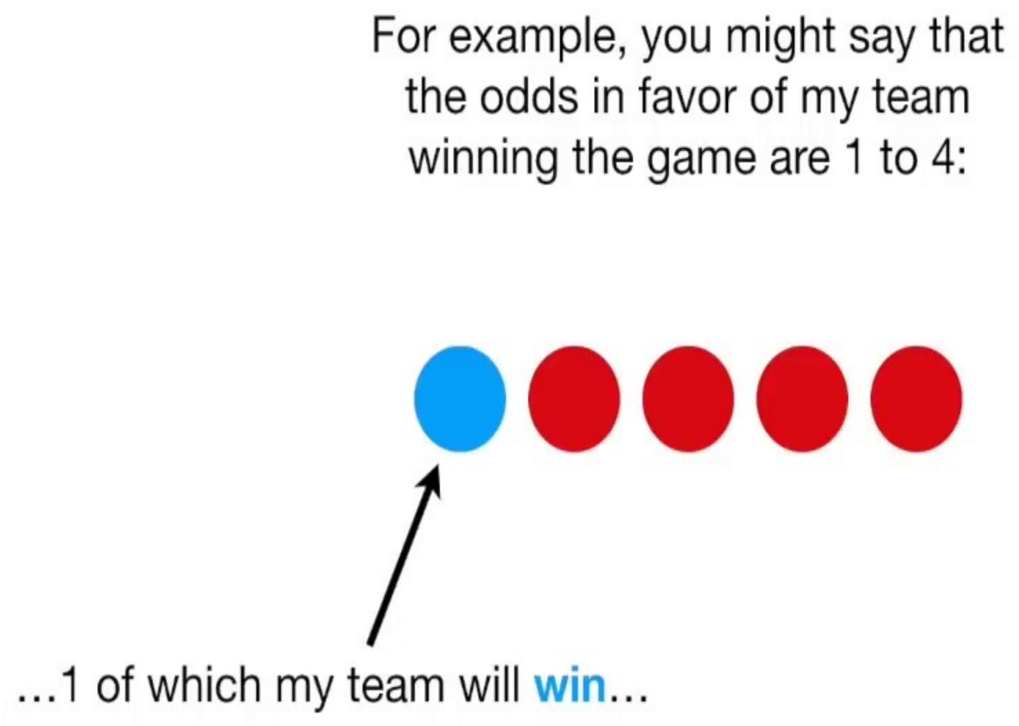
Odds Example

For example, you might say that
the odds in favor of my team
winning the game are 1 to 4:

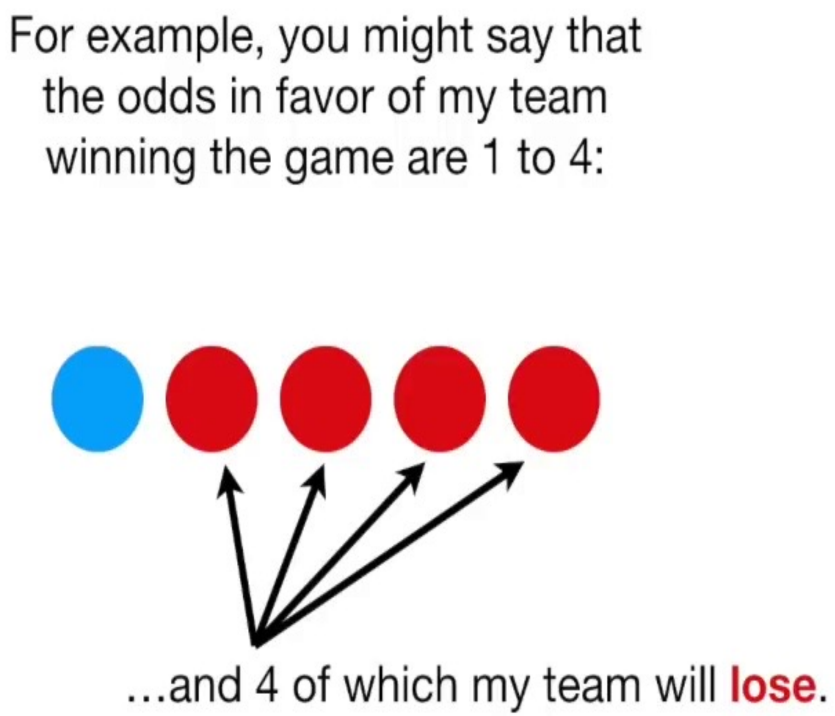


Visually, we have 5 games total…

Odds Example

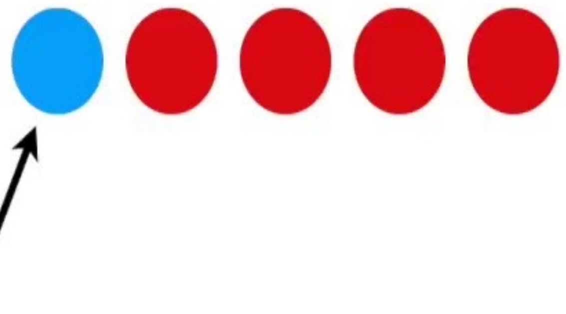


Odds Example



Odds Example

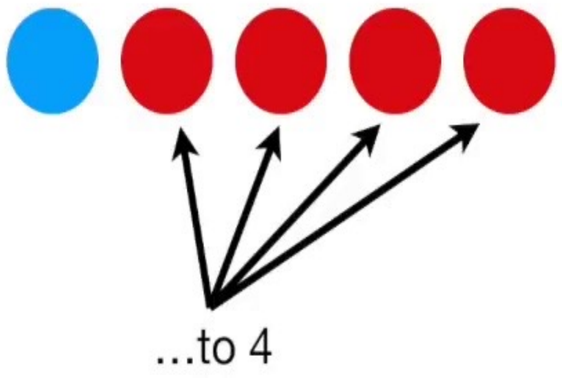
For example, you might say that the odds in favor of my team winning the game are 1 to 4:



So the odds are 1...

Odds Example

For example, you might say that
the odds in favor of my team
winning the game are 1 to 4:



Odds Example

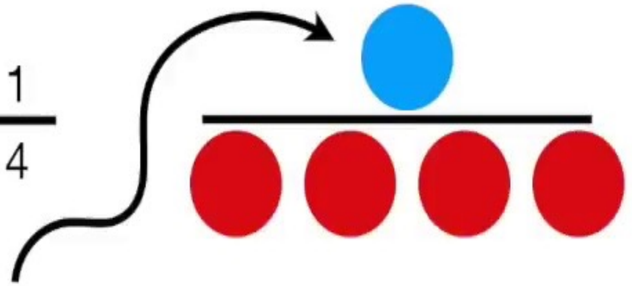
For example, you might say that the odds in favor of my team winning the game are 1 to 4:

Alternatively, we can write this as a fraction… $\frac{1}{4}$

Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

Alternatively, we can write this as a fraction... $\frac{1}{4}$

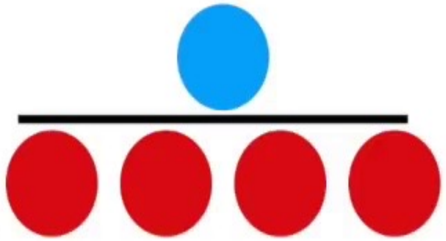


Visually, we have the 1 game my team **wins**...

Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

Alternatively, we can write this as a fraction… $\frac{1}{4}$



…divided by the 4 games that my team **loses**.

Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

Alternatively, we can write this as a fraction… $\frac{1}{4} = 0.25$

…if we do the math, we see that the odds are 0.25 that my team will win the game.

Odds Example

Here's another example: You might say that the odds in favor of my team winning the game are 5 to 3:

Alternatively, we can write this as a fraction… $\frac{5}{3} = 1.7$

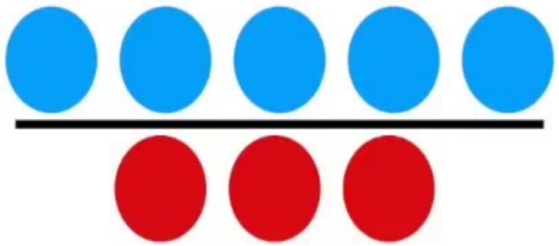
…if we do the math, we see that the odds are 1.7 that my team will win the game.

Note: Odds are not probabilities!!!

Odds vs Probability

The odds are the ratio of something happening (i.e. my team **winning**)...

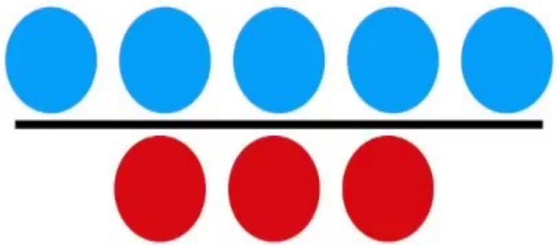
...to something not happening (i.e. my team **not winning**).



Odds vs Probability

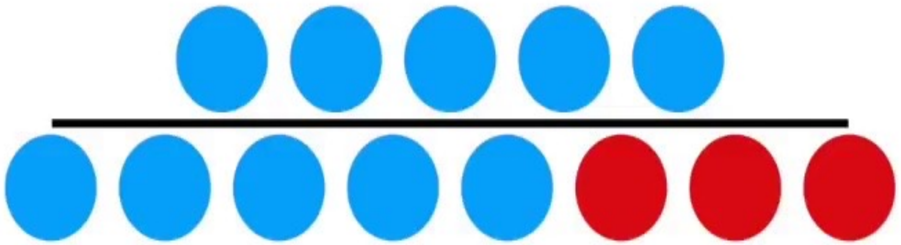
The odds are the ratio of something happening (i.e. my team **winning**)…

…to something not happening (i.e. my team **not winning**).

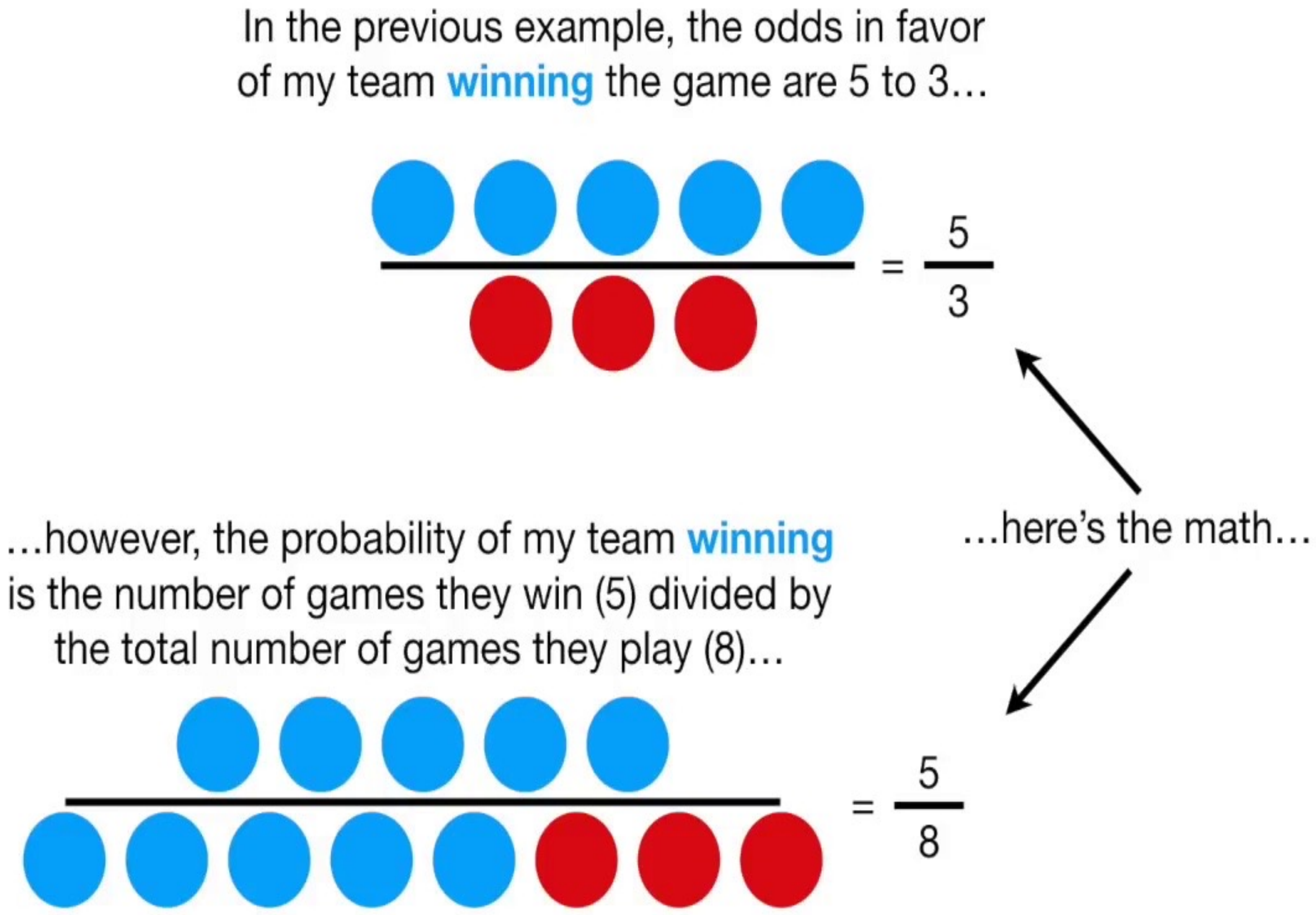


Probability is the ratio of something happening (i.e. my team **winning**)…

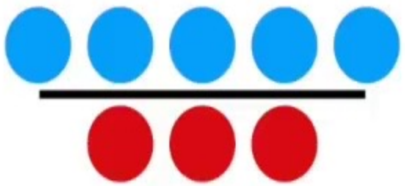
…to *everything* that could happen (i.e. my team **winning** *and* **losing**).



Odds vs Probability



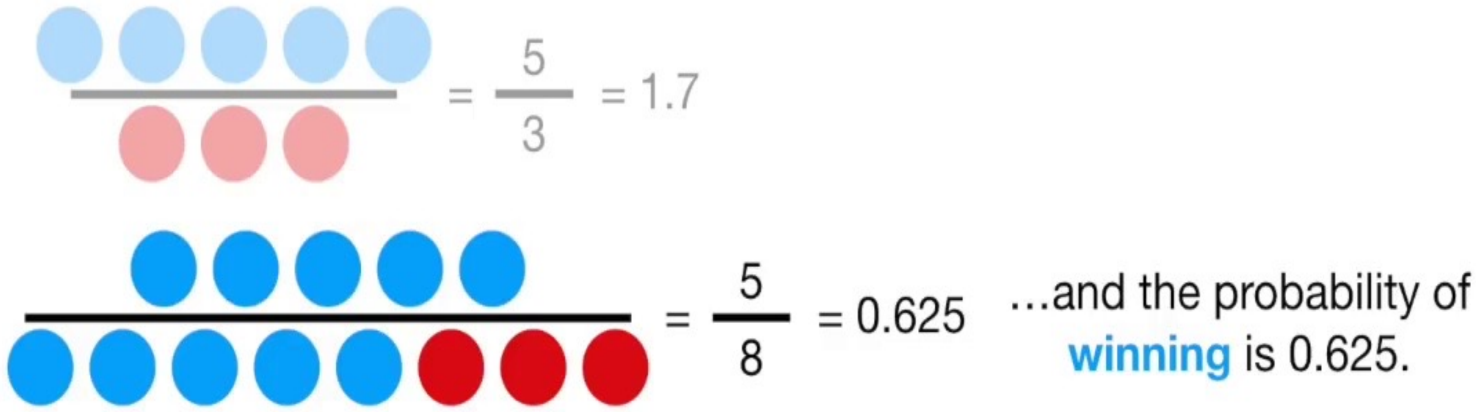
Odds from Probabilities



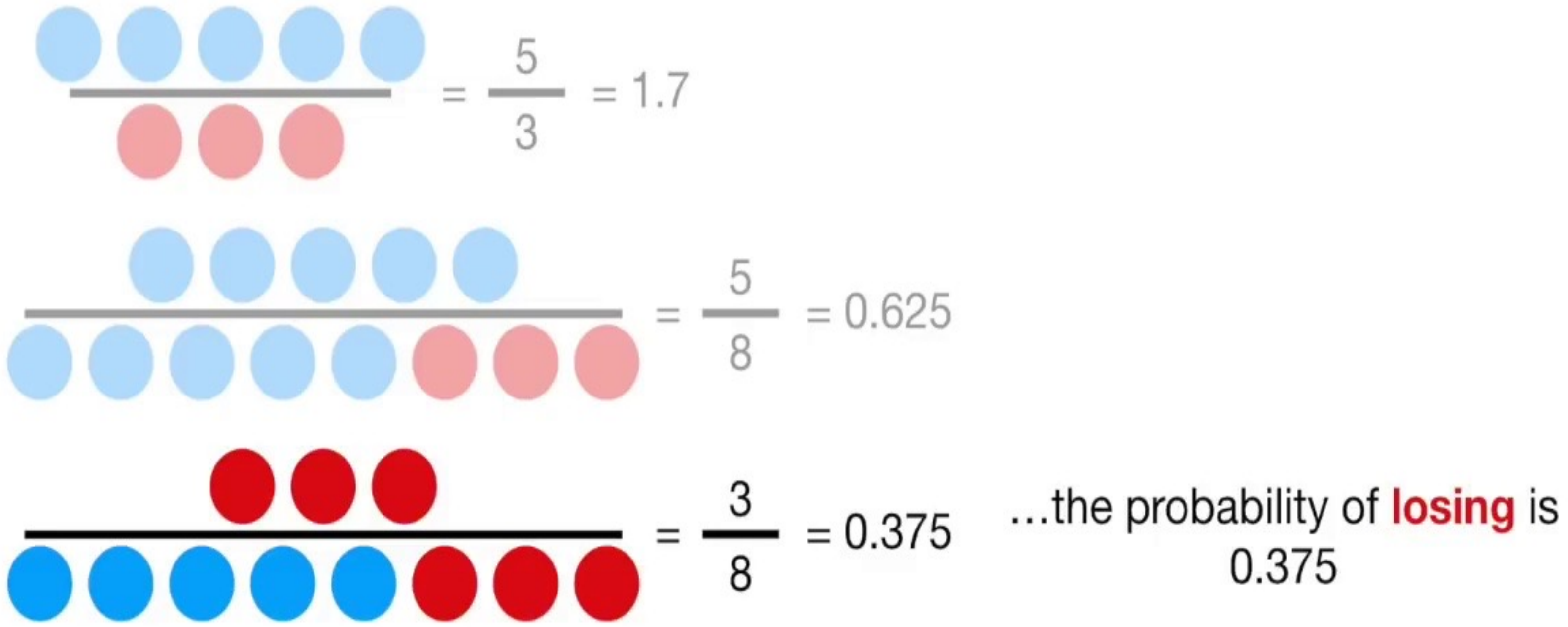
$$= \frac{5}{3} = 1.7$$

In the last example we saw that the odds of **winning** are 1.7…

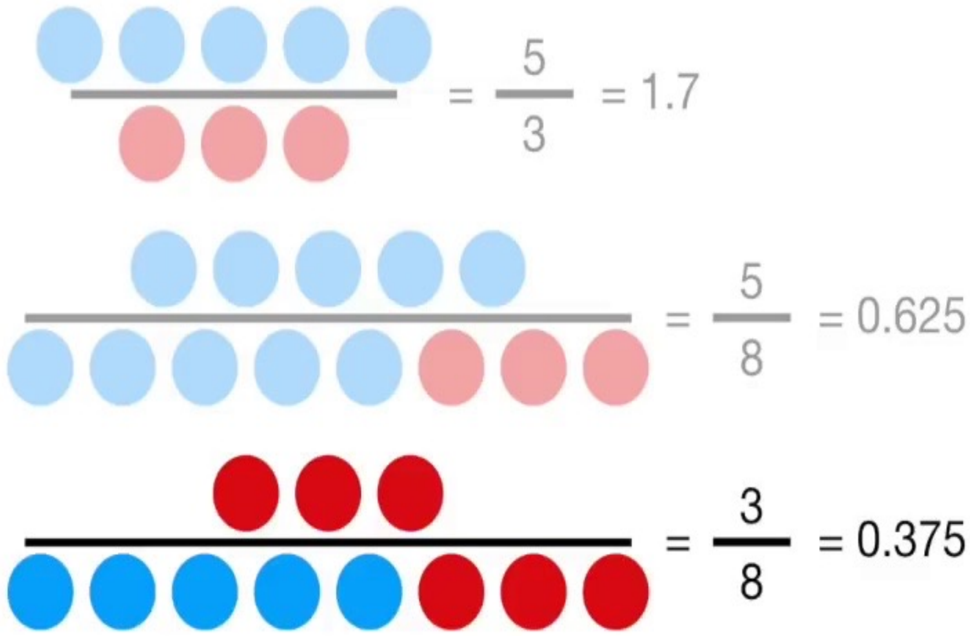
Odds from Probabilities



Odds from Probabilities



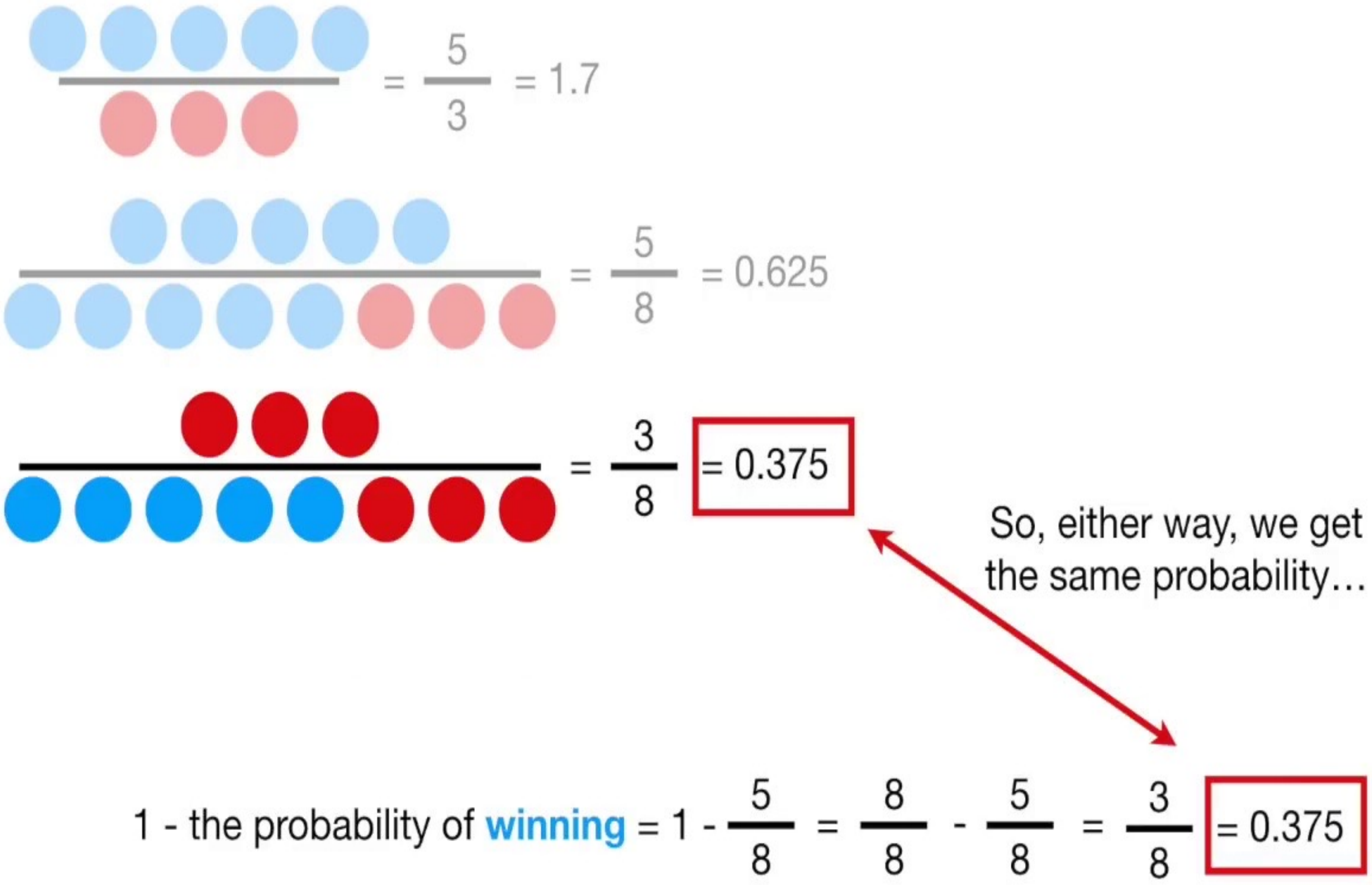
Odds from Probabilities



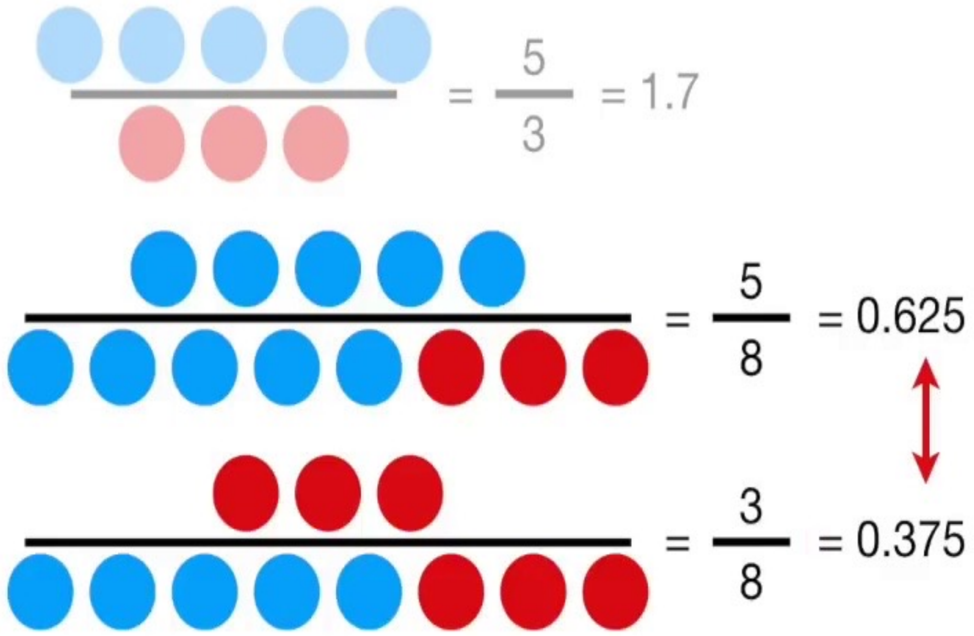
NOTE: We could also calculate the probability of **losing** as:

1 - the probability of **winning**

Odds from Probabilities



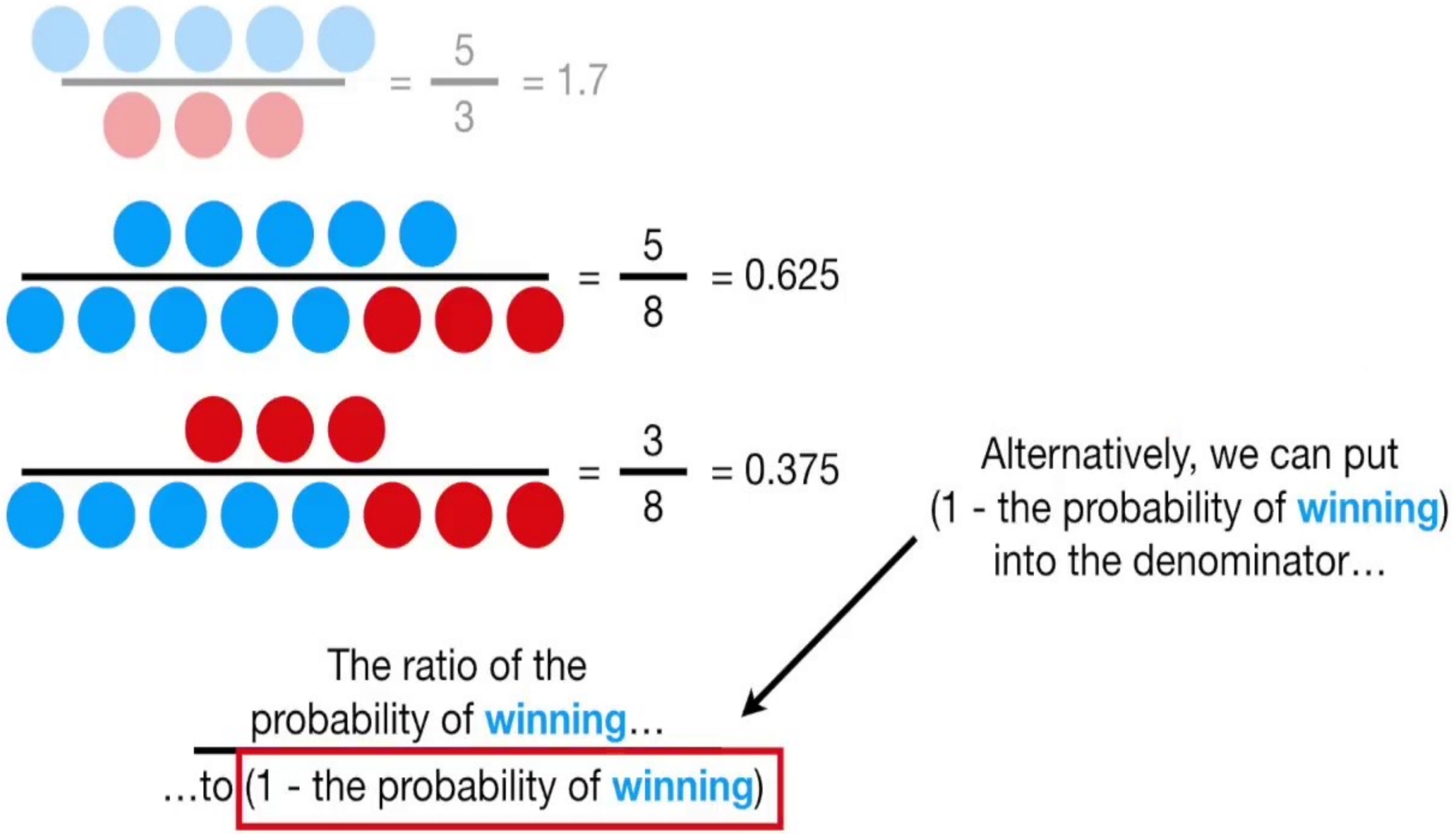
Odds from Probabilities



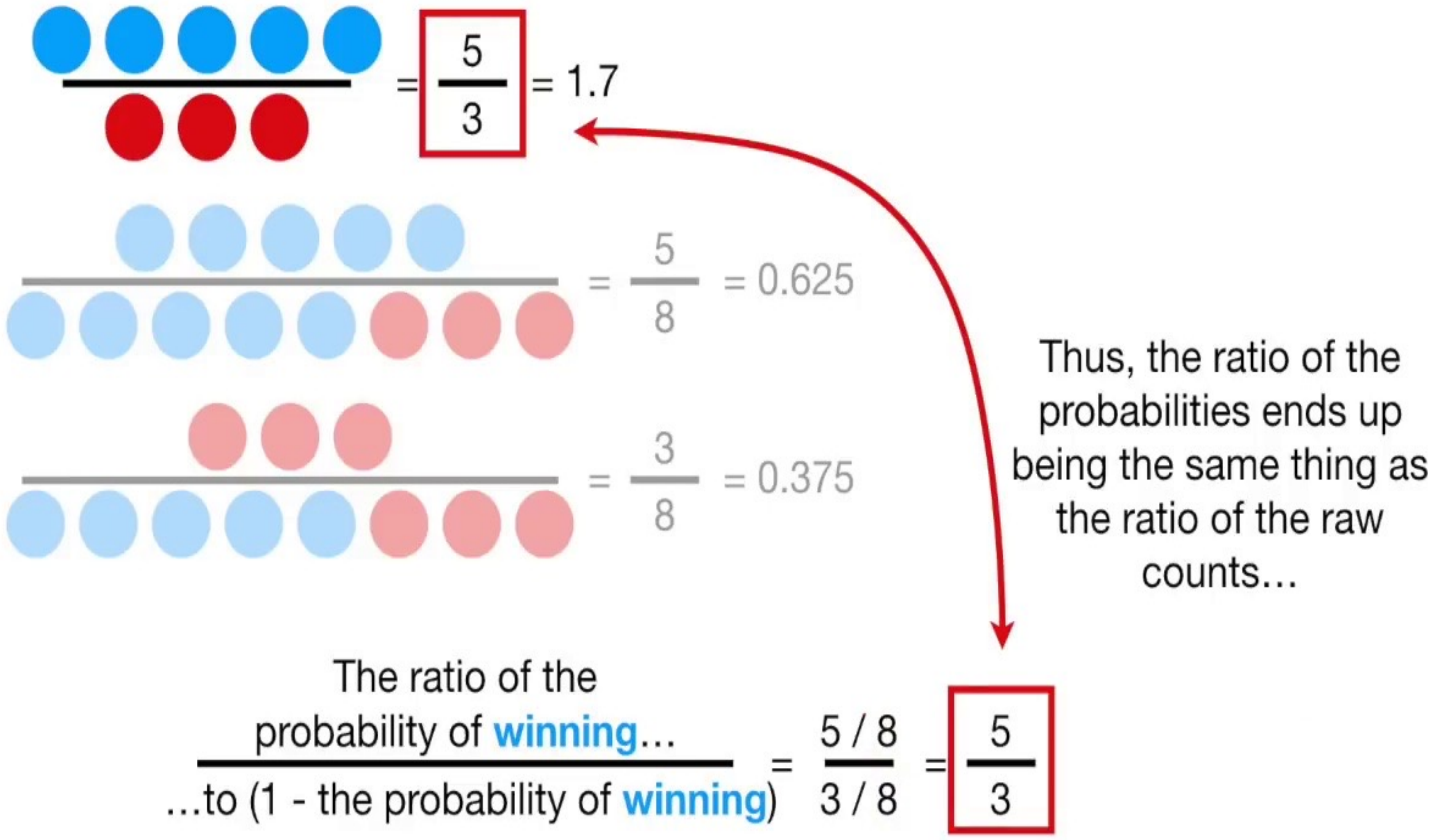
Now let's take the ratio of the probability of **winning** to the probability of **losing**…

$$\frac{\text{The ratio of the probability of \textbf{winning}…}}{\text{…to the probability of \textbf{losing}}}$$

Odds from Probabilities



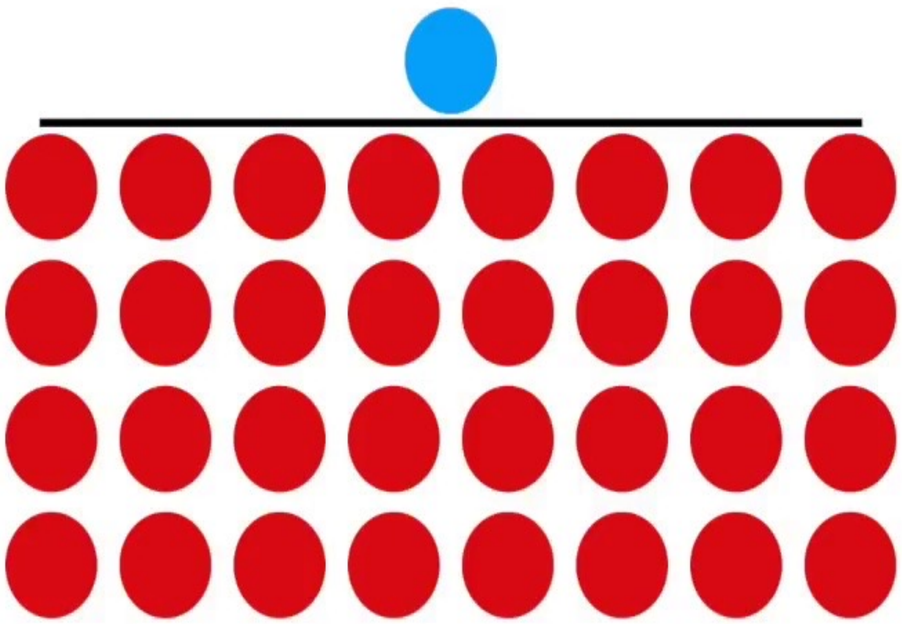
Odds from Probabilities



odds = p/(1-p)

Log of the Odds

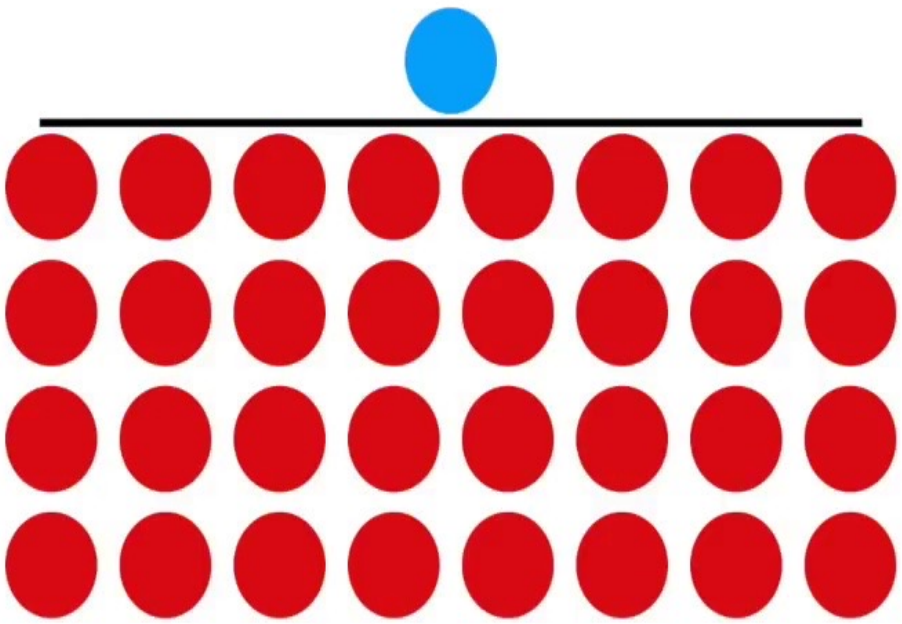
We can see that the worse my team is, the odds of **winning** get closer and closer to 0.



$$= \frac{1}{32} = 0.031$$

Log of the Odds

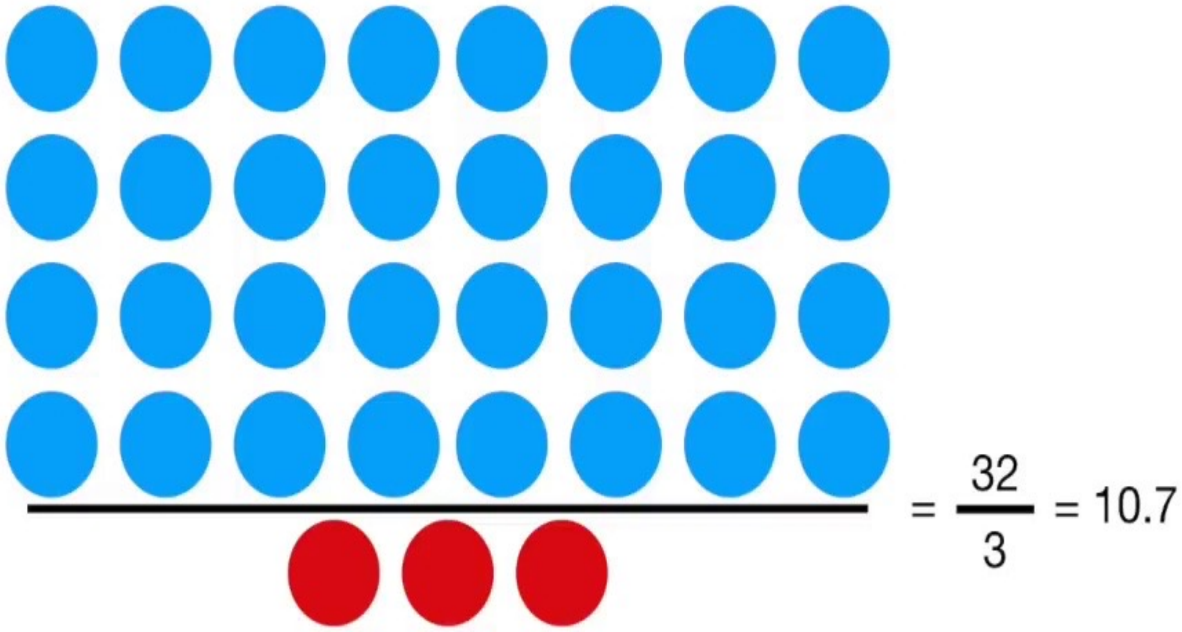
In other words, if the odds are *against* my team **winning**, then they will be between 0 and 1.



$$= \frac{1}{32} = 0.031$$

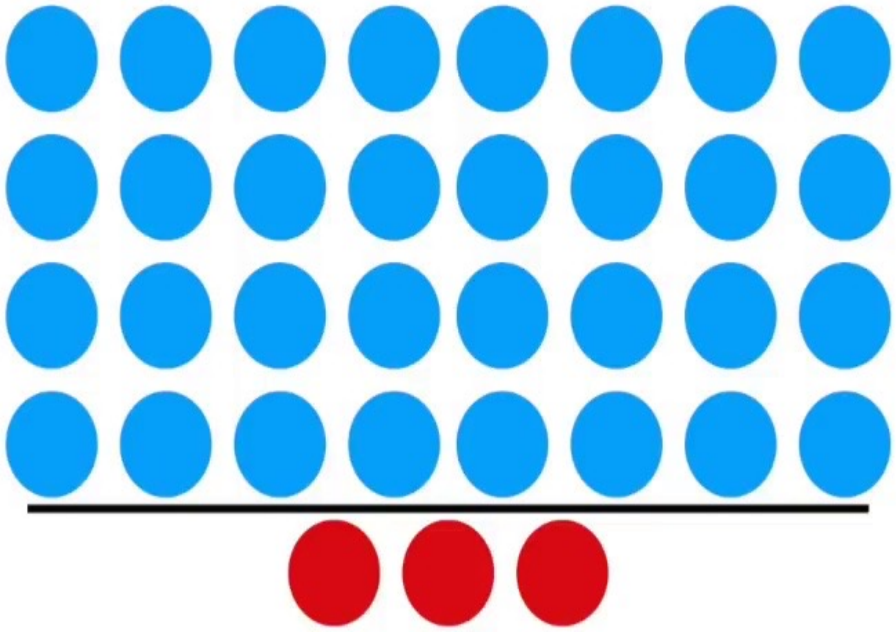
Log of the Odds

We can see that the better my team is, the odds of **winning** start at 1 and just go up and up.



Log of the Odds

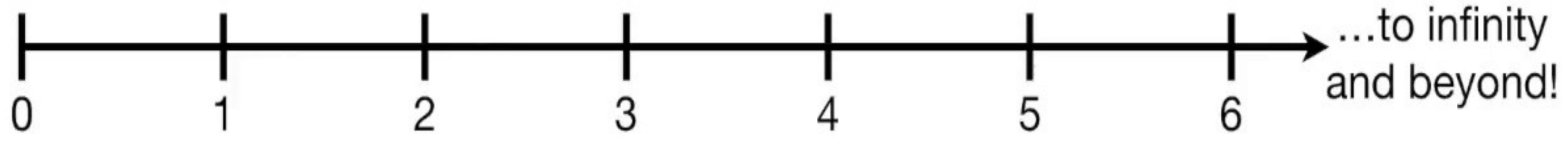
In other words, if the odds are *for* my team **winning**,
then they will be between 1 and infinity!



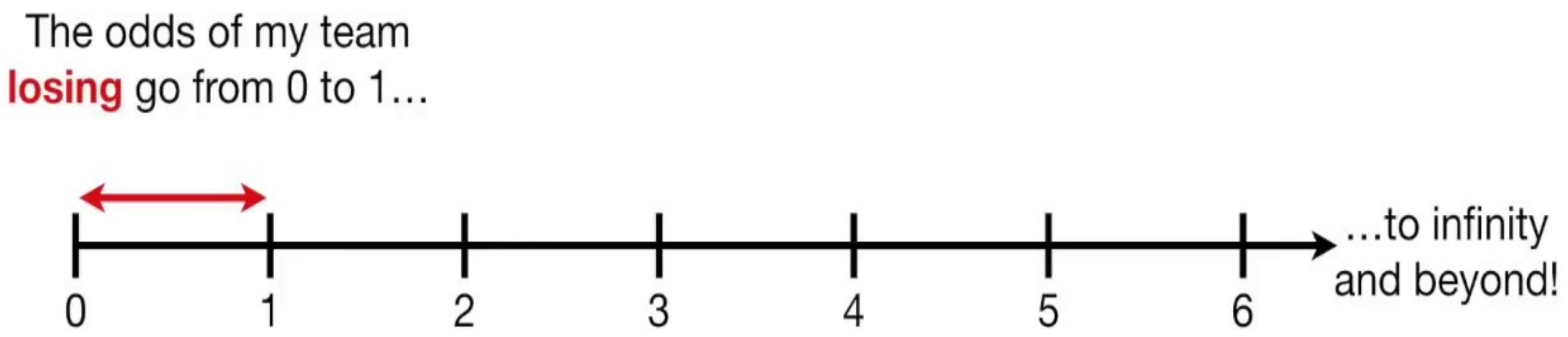
$$= \frac{32}{3} = 10.7$$

Log of the Odds

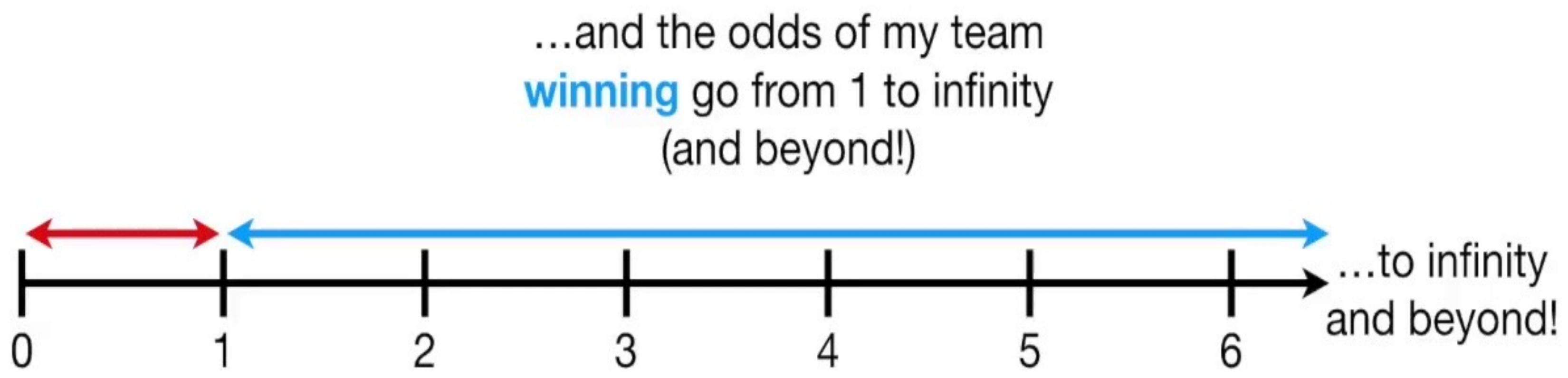
Another way to look at this is with a number line…



Log of the Odds

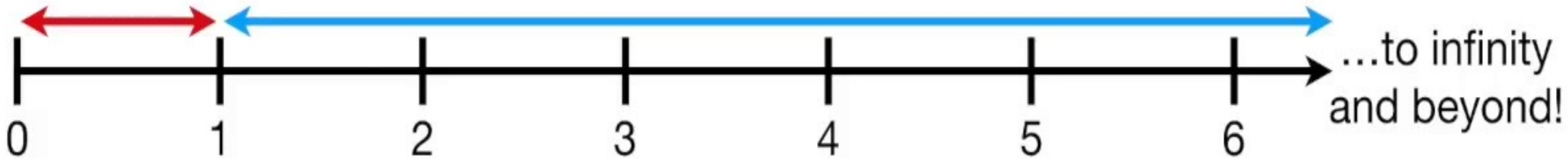


Log of the Odds

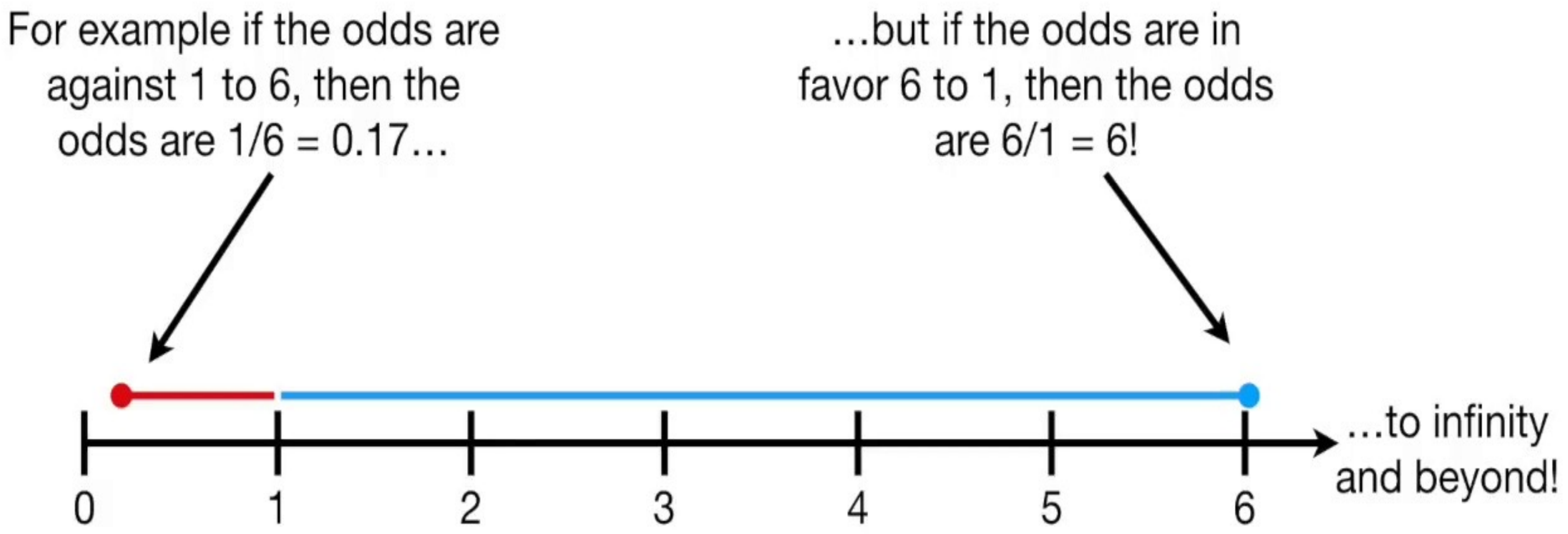


Log of the Odds

The asymmetry makes it difficult to compare the odds for or against my team winning.

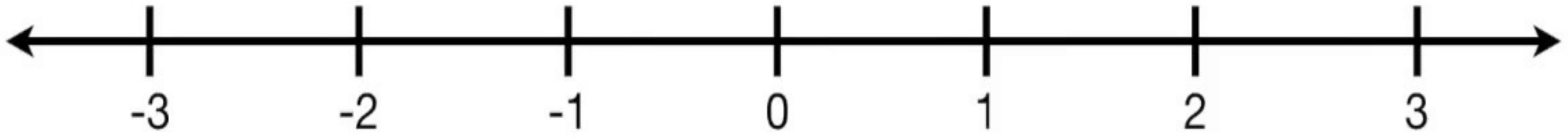


Log of the Odds

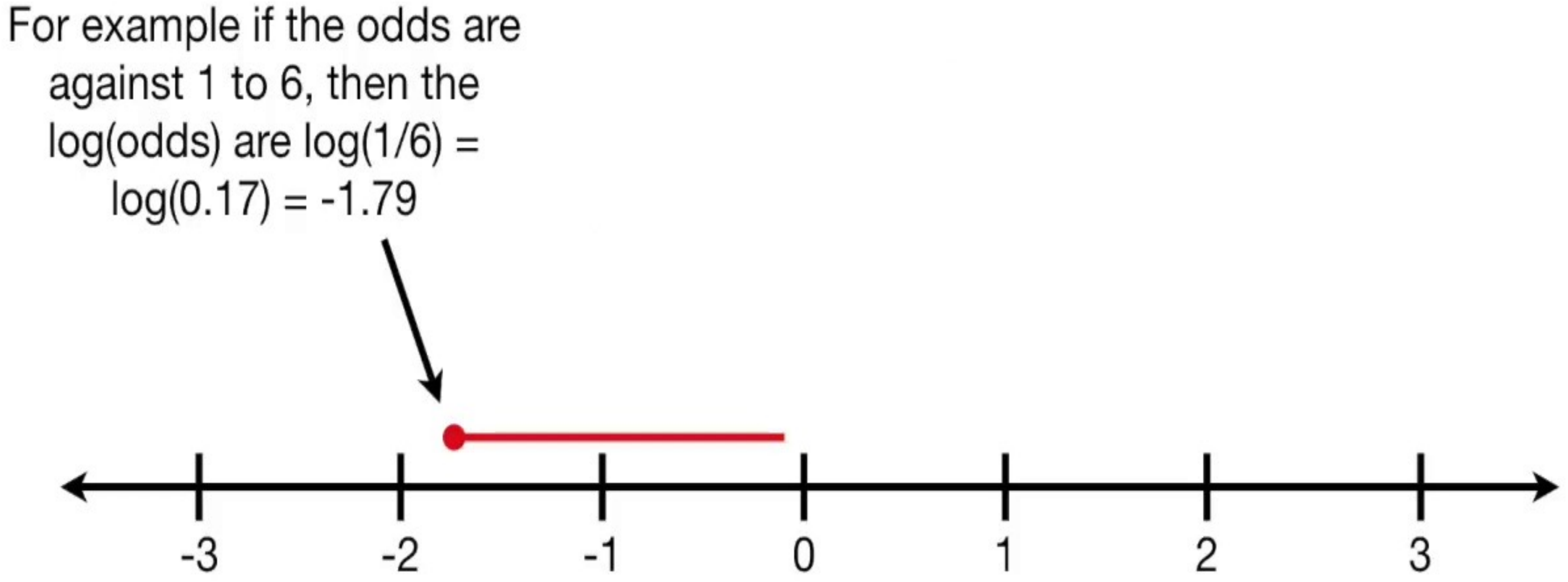


Log of the Odds

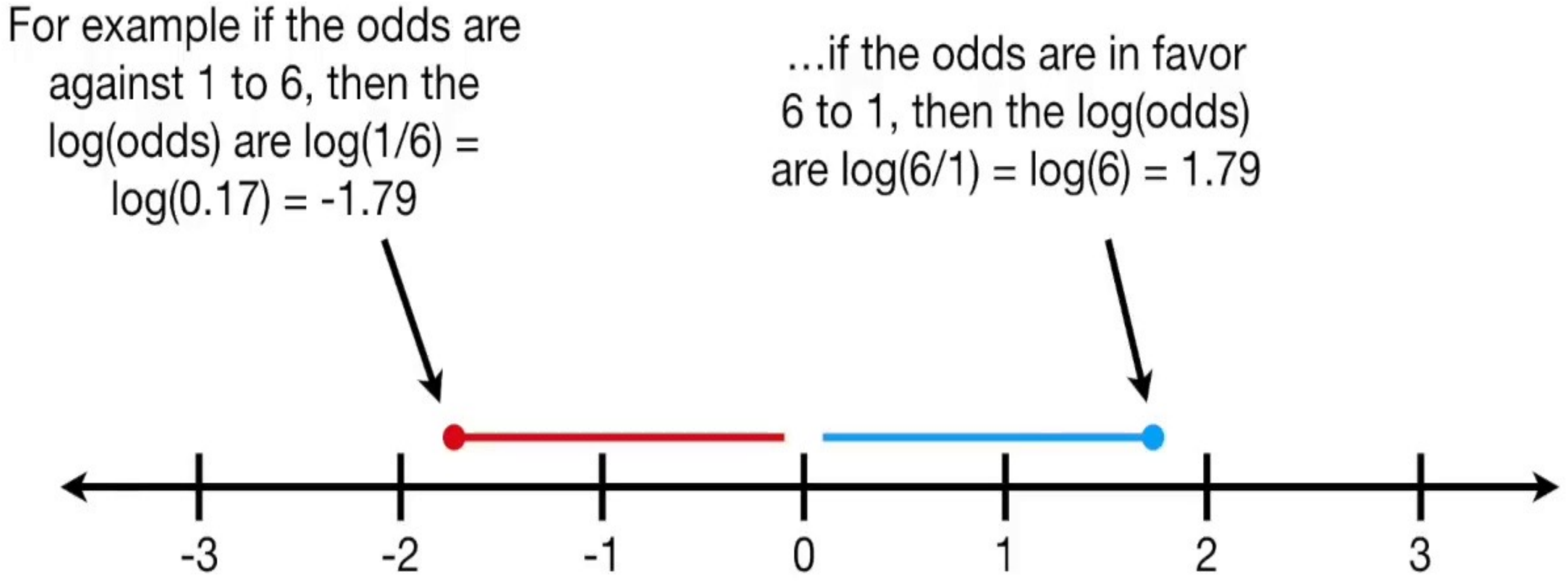
Taking the log() of the odds (i.e. log(odds)) solves this problem by making everything symmetrical.



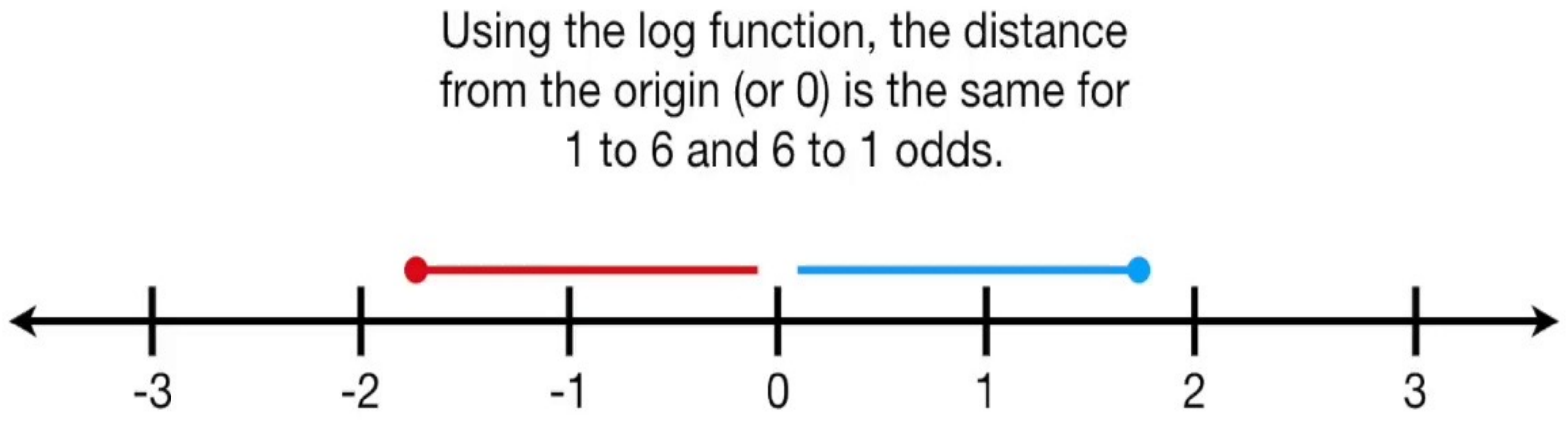
Log of the Odds



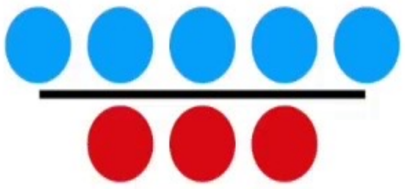
Log of the Odds



Log of the Odds



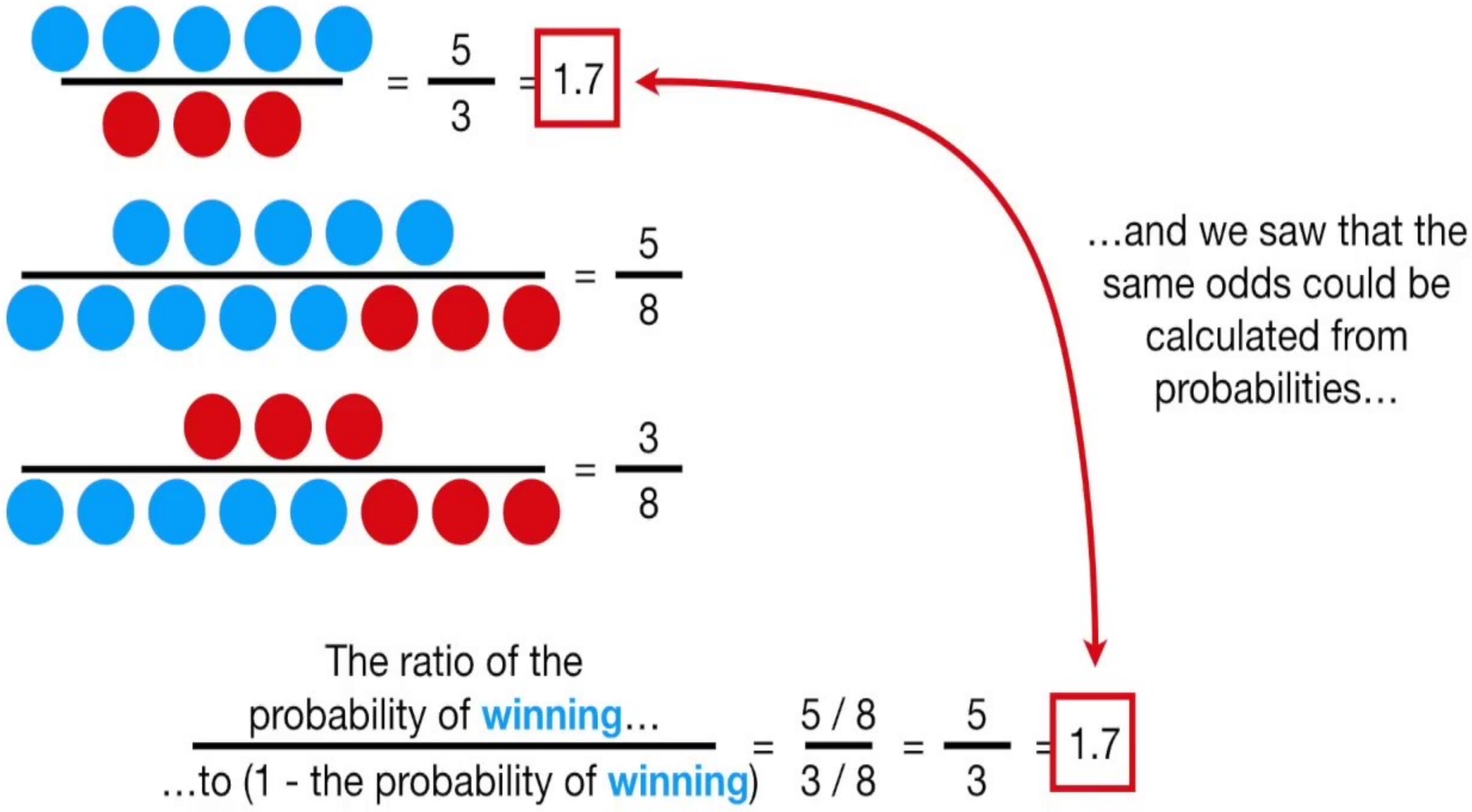
Odds and Log Odds



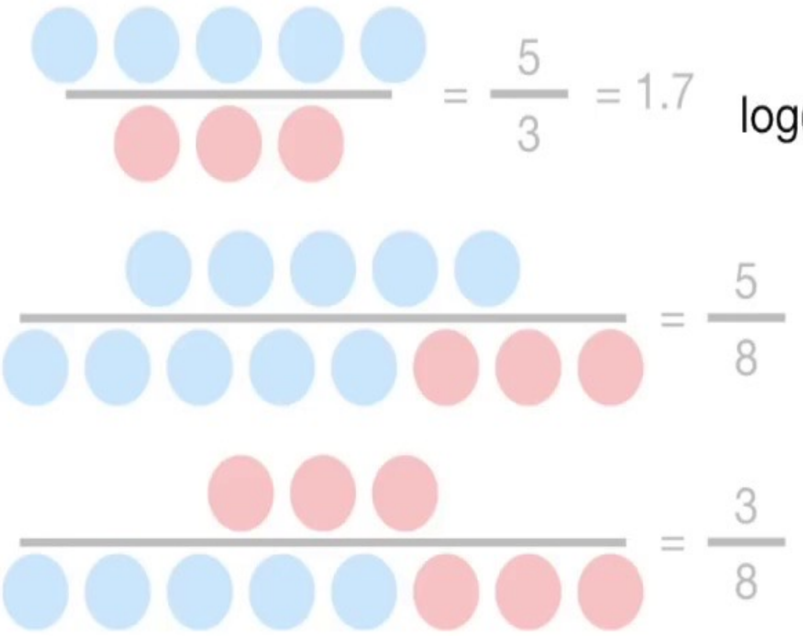
$$= \frac{5}{3} = 1.7$$

Earlier we saw that odds can be calculated from counts...

Odds and Log Odds



Odds and Log Odds

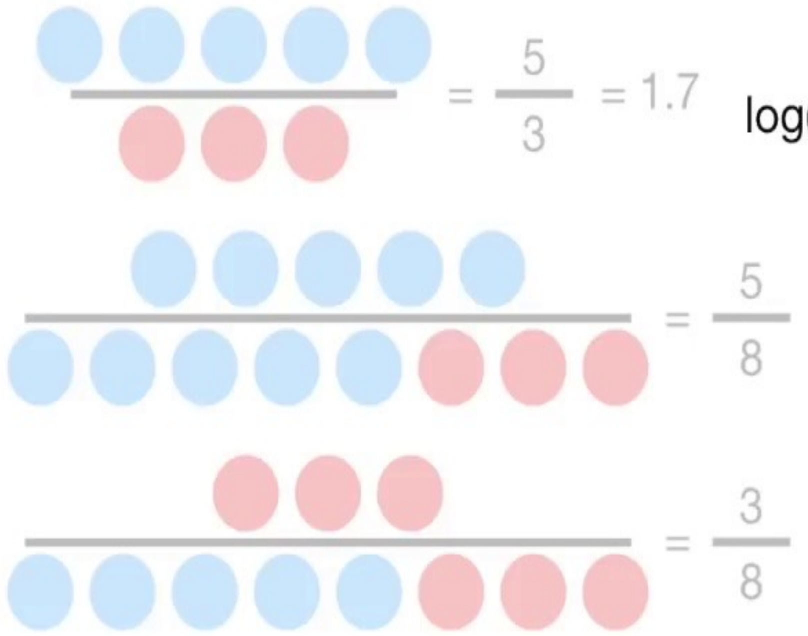


$$\log(\text{odds}) = \log\left(\frac{5}{3}\right) = \log\left(\frac{p}{(1-p)}\right) = \log(1.7) = 0.53$$

…and that means we can calculate the log of the odds with counts or probabilities - either way, we'll get the same value.

$$\frac{\text{The ratio of the probability of winning…}}{\text{…to (1 - the probability of winning)}} = \frac{5/8}{3/8} = \frac{5}{3} = 1.7$$

Odds and Log Odds



$$\log(\text{odds}) = \log\left(\frac{5}{3}\right) = \boxed{\log\left(\frac{p}{(1-p)}\right)} = \log(1.7) = 0.53$$

NOTE: The log of the ratio of the probabilities is called the **logit function** and forms the basis for logistic regression.

$$\frac{\text{The ratio of the probability of winning...}}{\text{...to (1 - the probability of winning)}} = \frac{5/8}{3/8} = \frac{5}{3} = 1.7$$

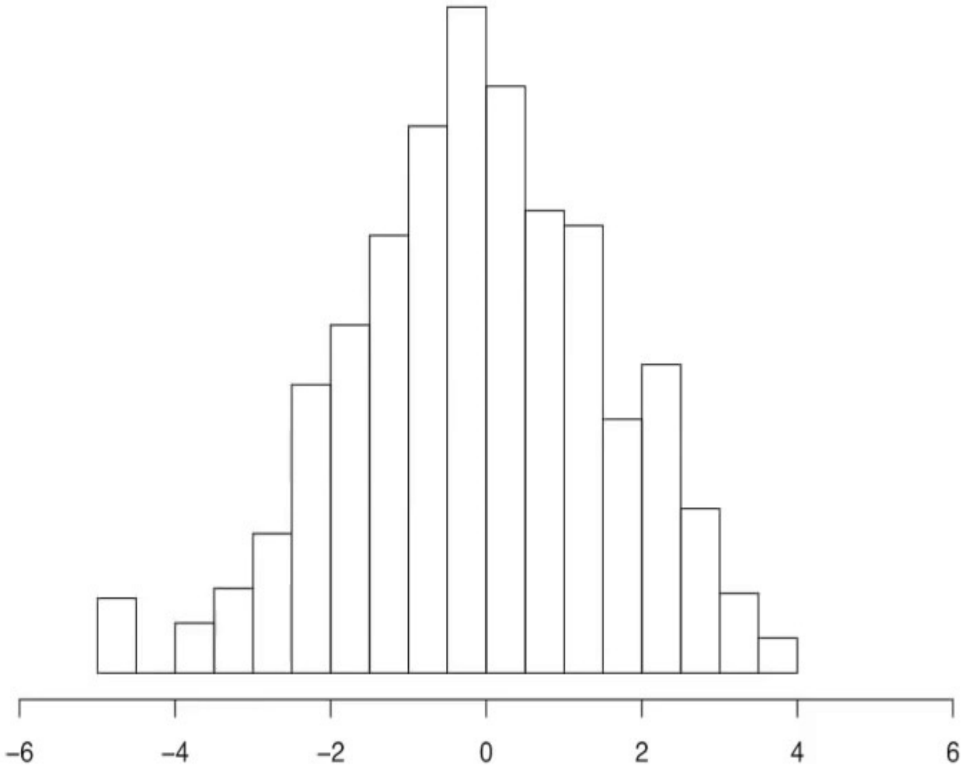
Odds and Log Odds

- Odds are the ratio of something happening to something not happening
- Log odds are the log of the odds

- What’s the big deal?

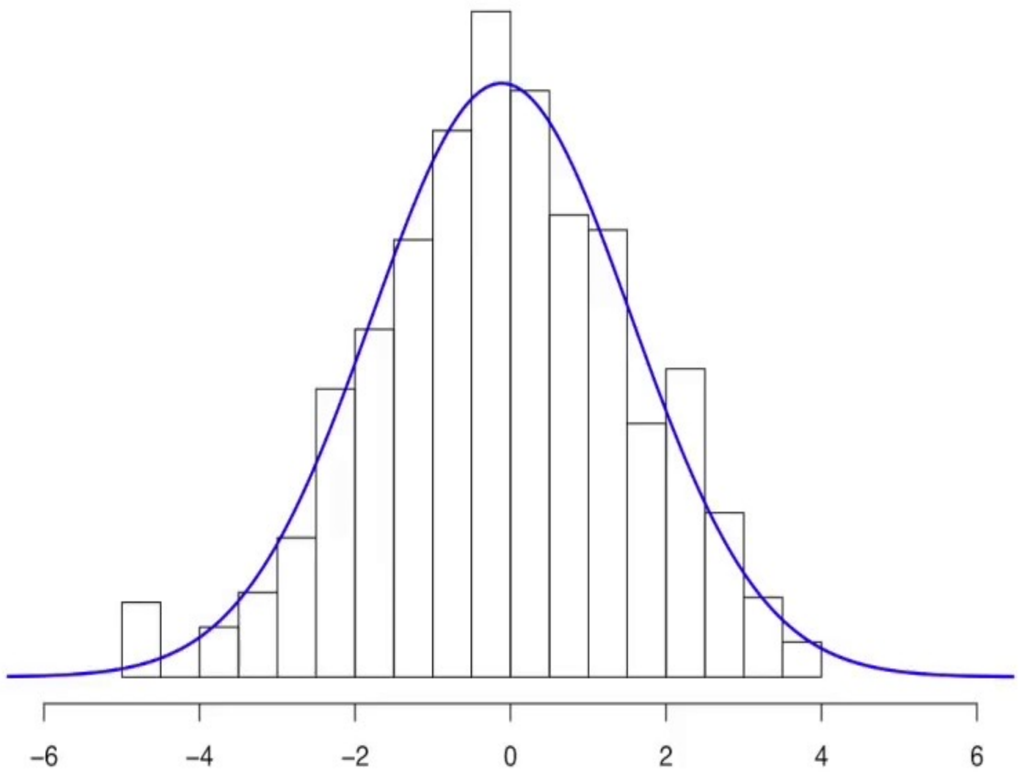
Odds and Log Odds

To show you what the big deal is all about, if I pick pairs of random numbers that add up to 100 (for example) and use them to calculate the log(odds) and draw a histogram…



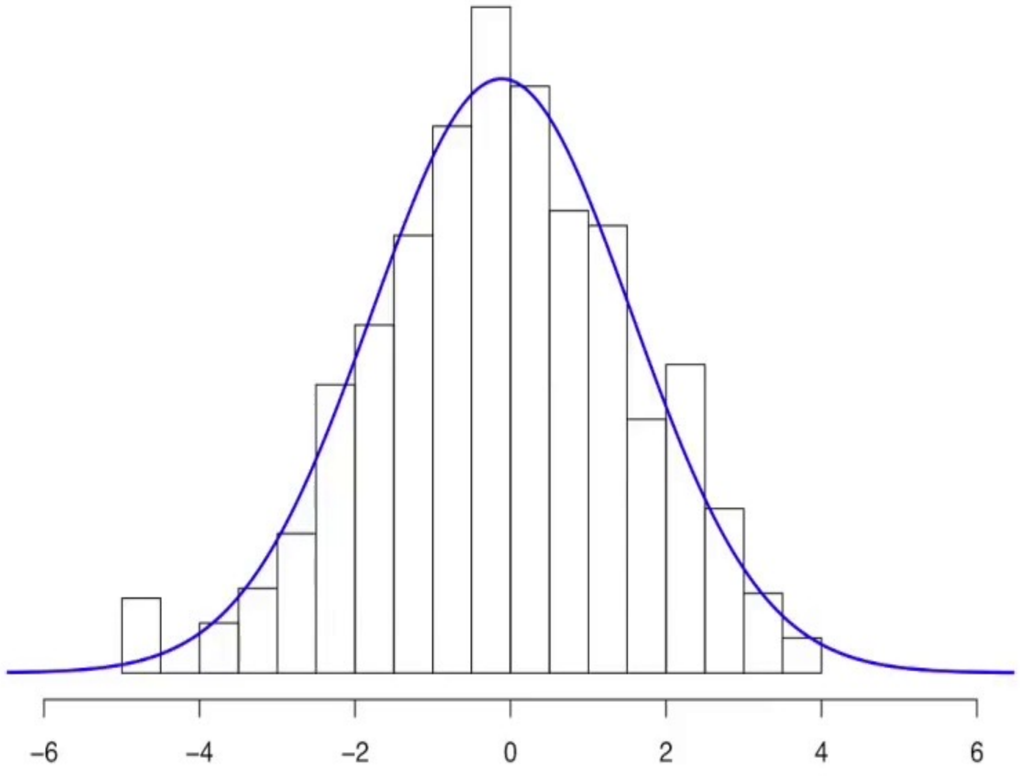
Odds and Log Odds

...the histogram is in the shape of a normal distribution!



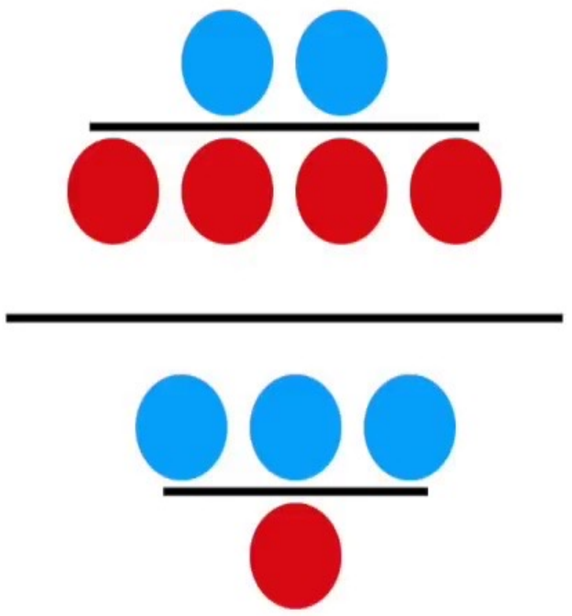
Odds and Log Odds

This makes the log(odds) useful for solving certain statistics problems - specifically ones where we are trying to determine probabilities about win/lose, or yes/no, or true/false types of situations.



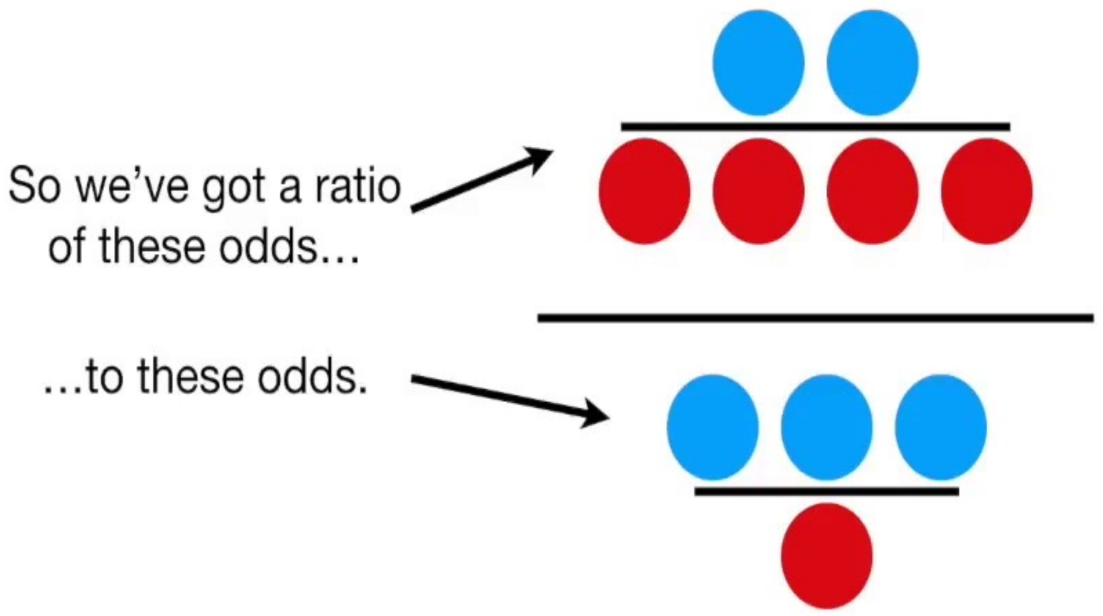
Odds Ratios

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

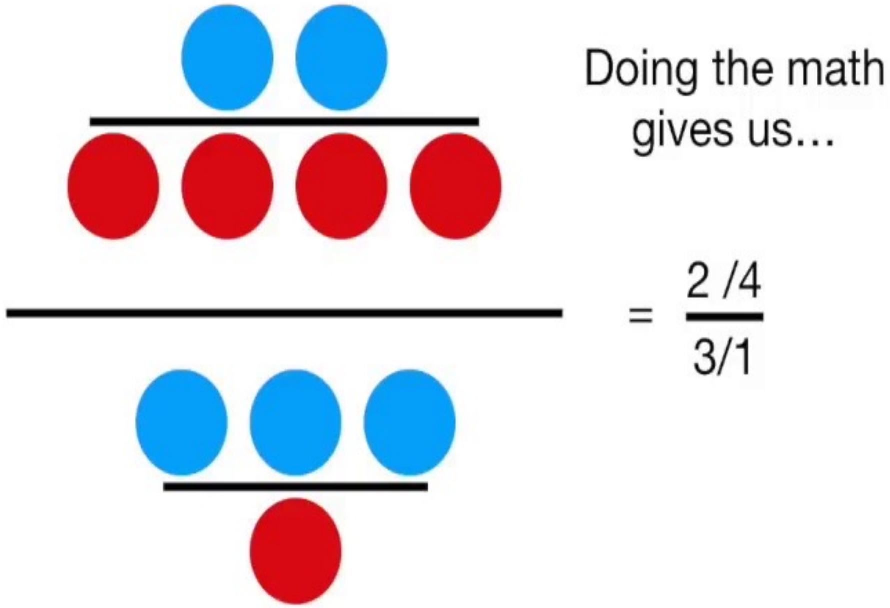


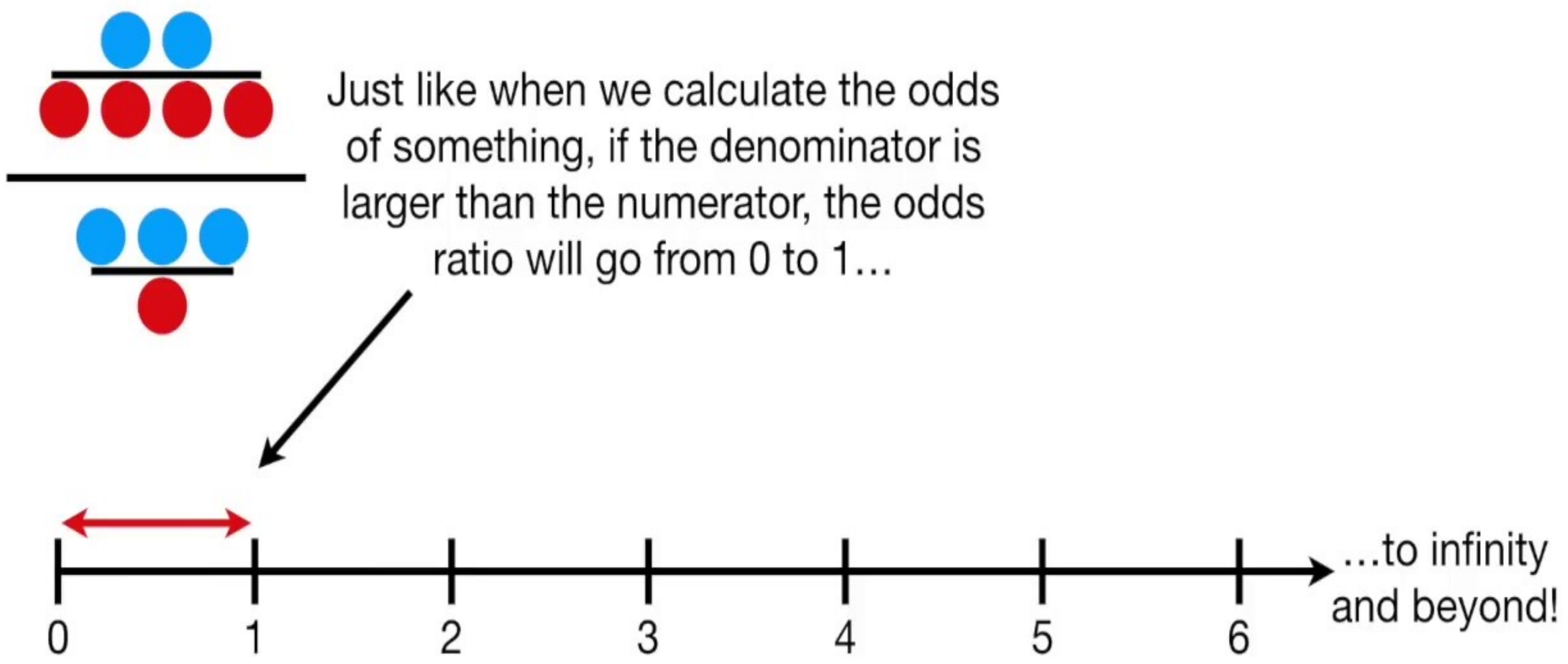
Odds Ratios

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

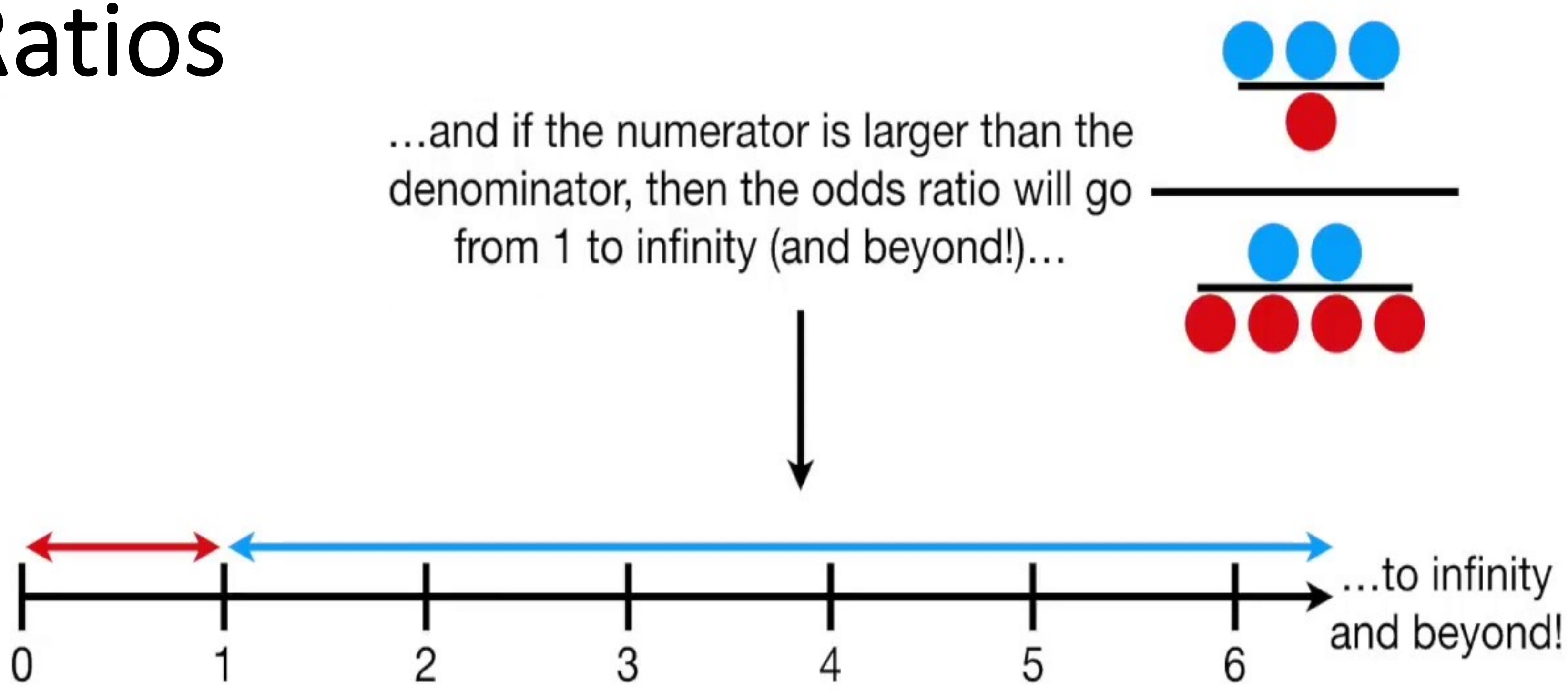


Odds Ratios




Just like when we calculate the odds of something, if the denominator is larger than the numerator, the odds ratio will go from 0 to 1…
0
1
2
3
4
5
6
…to infinity and beyond!

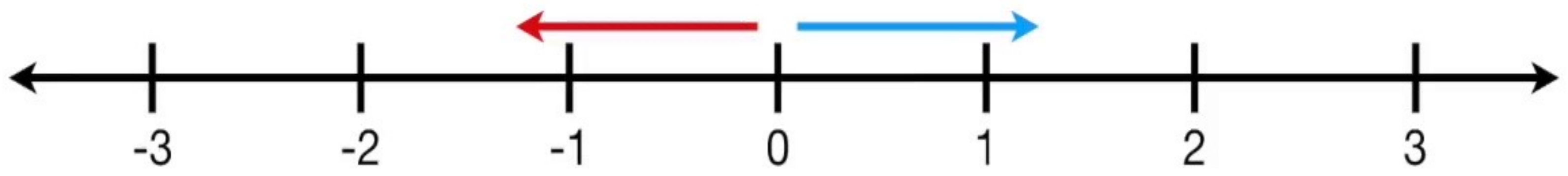
Odds Ratios

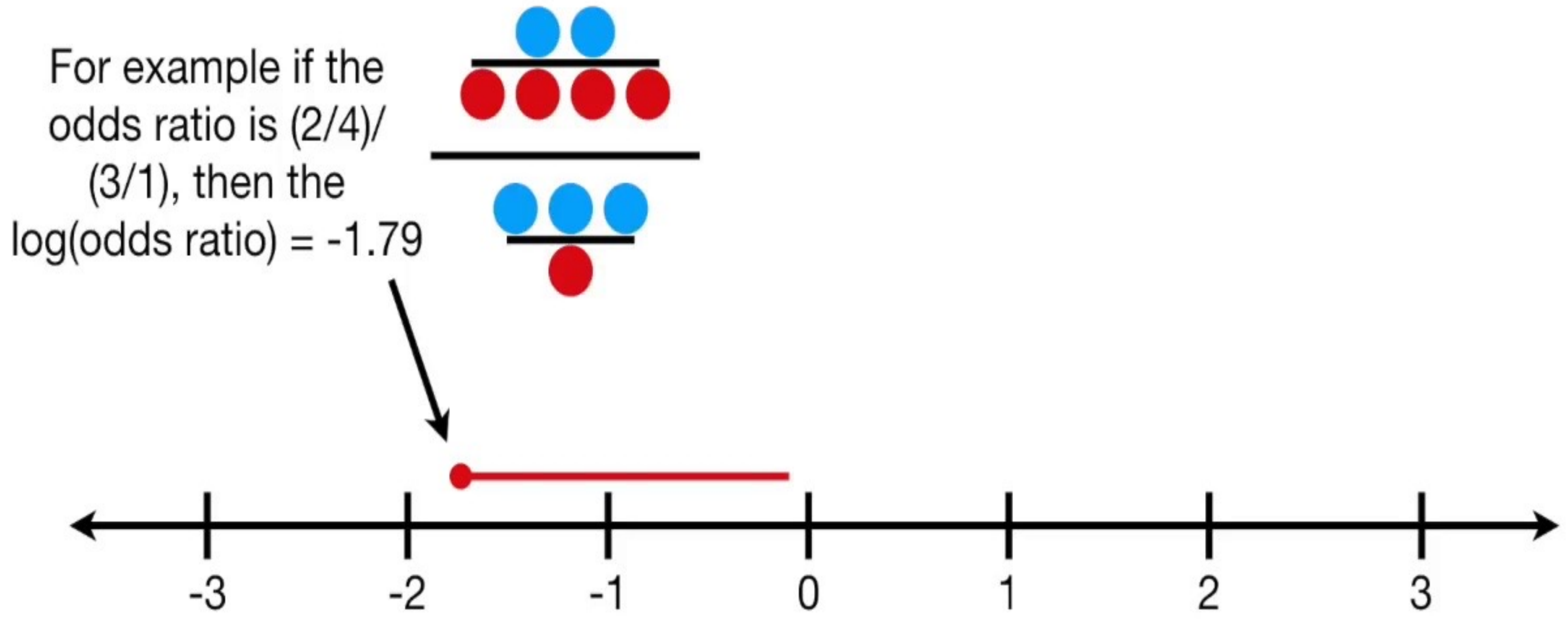


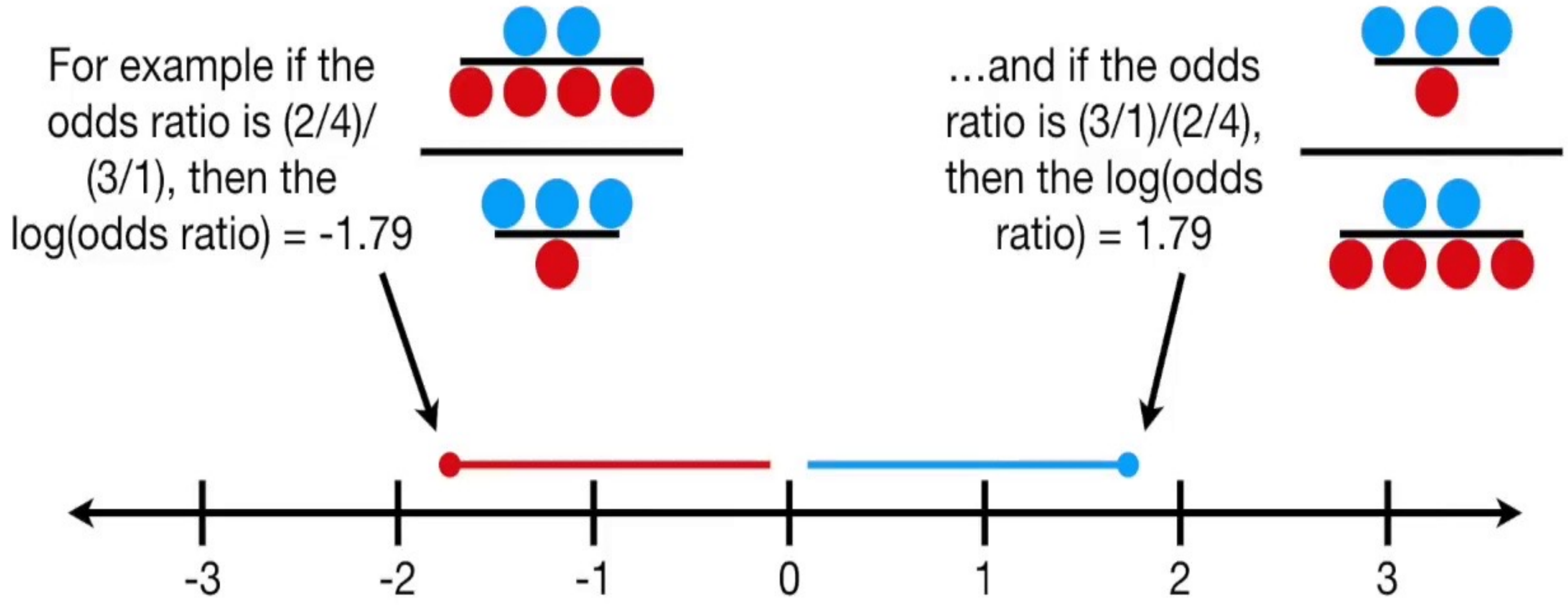
...and if the numerator is larger than the denominator, then the odds ratio will go from 1 to infinity (and beyond!)...
...to infinity and beyond!
0
1
2
3
4
5
6

Log of Odds Ratios

...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

Odds Ratios in Action

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

Odds Ratios in Action

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

We can use an "odds ratio" to determine if there is a relationship between the mutated gene and cancer.

If someone has the mutated gene, are the odds higher that they will get cancer?

Odds Ratios in Action

Given that a person has has the mutated gene, the odds that they have cancer are....

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

$$\rightarrow \frac{23}{117}$$

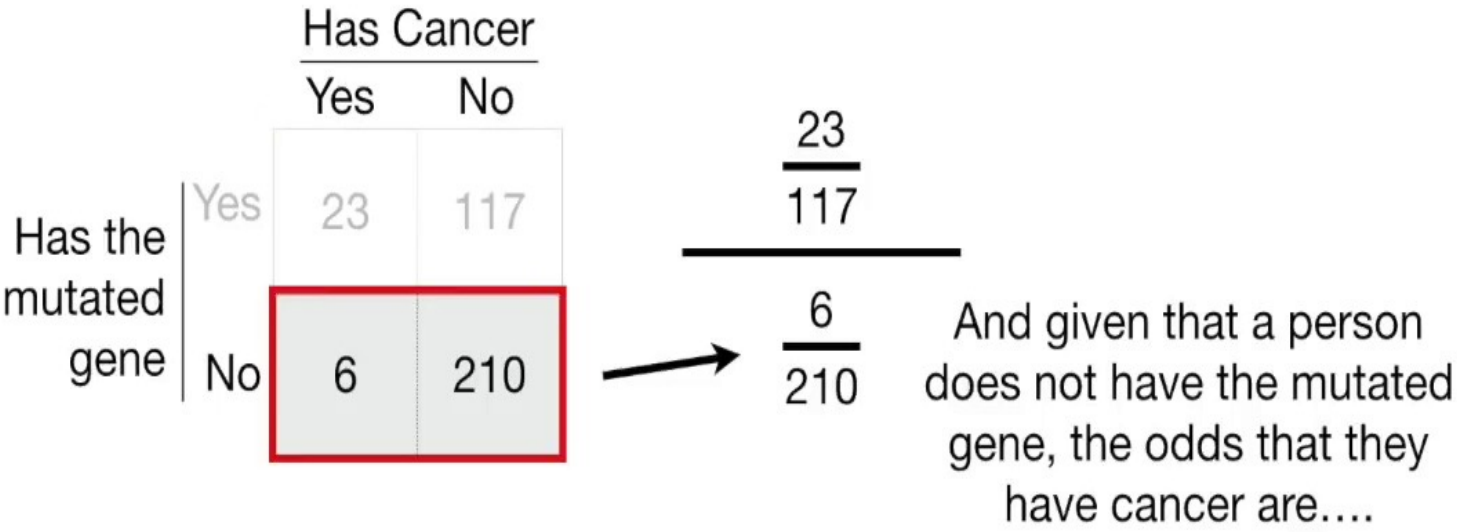
Odds Ratios in Action

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

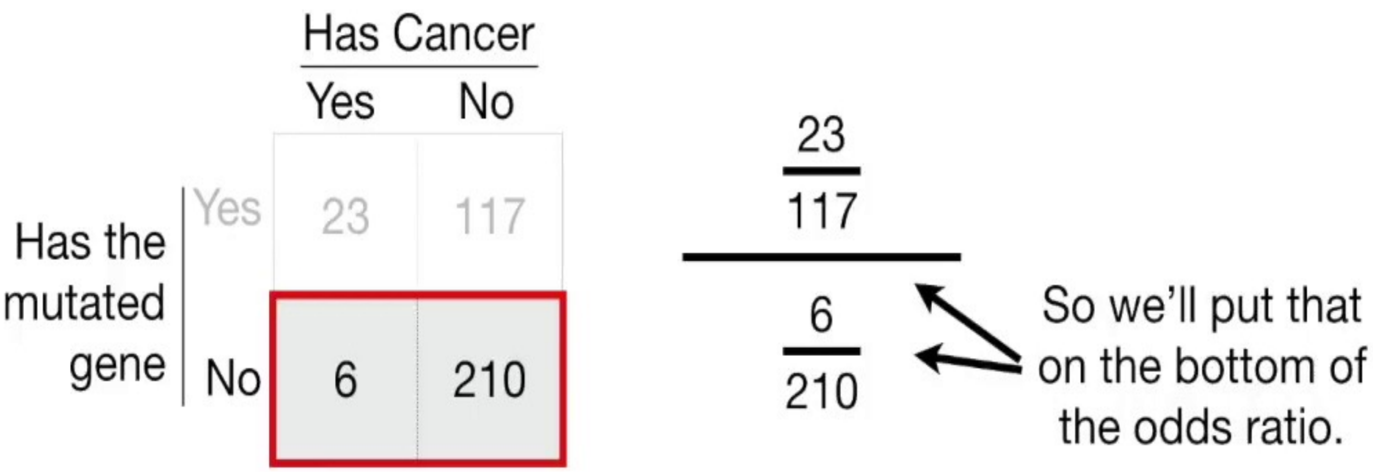
$$\frac{\frac{23}{117}}{\qquad}$$

So we'll put that on top of the odds ratio.

Odds Ratios in Action



Odds Ratios in Action



Odds Ratios in Action

Here's our odds ratio.

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

$$\boxed{\dfrac{\;\dfrac{23}{117}\;}{\;\dfrac{6}{210}\;}}$$

Odds Ratios in Action

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

We do the math…

$$\frac{\frac{23}{117}}{\frac{6}{210}} = \frac{0.2}{0.03} = 6.88$$

Odds Ratios in Action

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

...and the odds ratio tells us that the odds are 6.88 times greater that someone with the mutated gene will also have cancer.

$$\frac{\frac{23}{117}}{\frac{6}{210}} = \frac{0.2}{0.03} \boxed{= 6.88}$$

Odds Ratios in Action

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

$$\frac{\frac{23}{117}}{\frac{6}{210}} = \frac{0.2}{0.03} = 6.88$$

$$\boxed{\log(6.88) = 1.93}$$

...and the log(odds ratio) is 1.93.

Odds Ratios in Action

What does all this mean?

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

$$\frac{\frac{23}{117}}{\frac{6}{210}} = \frac{0.2}{0.03} = 6.88$$

$$\log(6.88) = 1.93$$

Odds Ratios in Action

...larger values mean that the mutated gene is a good predictor of cancer. Smaller values mean that the mutated gene is not a good predictor of cancer.

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

$$\frac{\frac{23}{117}}{\frac{6}{210}} = \frac{0.2}{0.03} \boxed{= 6.88}$$

$$\boxed{\log(6.88) = 1.93}$$

Odds Again

- Given some event with probability *p* of being 1, the odds of that event are given by:

$$\text{odds} = p \,/\, (1-p)$$

- When we go from Normal to High, the odds of being Sick triple:
- Odds ratio: 0.293/0.111 = 2.64
- 2.64 times more likely to be Sick with high values

Sick

Value	Yes	No	Total
Normal	402	3614	4016
High	101	345	446
Total	503	3959	4462

The odds of being sick if you have a Normal value are:

- Odds(Sick|Normal) = P(sick)/1-P(sick) =
  = (402/4016) / (1 - (402/4016))
  = 0.1001 / 0.8889 = 0.111

The odds of being not sick with a Normal value is the reciprocal:

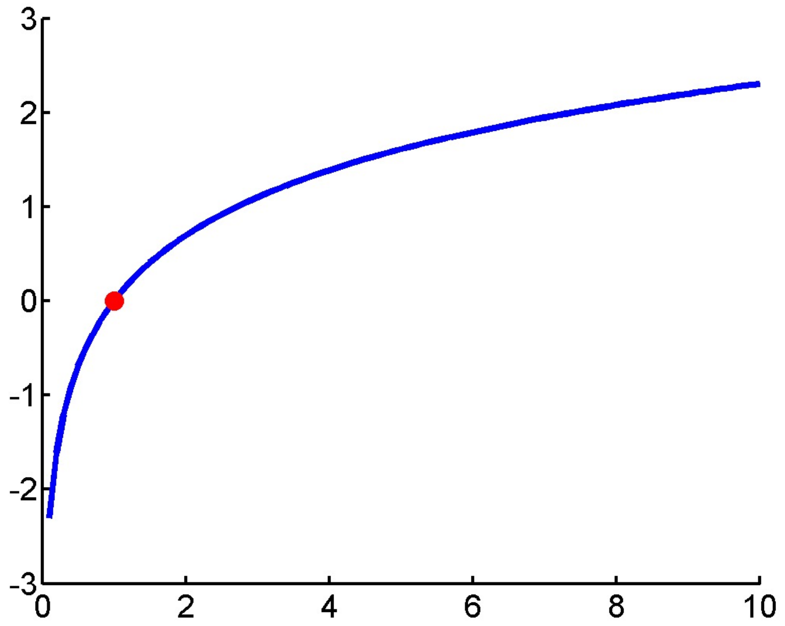
- Odds(not Sick|Normal) = 0.8999/0.1001 = 8.99

For the High value we have

- Odds(Sick|High) = 101/345 = 0.293
- Odds(not Sick|High) = 345/101 = 3.416

Logit Transform

- The logit is the natural log of the odds
- logit(p) = ln(odds) = ln(p/(1-p))



References

- Regression. Appendix D. Introduction to Data Mining.