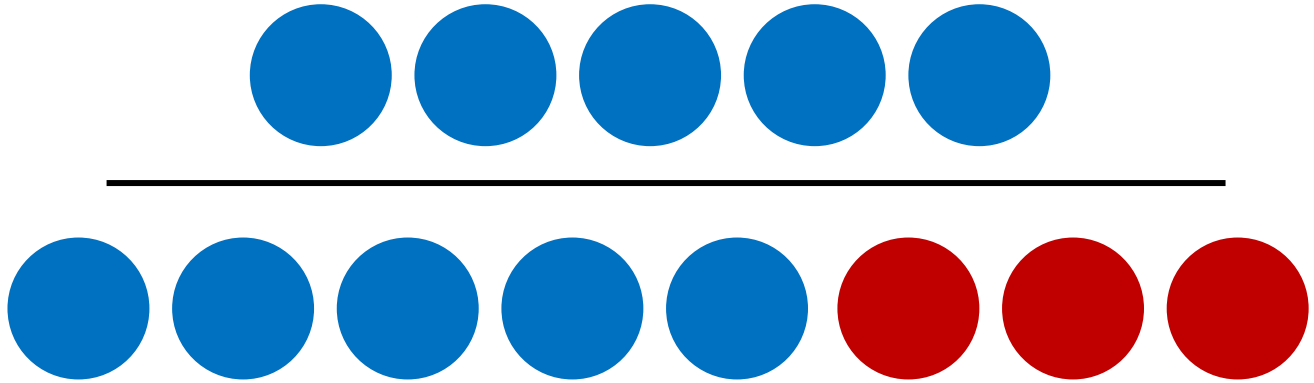
Distinguishing Fundamental Statistical Concepts for Logistic Regression

Yay

Is probability objective?

Probability

- The chance that an event of interest occurs.
- A measurement of how strongly we believe things about the world.
- A probability is calculated with the total number of outcomes of an event counted first.

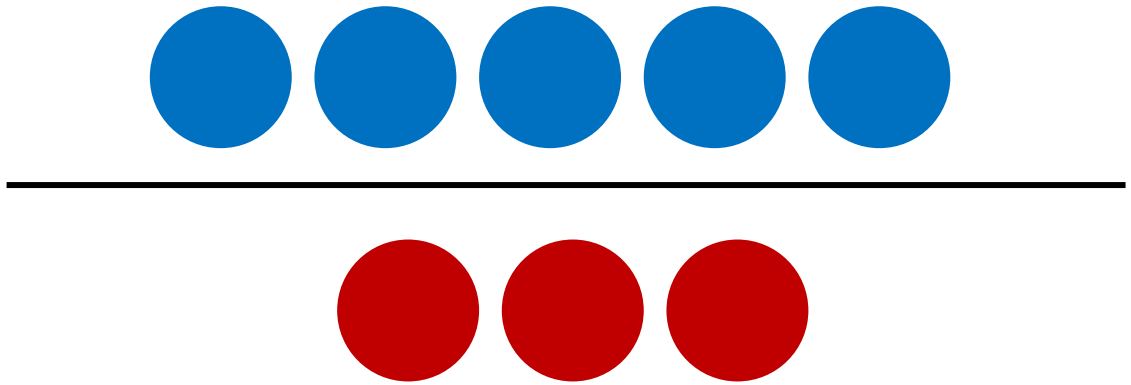


What would be the total number of outcomes for the following events?

“How likely will it rain tomorrow?”
“What chances do you think she will fall in love with him?”
“Do you think we will have time to stop and buy an ice cream?”

Odds

- A ratio of probabilities
- A ratio of something happening to not happening
- Not a probability, but allows to calculate probabilities in general situations.
- $\frac{p}{1-p}$ $(p = probabilty)$, $\frac{counts\ of\ an\ event\ happening}{counts\ of\ the\ event\ not\ happening}$



Odds

- A **bet** is the most obvious and familiar example of odds from which we can estimate a probability of a specific event.
  "You give me 10 dollars if I win, and I give you 5 if I lose."

$$p(win) + p(loose) = 1$$

$$\frac{p(win)}{p(loose)} = \frac{10}{5} = 2$$

*Here, we believe that we are 2 times more likely to win than lose.

$$p(win) = 2 * p(loose) = 2(1 - p(win))$$

$$p(win) = \frac{2}{3} = 67\%$$

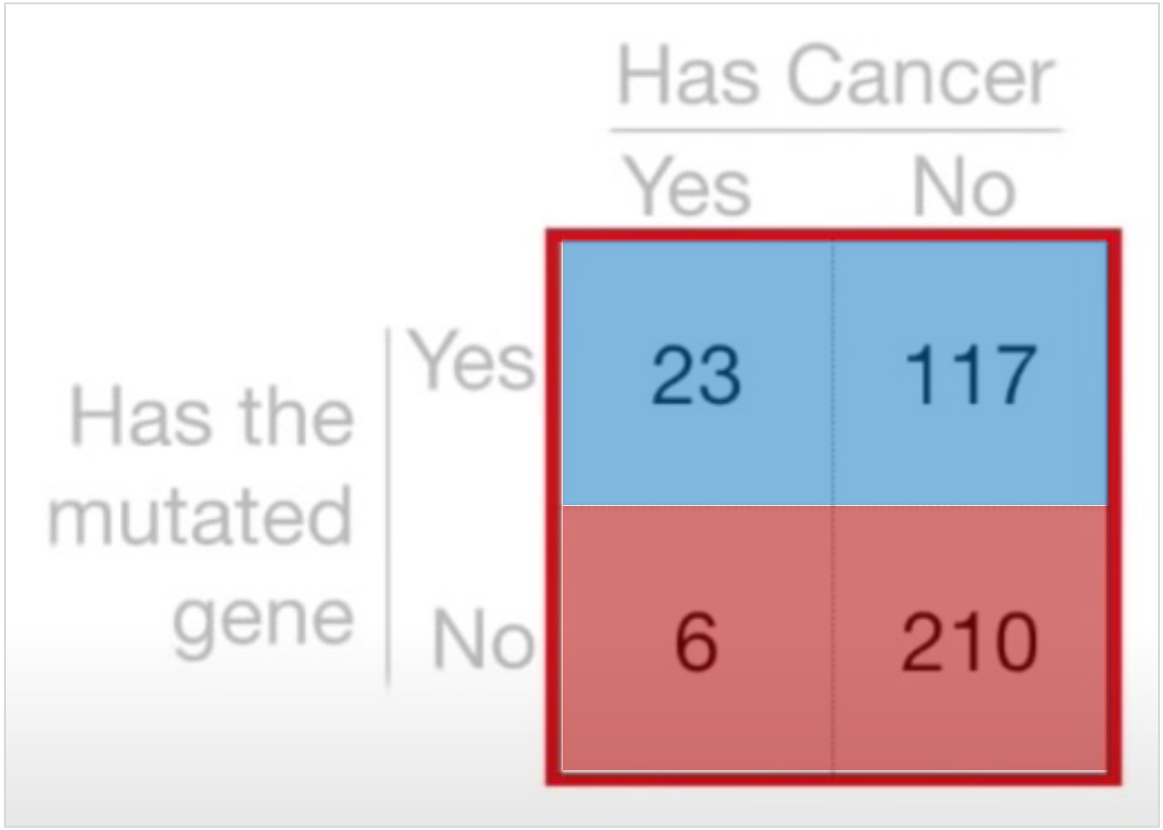
*Without having to know the total number of outcomes for the event of winning, the probability of winning is estimated.

Odds Ratio

- A ratio of odds
- Indicates a relationship between two independent events like R-squared
- Odds Ratio > 1, the considered factor is effective showing positive association.
  0< Odds Ratio < 1, the considered factor is ineffective showing negative association.

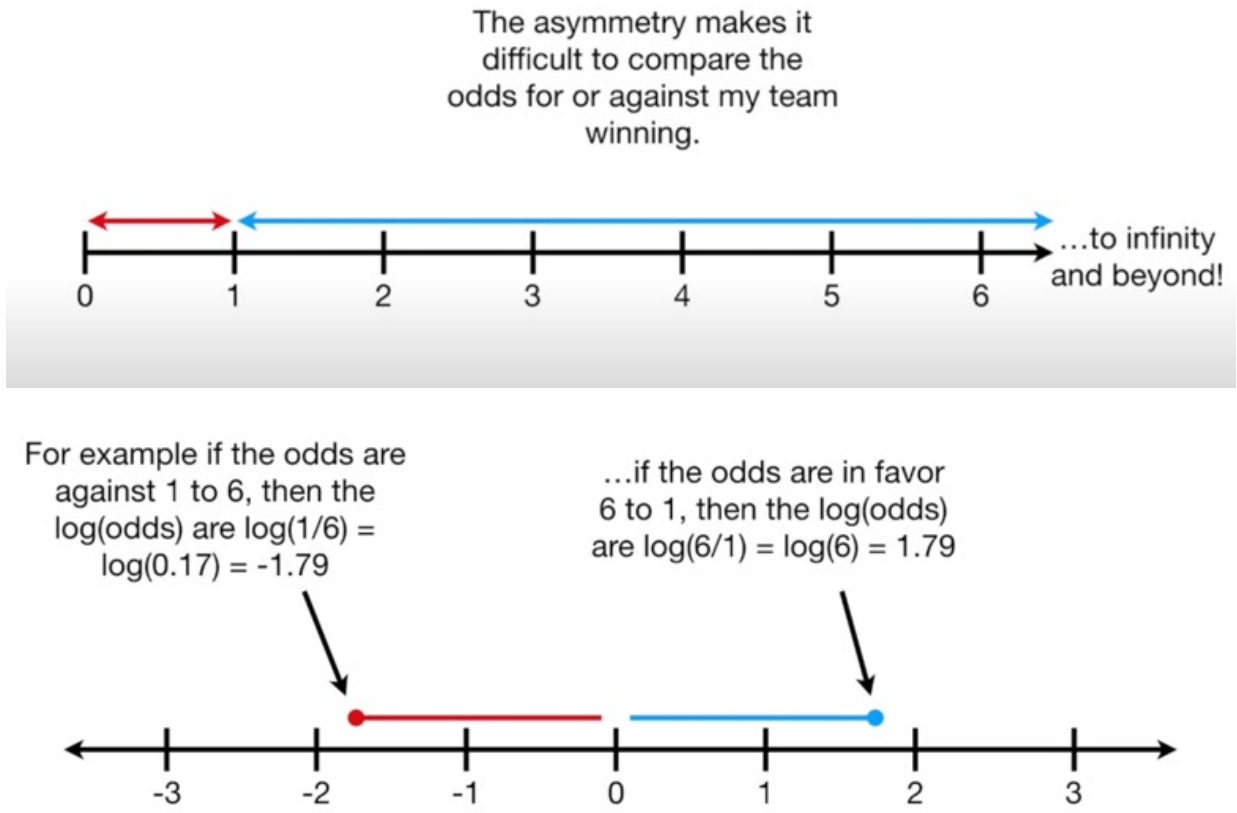
$$\frac{\frac{Having\ Cancer}{Not\ Having\ Cancer}}{\frac{Having\ Cancer}{Not\ Having\ Cancer}} = \frac{\frac{23}{117}}{\frac{6}{210}} = \frac{0.2}{0.03} = 6.88$$

*In this example, the odd ratio review
if “gene mutation” is an effective factor in cancer prediction.



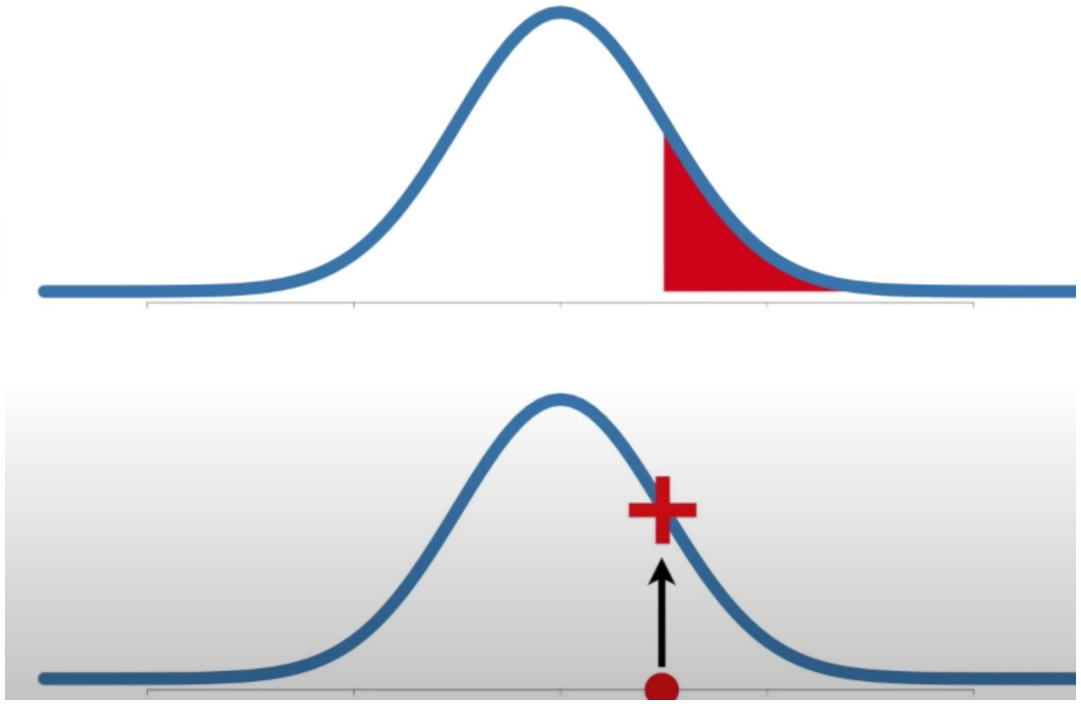
Logit Function

- A log of ratio of probabilities (=Odds)
- Scales the odds values on a log scale, showcasing symmetrical representation.



Likelihood vs. Probability

- Probability; Given a fixed distribution, what would be the area under the curve at a datapoint X?
- Likelihood; Given a set of fixed data points, what would be the distribution that best represent them?



All sum up to understand, Logistic Regression

- Logistic Regression is GLM that generalizes the concepts and abilities of regular linear models, but leveraged for classification problem.

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=i}^{J} \beta_j * x_{ij}$$

*p : logistic regression is transformed from the probability of y-variable(target)

References

- https://thestatsgeek.com/2015/01/03/interpreting-odds-and-odds-ratios/
- https://stats.stackexchange.com/questions/215349/why-use-odds-and-not-probability-in-logistic-regression
- https://towardsdatascience.com/the-concepts-behind-logistic-regression-6316fd7c8031
- https://www.youtube.com/playlist?list=PLblh5JKOoLUKxzEP5HA2d-Li7IJkHfXSe

EOD

Yay