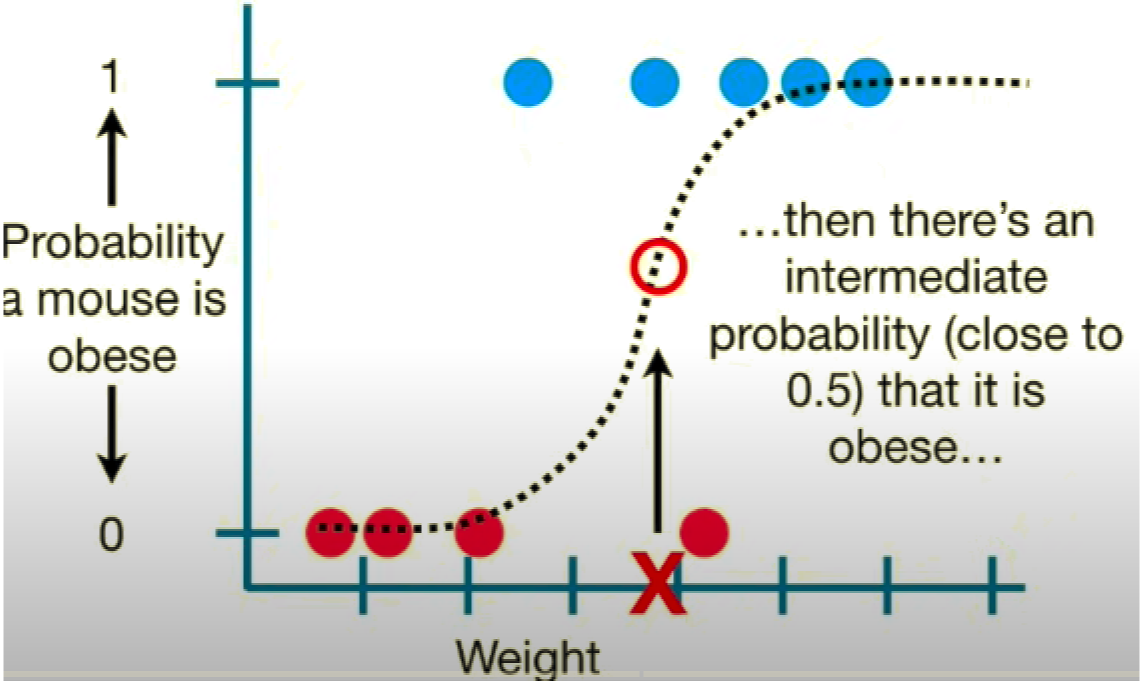
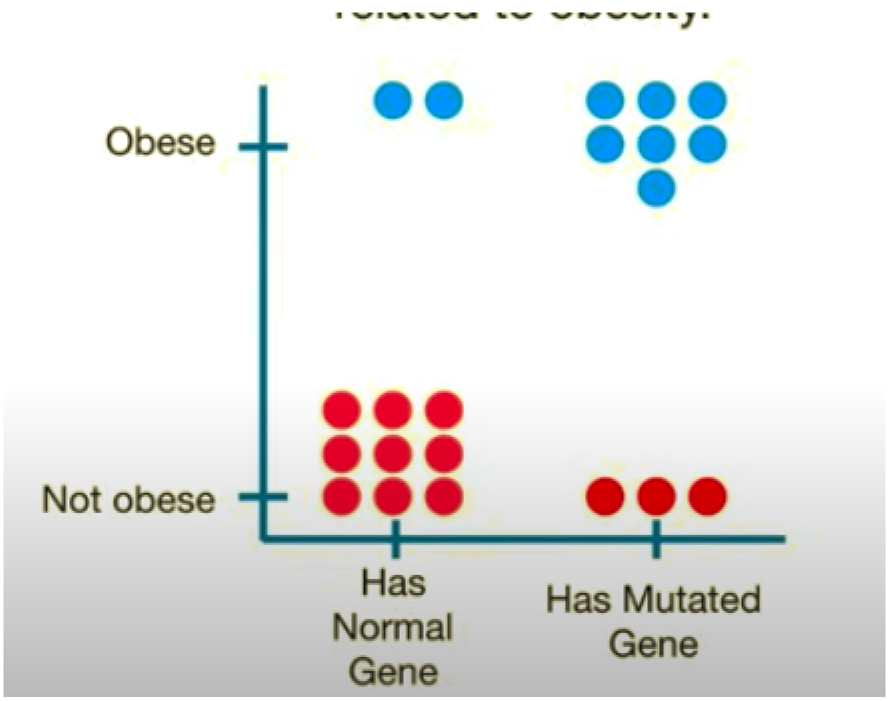
Logistic Regression

COEFFICIENTS

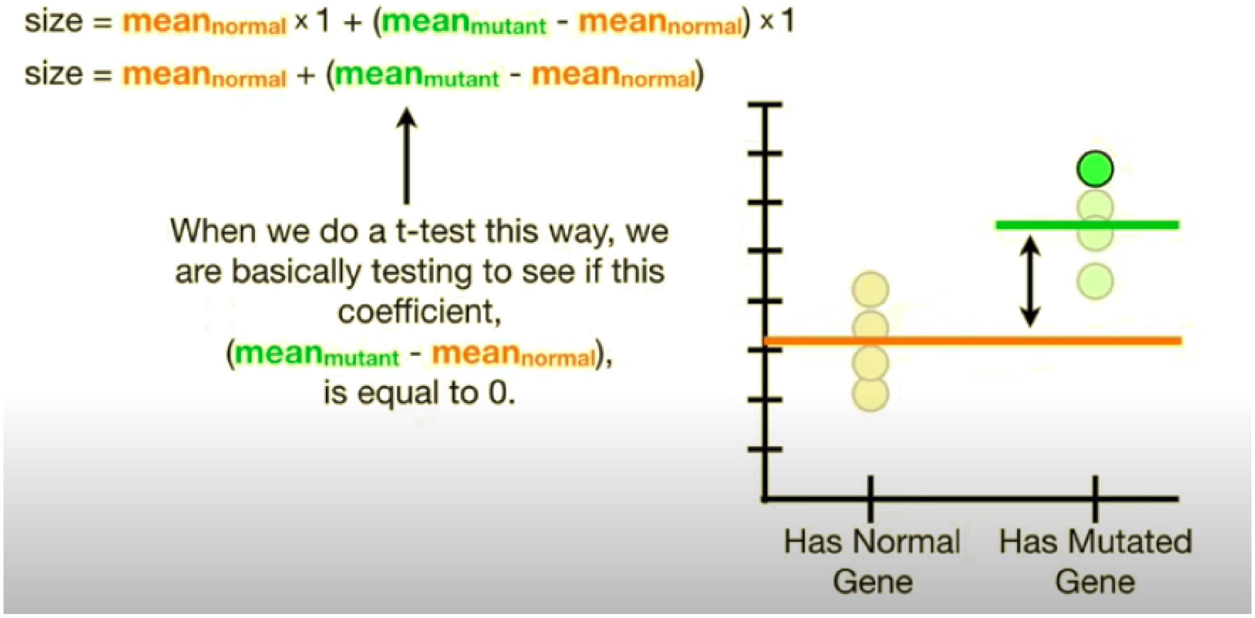


- The X axis determines the weight and corresponds with the probability of a mouse being obsese.
- One of the main differences between Linear and Logistic Regression is the values of the y-axis. In linear regression theoretically the values of the y-axis can be any number. In logistic regression the values on the y-axis must be between 0-1. This is due to the y axis being transformed from the 'probability of obesity' to the 'log odds of obesity'. In this sense it can go from - infinity to + infinity.



- Here we talk about logistic regression coefficients in the context of testing if a discrete variable like “whether or not a mouse has a mutated gene” is related to obesity.
- This type of logistic regression is very similar to how a t-test is done using linear models.

T-test with linear model



$$\text{size} = \text{mean}_{\text{normal}} \times 1 + (\text{mean}_{\text{mutant}} - \text{mean}_{\text{normal}}) \times 1$$

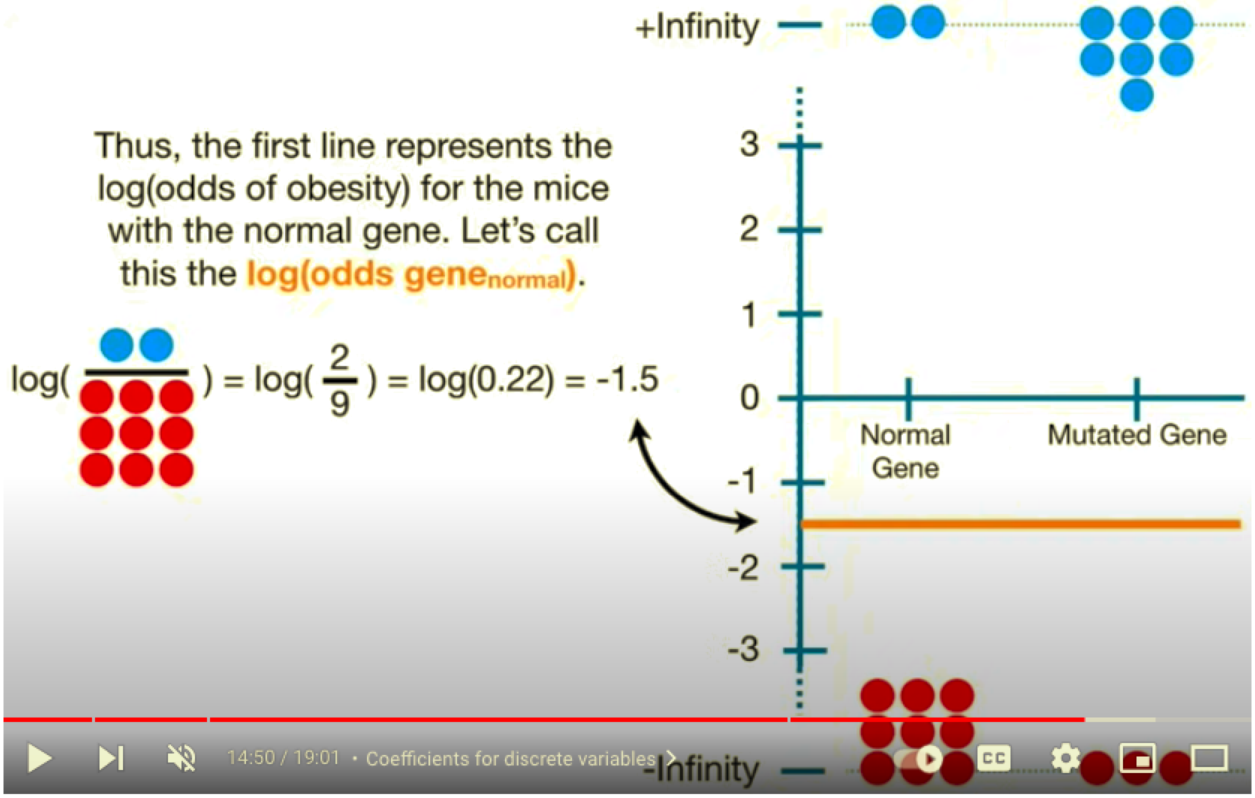
$$\text{size} = \text{mean}_{\text{normal}} + (\text{mean}_{\text{mutant}} - \text{mean}_{\text{normal}})$$

When we do a t-test this way, we are basically testing to see if this coefficient, $(\text{mean}_{\text{mutant}} - \text{mean}_{\text{normal}})$, is equal to 0.

- How does this apply to Logistic Regression

  First thing to do is transform the y-axis from the probability of being obese to the log(odds of obesity) then we fit two lines to the data.

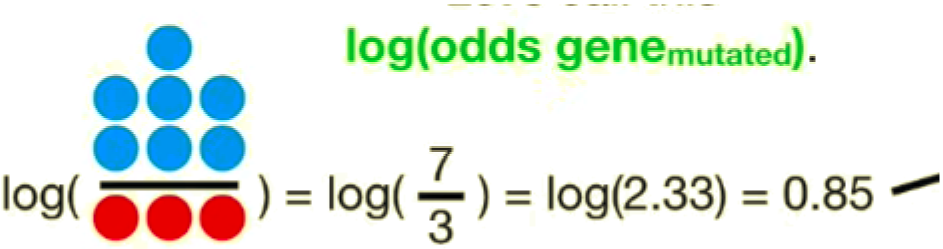
- Given the transformation, we take our first variable of having a normal gene and use it to calculate the log(odds of obesity) for mice with a normal gene. - pictured below



Thus, the first line represents the log(odds of obesity) for the mice with the normal gene. Let's call this the log(odds gene$_{normal}$).

log( ) = log($\frac{2}{9}$) = log(0.22) = -1.5

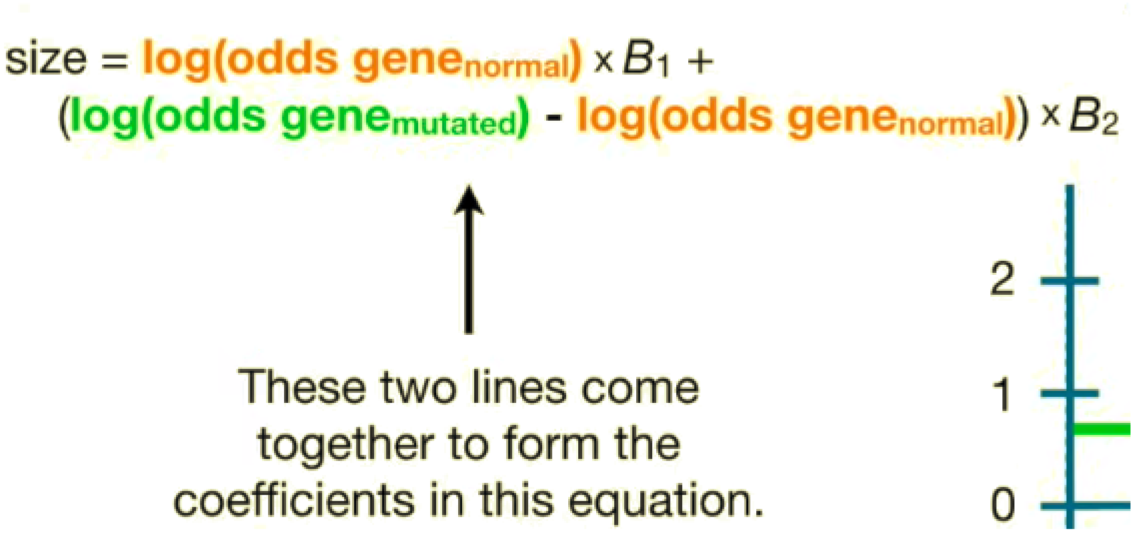
- Then we do the same for mice with the mutated gene. The log(odds of obesity) for a mouse with a mutant gene is 0.85



log(odds gene$_{mutated}$).

log( ) = log($\frac{7}{3}$) = log(2.33) = 0.85

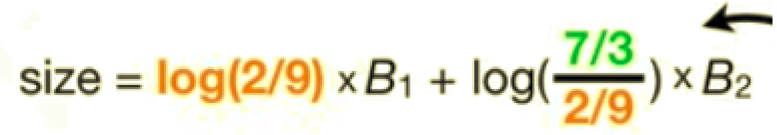
- These two lines come together to form the coefficients in the equation.



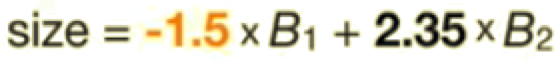
$$\text{size} = \log(\text{odds gene}_{\text{normal}}) \times B_1 + (\log(\text{odds gene}_{\text{mutated}}) - \log(\text{odds gene}_{\text{normal}})) \times B_2$$

These two lines come together to form the coefficients in this equation.

- This provides a log(odds ratio) and tells us on a log scale how much having the mutated gene increases or decrease the odds of a mouse being obese. Below the numbers are substituted in.



$$\text{size} = \log(2/9) \times B_1 + \log\left(\frac{7/3}{2/9}\right) \times B_2$$

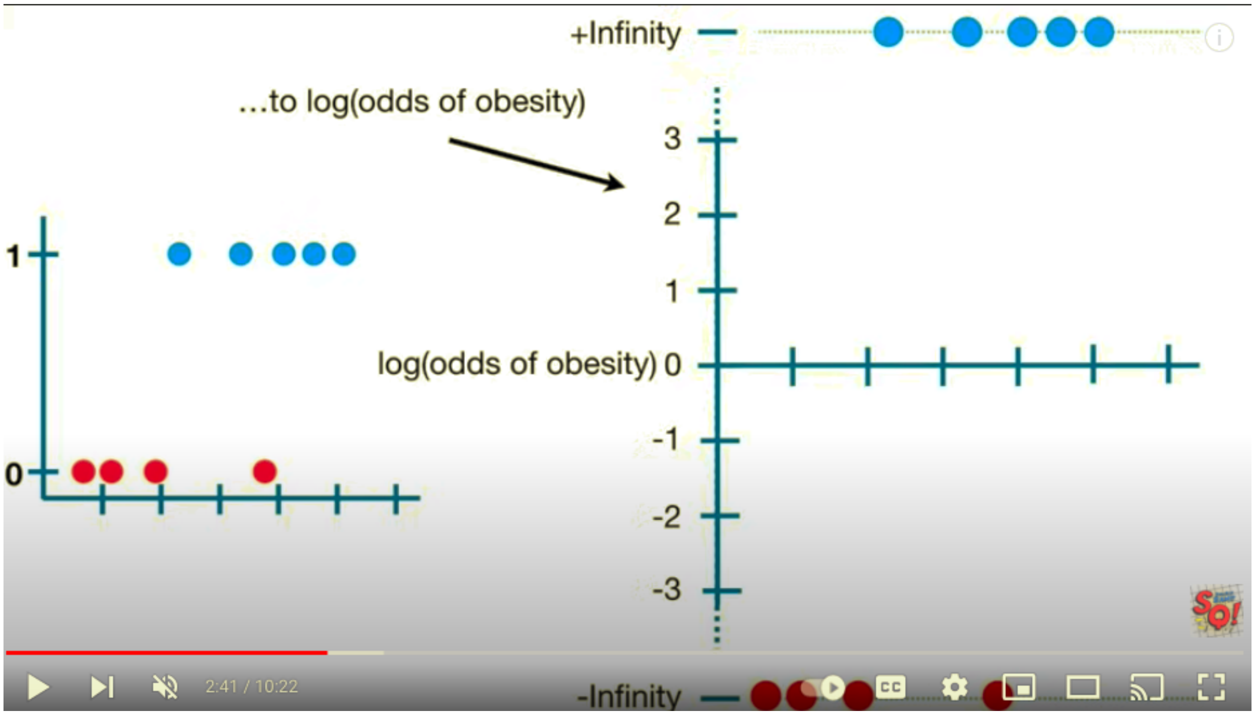


$$\text{size} = -1.5 \times B_1 + 2.35 \times B_2$$

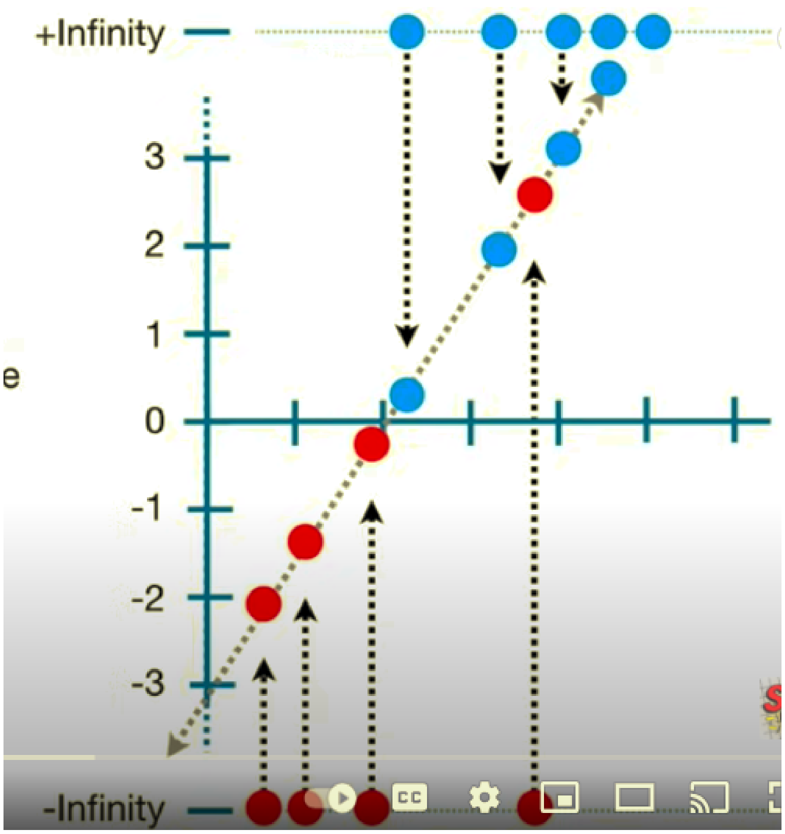
- The -1.5 or the intercept is the log(odds gene normal) and the gene mutant term - 2.35 log(odds ratio) tells us on a log scale, how much having the mutated gene increases or decreases the odds of being obese.
- This explains how some of the linear model concepts apply to logistic regression, and how t-tests apply to logistic regression as well.
- Logistic regression is the exact same as linear regression in the way it identifies coefficients, only the coefficients are in terms of the log(odds).

MAXIMUM LIKELIHOOD

- Fitting the sigmoid curve to where it is optimized to the data.
- In linear regression we fit a line to our data points using least squares regression. In other words we measure the residuals, the distances between the data and the line. Then we square them so that negative values do not cancel out positive values and then add them all up. The line with the smallest sum of squared residuals - the least squares - is the line chosen to fit best.
- As we know in logistic regression we want to do the same with a sigmoid curve. Therefore we transform the y-axis from the probability of obesity to the log odds of obesity.



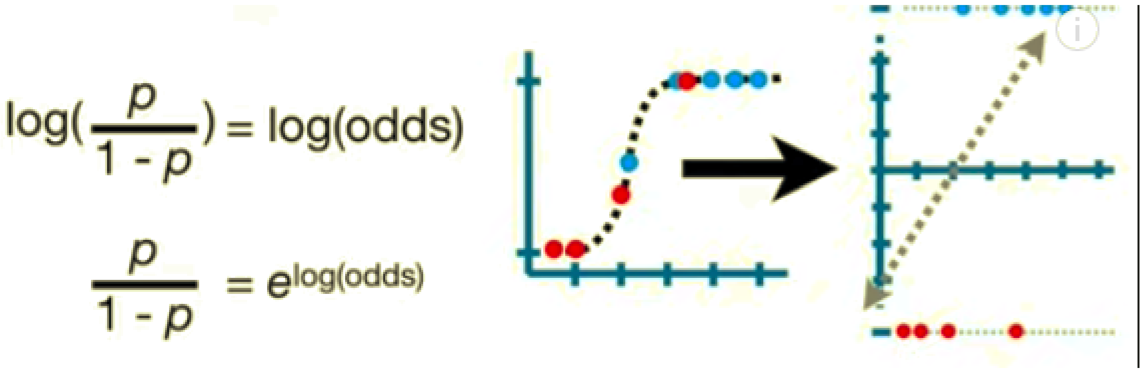
- The problem is that the transformation pushes the raw data to positive and negative infinity. So the next step is to project the data onto a candidate line.



- Then we transform the candidate log odds to candidate probabilities using the formula.

  Just a transformation from probability to log odds.

  Example - The lowest most red point has a log(odds) value of -2.1



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

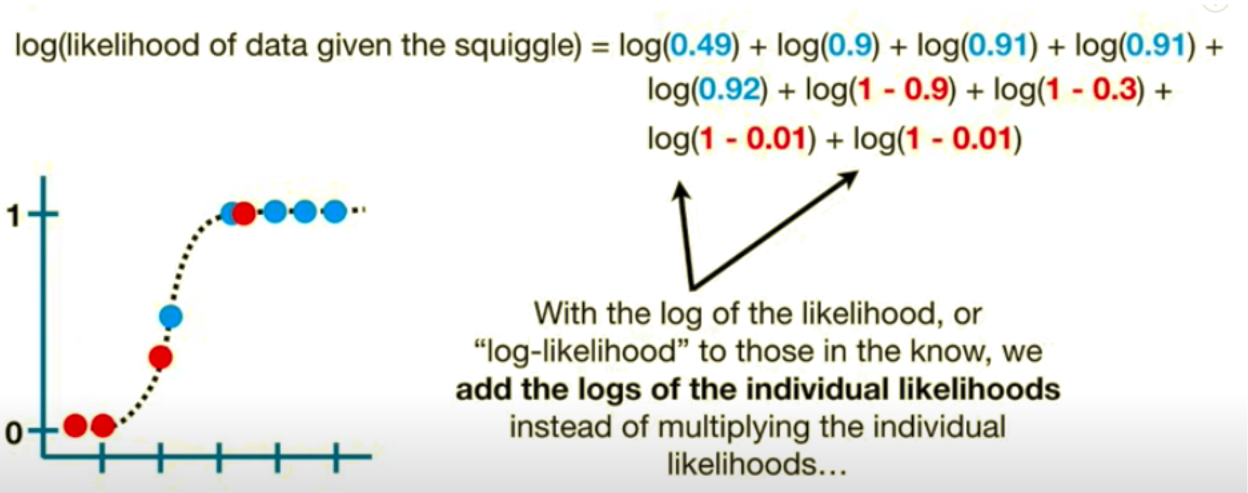
$$\frac{p}{1-p} = e^{\log(\text{odds})}$$

- For example when passing in the lowest red point as -2.1 for log odds this feeds back a probability value of 0.1

- In other words, 10 percent probability of obesity, and 90 percent probability of no obesity.
- In order to find the log likelihood, the probabilities are logged and added.

  In this case the log-likelihood of the sigmoid curve is 3.77

  We want log-likelihood to be as close to zero as possible



log(likelihood of data given the squiggle) = log(0.49) + log(0.9) + log(0.91) + log(0.91) + log(0.92) + log(1 - 0.9) + log(1 - 0.3) + log(1 - 0.01) + log(1 - 0.01)

With the log of the likelihood, or “log-likelihood” to those in the know, we **add the logs of the individual likelihoods** instead of multiplying the individual likelihoods…