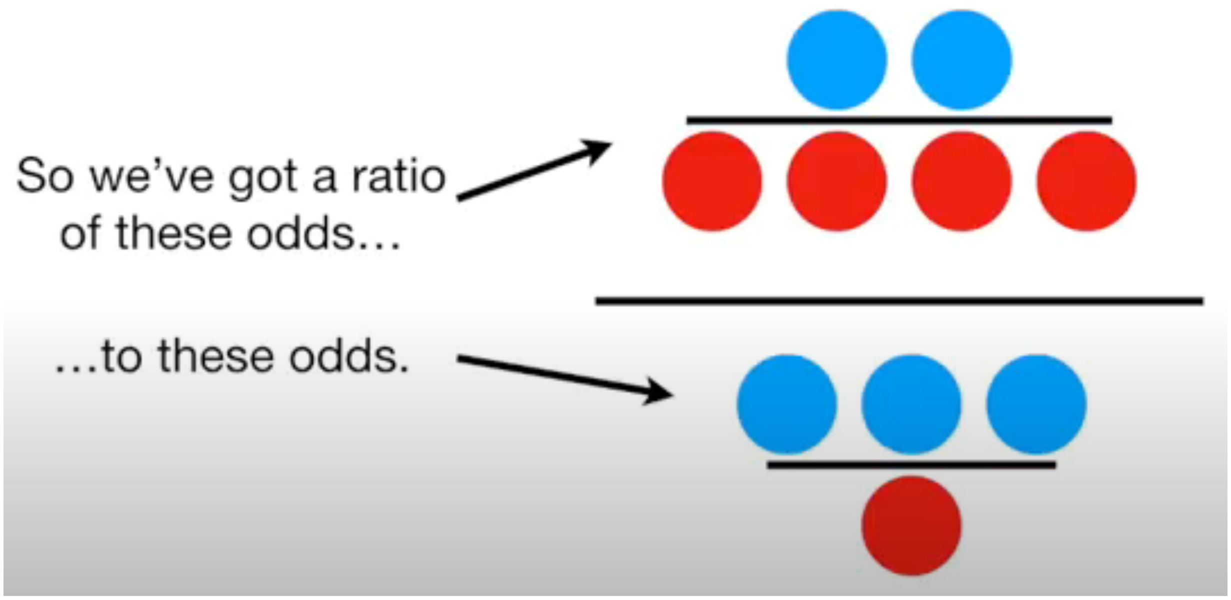
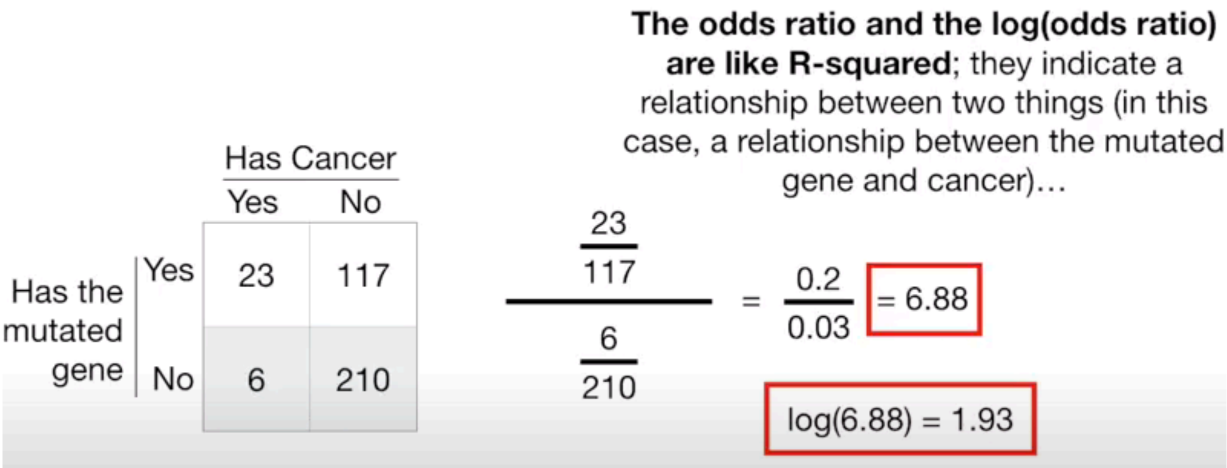
Odds Ratio
by StatQuest

when people say ‘odds ratio’, they are talking about a ‘ratio of odds’



***Remember key point with odds, if denominator is larger than the numerator, then the odds will go from 0 to 1
if the numerator is larger than the denominator, then the odds ratio will go from 1 to infinity
just like log(odds), taking the log(odds ratio) will make things nice and symmetrical

Odds ratio in action:

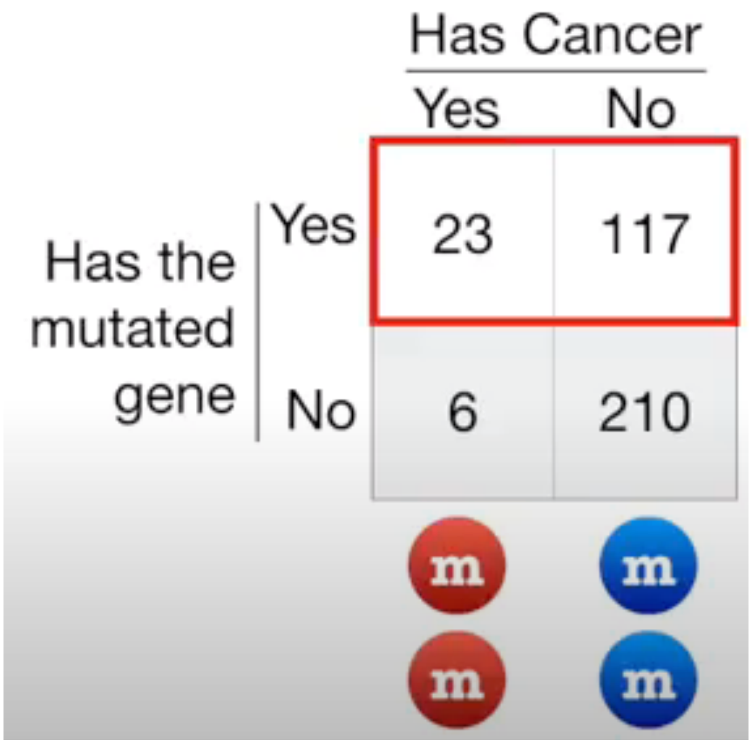


values correspond to effect size
larger values mean that the mutated gene is a good predictor of cancer
smaller values meant that the mutated gene is not a good predictor of cancer
**this requires us to know if the relationship is statistically significant
three ways to do this:

1. Fisher’s Exact Test - calculate a p-value
2. Chi-Square Test - calculate a p-value
3. The Wald Test - calculate a p-value and a confidence interval (CI)

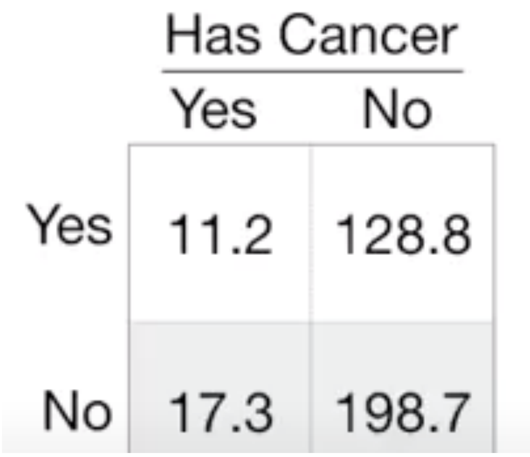
*no general concensus on which is best

Fisher’s Exact Test using the cancer and mutated gene dataset
imagine each person with cancer as a red m&m
and each person without cancer as a blue m&m



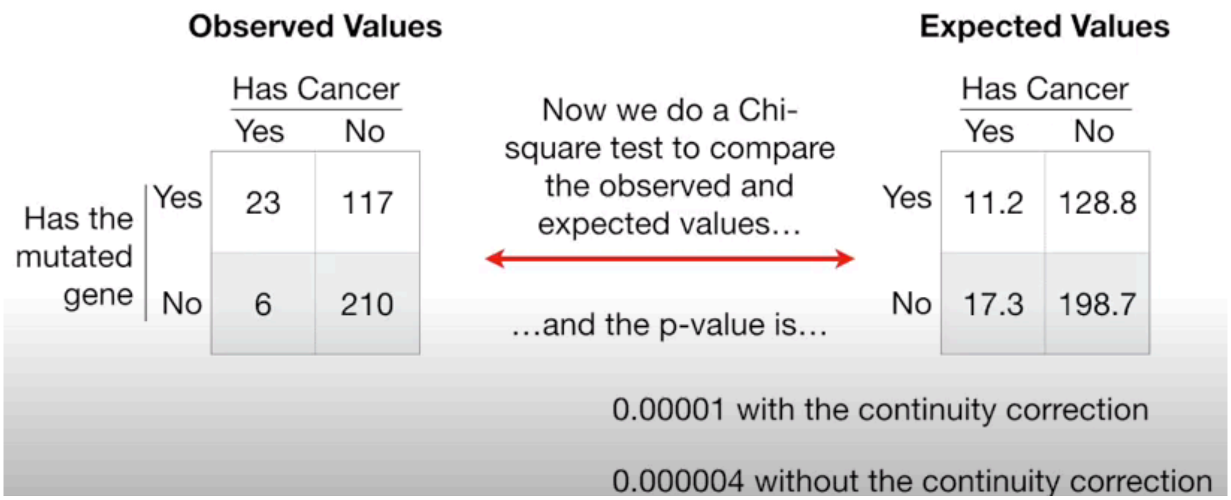
now work out the p-value for grabbing a handful of 23 red m&ms and 117 blue m&ms
*for more details look at StatQuest tutorial

Chi-square test
compares the observed values to expected values that assume there is no relationship between the mutated gene and cancer
calculate teh probability of having cancer as the total number of people with cancer divided by the total number of people
29/356 = p(has cancer) = 0.08
so if the gene is not associated with the 140 people with the mutated gene then the probability of having cancer x the 140 people with the mutated gene = 11.2
**this states that our expected values are:

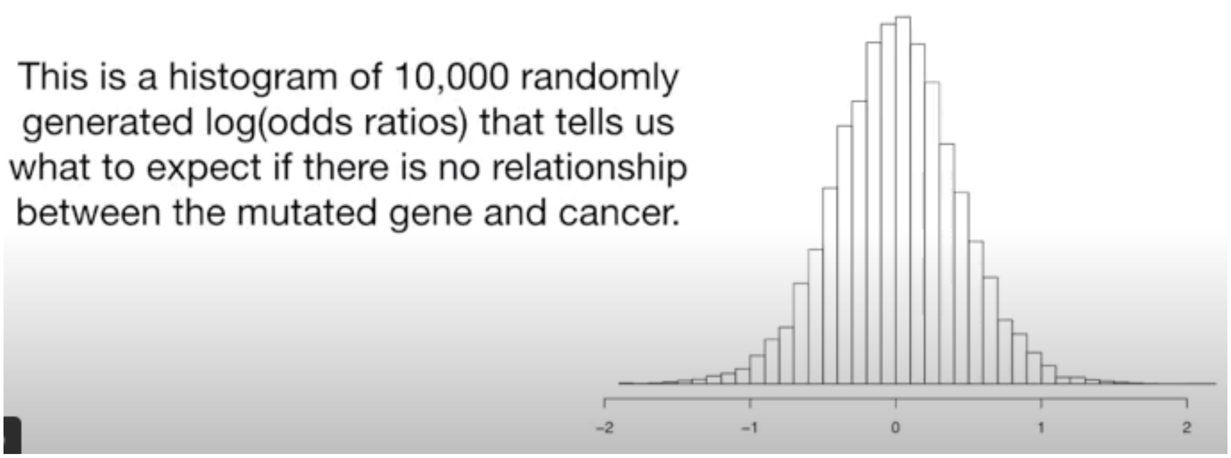


	Has Cancer: Yes	Has Cancer: No
Yes	11.2	128.8
No	17.3	198.7

now we have our observed and expected values



Wald’s Test
commonly used to determine the significance of odds-ratios in logistic regression and to calculate confidence intervals
takes advantage of the fact that log(odds ratios) are normally distributed



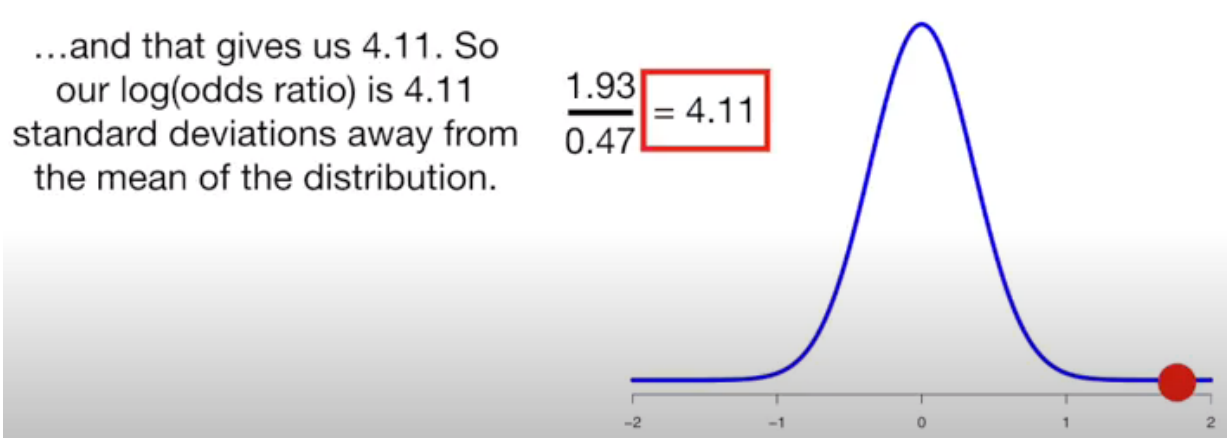
this is based off a matrix of random values that did not depend on a relationship between the mutated gene and cancer
*key point is centered on 0
when there is no difference in the odds, the log(odds ratio)=0
this visual states that there is no relationship between the two variables (in our case between the mutated gene and cancer)
standard deviation of 10,000 log(odds ratio) is 0.43
the more common way to estimate the standard deviation from the observed values is by
taking the square root of 1 over the sum of each observed value
example

$$\sqrt{\frac{1}{23} + \frac{1}{117} + \frac{1}{6} + \frac{1}{210}}$$

this equals 0.47
what the Wald Test does is look to see how many standard deviations the observed log(odds ratio) is from 0
now calculate the log(odds ratio)

$$\text{log(odds ratio)} = \log\left(\frac{\frac{23}{117}}{\frac{6}{210}}\right) = \log(6.88) = 1.93$$

tells us where the log(odds ratio) goes on the curve



we want to know how many standard deviations the log(odds ratio) is away from

the center
in this case it is 4.11
general rule of thumb with normal distributions is that anything further than 2 standard deviations from the mean will have a p-value <0.05
so in our example we know our log(odds ratio) is statistically significant (ie not by chance)

If the above tests worked as expected, 5% should have p-values <0.05.
use the test that is most commonly used in the field that you are evaluating
if p-value is borderline, it may be advantageous to check all three of these tests

The odds ratio (and log(odds ratio)) tells us if there is a strong or weak relationship between two things
like in our above example, whether or not having a mutated gene increases the odds of having cancer