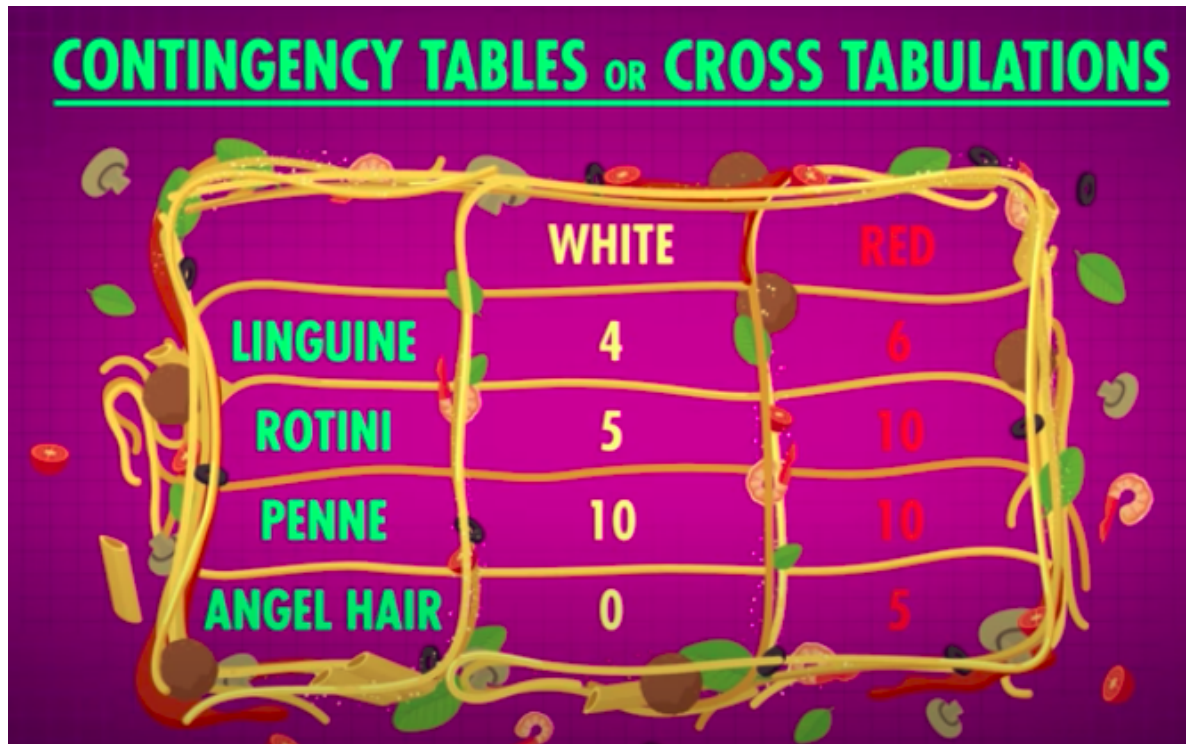


Chi Square Tests: Crash Course Statistics

- Frequency tables look at two or more categorical variables, and we call them contingency tables or cross tabulations.



	WHITE	RED
LINGUINE	4	6
ROTINI	5	10
PENNE	10	10
ANGEL HAIR	0	5

STATISTICAL TEST : $\frac{\text{OBSERVED DIFFERENCE - WHAT WE EXPECT IF THE NULL IS TRUE}}{\text{AVERAGE VARIATION}}$

- GOODNESS OF FIT TEST
- One way to know if you are working with a goodness of fit test is if there is only one row. This lets us know that we are only considering one variable. Like in our case character class.
- One thing we should always check when doing a chi-square test is whether the expected frequency for every cell is greater than 5. If the frequency is less than 5 the result of the test can be off. 5 is arbitrary, but widely accepted.
- For example, a new game called league of lemurs and you can decide on a certain type of character within the game. Categories are healers, fighters, assassins, and tanks.
- NULL HYPOTHESIS IS AS FOLLOWS FOR EACH CATEGORY - on average

$$H_0 : P_{\text{HEALER}} = 15\%, P_{\text{TANK}} = 20\%, P_{\text{ASSASSIN}} = 20\%, P_{\text{FIGHTER}} = 45\%$$

- The alternative hypothesis is that at least one of these average proportions is incorrect, in whether it holds in the top ranks of players.
- We acquire observed data with different info then what we expected based on the percentages.

OBSERVED DATA

HEALER	TANK	ASSASSIN	FIGHTER
25	35	50	90

EXPECTED DATA

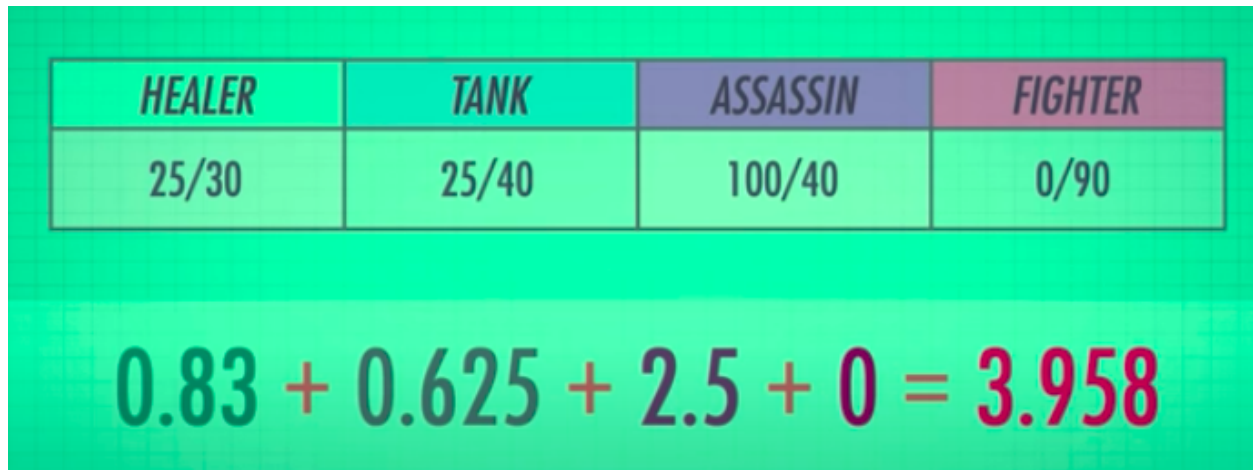
HEALER	TANK	ASSASSIN	FIGHTER
15% OF 200 = 30	20% OF 200 = 40	20% OF 200 = 40	45% OF 200 = 90

- We have to ask whether they are different enough for us to consider whether it is statistically significant.
- Chi squared formula is below. Example would be
 $(25-30)^2/30 + \dots$
 .83

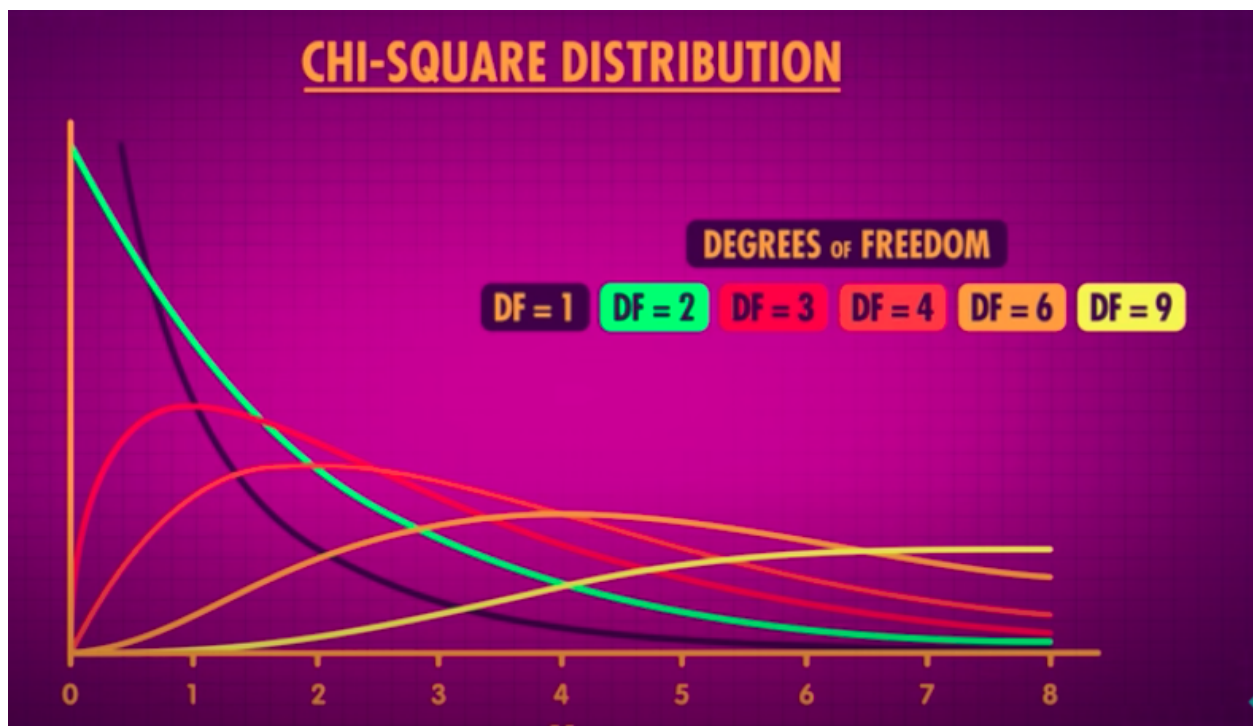
$$\frac{(OBS\ VAL_1 - EXP\ VAL_1)^2}{EXP\ VAL_1} + \frac{(OBS\ VAL_2 - EXP\ VAL_2)^2}{EXP\ VAL_2} + \dots$$

- A deviation of 1 is not a big deal if the expected count is 2000, but if it's 10, that deviation of 1 matters more.

- For example, if you are overcharged a dollar on a 2000 dollar laptop, versus a 10 dollar meal.

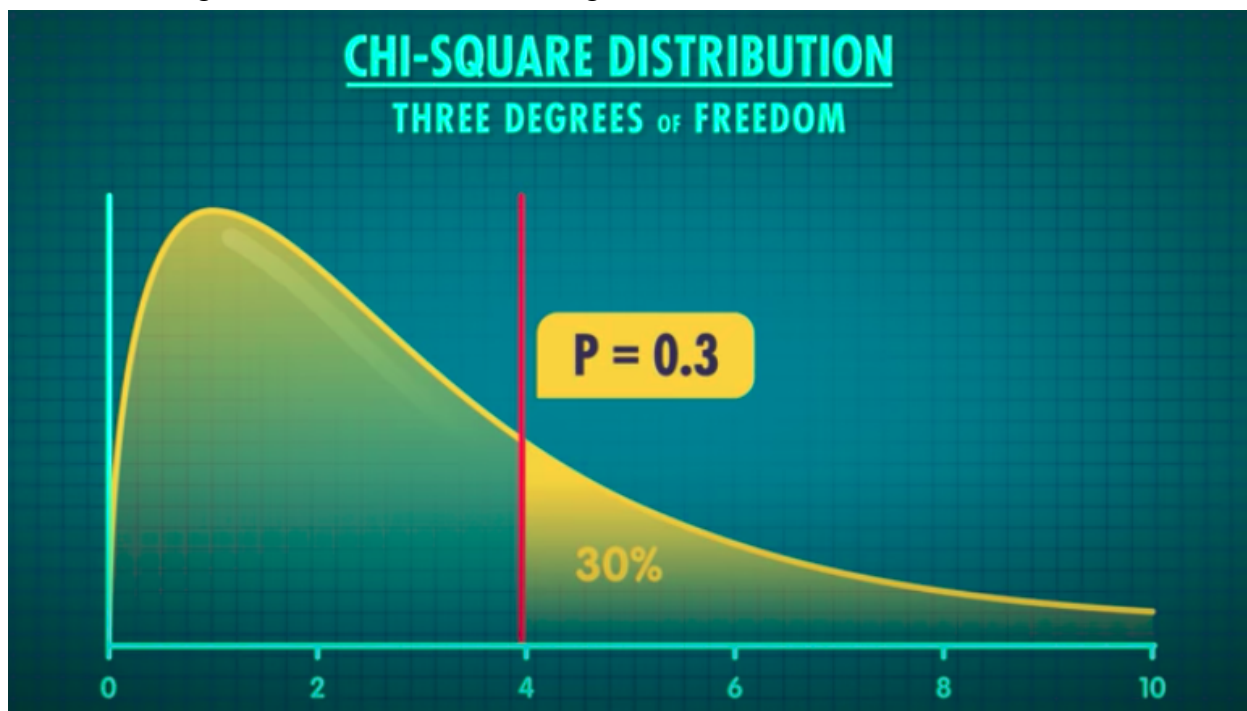


- After passing in numbers to the formula we get a statistic of 3.958
- This helps us quantify how well our observed data fits our expected data.



- Like a T statistic a chi square statistic has a distribution that we can use to find the p-value. Like T distributions Chi squared distributions change their shape with degrees of freedom change.
- To find degrees of freedom we consider what kind of information we have. In the league of lemurs we have 4 columns of healer, tank, assassin, and fighter. That means we have 4 independent pieces of information.

- As soon as we know the total counts - in this case the 200 players recorded. The 4 values are no longer independent because if we know 3 of the values we can get the 4th through basic algebra. Therefore we have 3 degrees of freedom.



- Using our Chi Squared Distribution with 3 degrees of freedom we can now find our P value. Our p-value here is 0.3 , so if our cutoff was .05 we would fail to reject the null. The sample that we took failed to give us any significant evidence that the game developer percentages were wrong.
- Chi squared can handle more than one categorical variable with the test of independence.

TEST OF INDEPENDENCE.

- Look to see whether being a member of one category is independent of the other.
- For example, we look at a survey to see the following.
What Hogwarts house are you in?
Do you like pineapple on pizza?
- What we want to know is whether pineapple on pizza preference is independent from Hogwarts house. Does liking pineapple on pizza affect the probabilities of you identifying with each of the houses?
- We take a random sample of 1000 records.



	<i>GRYFFINDOR</i>	<i>HUFFLEPUFF</i>	<i>RAVENCLAW</i>	<i>SLYTHERIN</i>
<i>NO</i>	79	122	204	74
<i>YES</i>	82	130	240	69



- Unlike our Chi Squared goodness of fit test we are not specifying an exact distribution for Hogwarts houses and comparing our two groups yes/no.
- In this situation we are not concerned about the exact distribution, but we just want to know whether it is different for people who like pineapple on pizza.

TEST OF HOMOGENEITY

- Looking at whether it's likely that different samples come from the same population
- For example, looking at whether two samples of water come from the same lake based on the count of fish, algae, and bacteria that you find in them.
- Next step we need observed frequencies, and expected frequencies which we need to calculate.

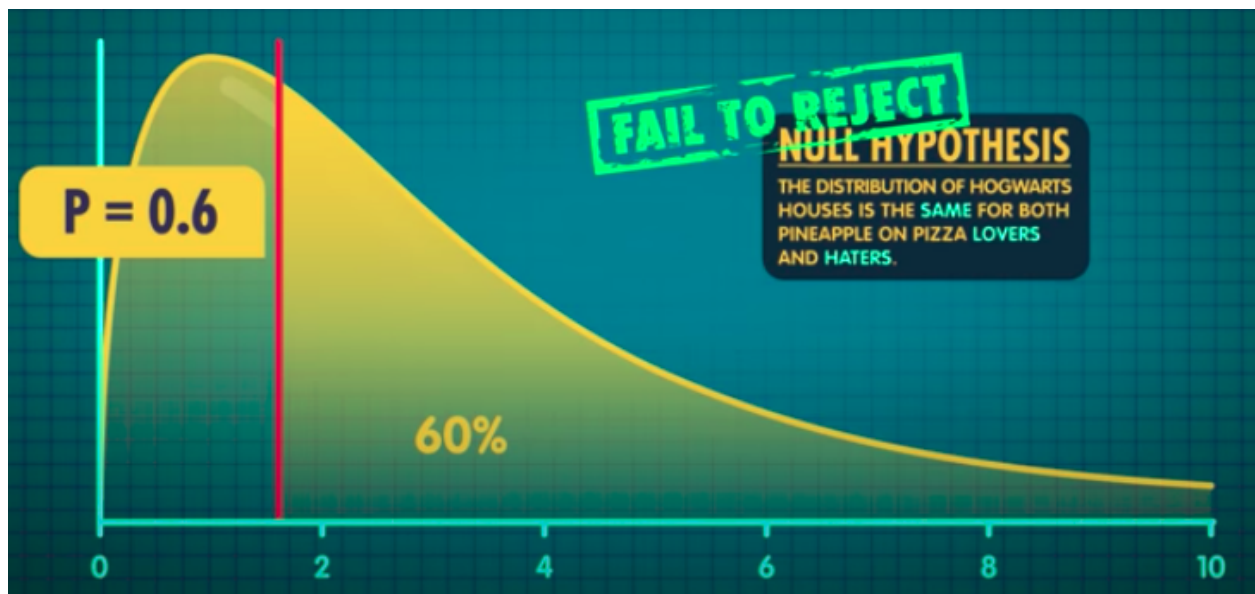
	<i>GRYFFINDOR</i>	<i>HUFFLEPUFF</i>	<i>RAVENCLAW</i>	<i>SLYTHERIN</i>	
<i>NO</i>	79	122	204	74	479
<i>YES</i>	82	130	240	69	521
	161	252	444	143	1000

- In general the formula for degrees of freedom is....
 $(\text{rows} - 1)(\text{columns} - 1)$
 $(2-1)(4-1) = 3$ degrees of freedom.
- The null hypothesis in this case is the amount of people who like pineapple pizza is the same distribution across houses.
- We must calculate our expected frequencies for all cells, and then utilize the chi squared formula.

- The way that we calculate likelihood for each cell is the following:
(total of column * total of row) / sample size
 $(161 * 479) / 1000 = 77.12$

	GRYFFINDOR	HUFFLEPUFF	RAVENCLAW	SLYTHERIN	
NO	77.12	120.71	212.68	68.5	479
YES	83.88	131.29	231.32	74.5	521
	161	252	444	143	1000

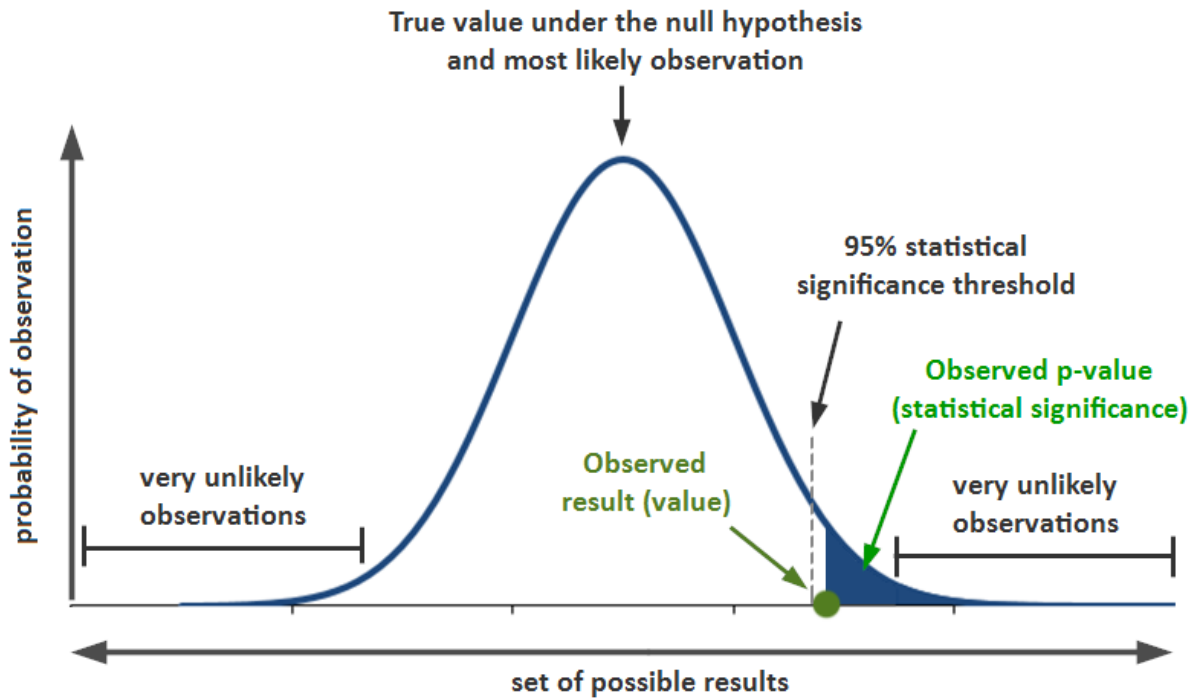
- Our chi square output is 1.6 with three degrees of freedom. We can see that our p-value of .6 is very large compared to our p-value of 0.05.
Therefore we fail to reject the null hypothesis. If the null were true we would expect the see numbers as or more different than ours 60% of the time.



KRISH NAIK - P - VALUE, T TEST, CHI SQUARE, ANOVA TEST WHEN TO USE

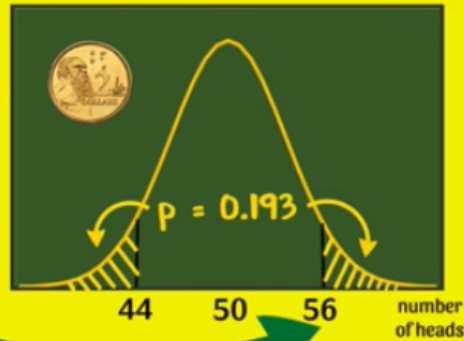
- P-value, significance value, and alpha are all the same thing.

- We must always establish the null that there is no difference, where the alternate accepts the difference.
- If a test provides a value less than .05 then we reject the null hypothesis.
& vice versa if the test provides a value that is greater than .05 then we accept the null hypothesis.



- The p-value is .05 or 5%
- One categorical variable
- Chi squared test is done with two categorical features.
- With one continuous variable such as weight we utilize a T-test
- With two continuous variables we (height, weight) use a correlation test. Then we apply a T-test.
- If we have categorical variables, paired with continuous, and the categorical variable has outcomes of more than two features then we use an Anova test.
- The p-value is effectively saying if the null hypothesis is true then there is a 0.05 or 5% chance of observing a difference as large or larger than what we observed in our sample.

100 coin flips
expecting: 50 heads
sample: 56 heads



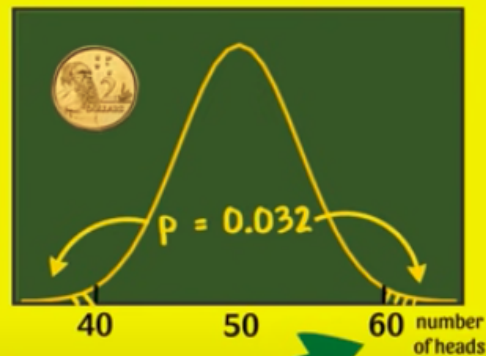
Under the null hypothesis, the **p-value** is the probability of getting a sample AS OR MORE extreme than our own.

$$p = 0.193$$

If the coin was fair, the probability of getting 56 heads (or a sample more extreme) is 19.3%

H_0 : the coin is fair

100 coin flips
expecting: 50 heads
sample: 60 heads



$$p = 0.032$$

If the coin was fair, the probability of getting 60 heads (or a sample more extreme) is 3.2%