Chi Square Statistics

Tushar B. Kute,
http://tusharkute.com

Goodness of fit

- The goodness of fit test is used to test if sample data fits a distribution from a certain population (i.e. a population with a normal distribution or one with a Weibull distribution).
- In other words, it tells you if your sample data represents the data you would expect to find in the actual population. Goodness of fit tests commonly used in statistics are:
  - The chi-square.
  - Kolmogorov-Smirnov.
  - Anderson-Darling.
  - Shipiro-Wilk.

Chi Square Test

- There are two types of chi-square tests. Both use the chi-square statistic and distribution for different purposes:
  - A chi-square goodness of fit test determines if sample data matches a population.
  - A chi-square test for independence compares two variables in a contingency table to see if they are related.
  - In a more general sense, it tests to see whether distributions of categorical variables differ from each another.

Chi Square Test

- The formula for the chi-square statistic used in the chi square test is:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- The subscript “c” is the degrees of freedom. “O” is your observed value and E is your expected value. It’s very rare that you’ll want to actually use this formula to find a critical chi-square value by hand.
- The summation symbol means that you’ll have to perform a calculation for every single data item in your data set.

Chi Square Test

- A chi-square statistic is one way to show a relationship between two categorical variables. In statistics, there are two types of variables:
  - numerical (countable) variables and non-numerical (categorical) variables.
- The chi-squared statistic is a single number that tells you how much difference exists between your observed counts and the counts you would expect if there were no relationship at all in the population.

Chi Square Test

- There are a few variations on the chi-square statistic. Which one you use depends upon how you collected the data and which hypothesis is being tested.
- However, all of the variations use the same idea, which is that you are comparing your expected values with the values you actually collect.
- One of the most common forms can be used for contingency tables:

Contingency Table

- Contingency tables (also called crosstabs or two-way tables) are used in statistics to summarize the relationship between several categorical variables.
- A contingency table is a special type of frequency distribution table, where two variables are shown simultaneously.

Contingency Table

- A Contingency table (also called crosstab) is used in statistics to summarise the relationship between several categorical variables.
- Here, we take a table that shows the number of men and women buying different types of pets.

	dog	cat	bird	total
men	207	282	241	730
women	234	242	232	708
total	441	524	473	1438

- The aim of the test is to conclude whether the two variables(gender and choice of pet) are related to each other.

Contingency Table: Chi Square

- Where O is the observed value, E is the expected value and “i” is the “i^{th}” position in the contingency table.

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Contingency Table: Chi Square

- A low value for chi-square means there is a high correlation between your two sets of data.
- In theory, if your observed and expected values were equal (“no difference”) then chi-square would be zero — an event that is unlikely to happen in real life.
- Deciding whether a chi-square test statistic is large enough to indicate a statistically significant difference isn’t as easy it seems.
- It would be nice if we could say a chi-square test statistic >10 means a difference, but unfortunately that isn’t the case.

Contingency Table: Chi Square

- You could take your calculated chi-square value and compare it to a critical value from a chi-square table.
- If the chi-square value is more than the critical value, then there is a significant difference.
- You could also use a p-value. First state the null hypothesis and the alternate hypothesis. Then generate a chi-square curve for your results along with a p-value
- Small p-values (under 5%) usually indicate that a difference is significant (or “small enough”).

Contingency Table: Chi Square

- Tip: The Chi-square statistic can only be used on numbers.
- They can’t be used for percentages, proportions, means or similar statistical values.
- For example, if you have 10 percent of 200 people, you would need to convert that to a number (20) before you can run a test statistic.

Chi Square: p-values

- A chi square test will give you a p-value. The p-value will tell you if your test results are significant or not. In order to perform a chi square test and get the p-value, you need two pieces of information:
  - Degrees of freedom. That's just the number of categories minus 1.
  - The alpha level(α). This is chosen by you, or the researcher. The usual alpha level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

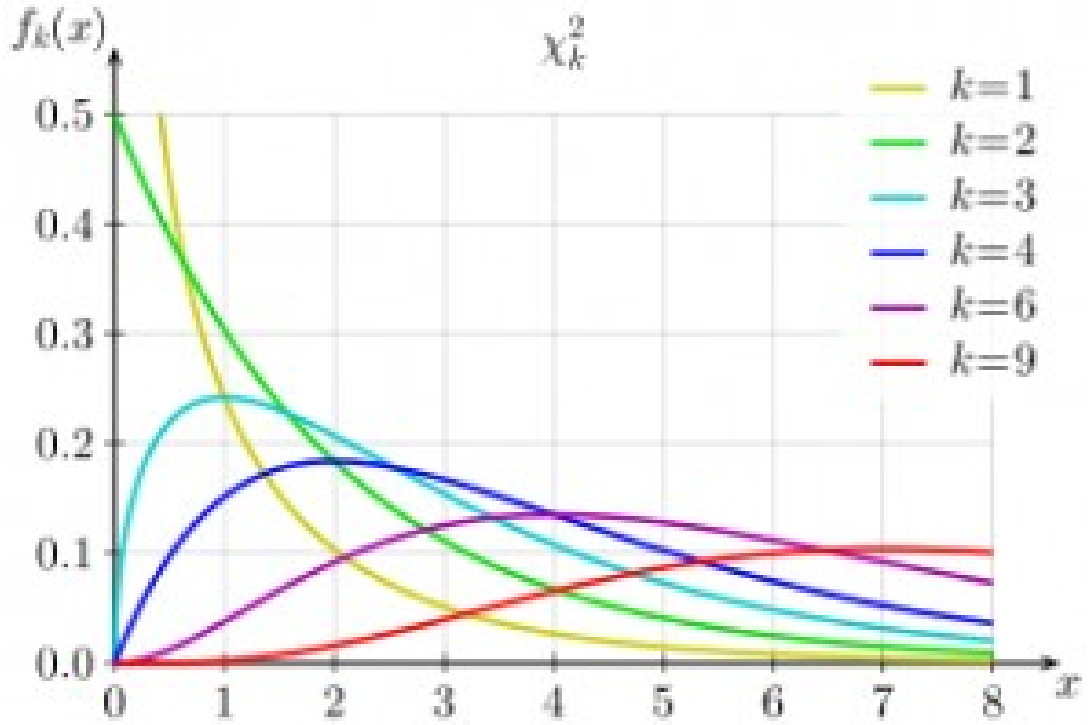
Chi Square: p-values

- In elementary statistics or AP statistics, both the degrees of freedom(df) and the alpha level are usually given to you in a question.
- You don't normally have to figure out what they are. You may have to figure out the df yourself, but it's pretty simple: count the categories and subtract 1.
- Degrees of freedom are placed as a subscript after the chi-square (X^2) symbol. For example, the following chi square shows 6 df:

  X^2_6.
- And this chi square shows 4 df:

  X^2_4.

Chi Square Distribution



Calculate chi square statistic

- Example question: 256 visual artists were surveyed to find out their zodiac sign.
- The results were: Aries (29), Taurus (24), Gemini (22), Cancer (19), Leo (21), Virgo (18), Libra (19), Scorpio (20), Sagittarius (23), Capricorn (18), Aquarius (20), Pisces (23).
- Test the hypothesis that zodiac signs are evenly distributed across visual artists.

Calculate chi square statistic

- Step 1: Make a table with columns for “Categories,” “Observed,” “Expected,” “Residual (Obs-Exp)”, “(Obs-Exp)2” and “Component (Obs-Exp)^2 / Exp.”

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp

Calculate chi square statistic

- Step 2: Fill in your categories. Categories should be given to you in the question. There are 12 zodiac signs, so:

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries					
Taurus					
Gemini					
Cancer					
Leo					
Virgo					
Libra					
Scorpio					
Sagittarius					
Capricorn					
Aquarius					
Pisces					

Calculate chi square statistic

- Step 3: Write your counts. Counts are the number of each items in each category in column 2. You're given the counts in the question:

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29				
Taurus	24				
Gemini	22				
Cancer	19				
Leo	21				
Virgo	18				
Libra	19				
Scorpio	20				
Sagittarius	23				
Capricorn	18				
Aquarius	20				
Pisces	23				

Calculate chi square statistic

- Step 4: Calculate your expected value for column 3. In this question, we would expect the 12 zodiac signs to be evenly distributed for all 256 people, so 256/12=21.333. Write this in column 3.

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333			
Taurus	24	21.333			
Gemini	22	21.333			
Cancer	19	21.333			
Leo	21	21.333			
Virgo	18	21.333			
Libra	19	21.333			
Scorpio	20	21.333			
Sagittarius	23	21.333			
Capricorn	18	21.333			
Aquarius	20	21.333			
Pisces	23	21.333			

Calculate chi square statistic

- Step 5: Subtract the expected value (Step 4) from the Observed value (Step 3) and place the result in the “Residual” column. For example, the first row is Aries: 29-21.333=7.667.

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667		
Taurus	24	21.333	2.667		
Gemini	22	21.333	0.667		
Cancer	19	21.333	-2.333		
Leo	21	21.333	-0.333		
Virgo	18	21.333	-3.333		
Libra	19	21.333	-2.333		
Scorpio	20	21.333	-1.333		
Sagittarius	23	21.333	1.667		
Capricorn	18	21.333	-3.333		
Aquarius	20	21.333	-1.333		
Pisces	23	21.333	1.667		

Calculate chi square statistic

- Step 6: Square your results from Step 5 and place the amounts in the $(Obs-Exp)^2$ column.

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	
Taurus	24	21.333	2.667	7.112889	
Gemini	22	21.333	0.667	0.44889	
Cancer	19	21.333	-2.333	5.442889	
Leo	21	21.333	-0.333	0.110889	
Virgo	18	21.333	-3.333	11.108889	
Libra	19	21.333	-2.333	5.442889	
Scorpio	20	21.333	-1.333	1.776889	
Sagittarius	23	21.333	1.667	2.778889	
Capricorn	18	21.333	-3.333	11.108889	
Aquarius	20	21.333	-1.333	1.776889	
Pisces	23	21.333	1.667	2.778889	

Calculate chi square statistic

- Step 7: Divide the amounts in Step 6 by the expected value (Step 4) and place those results in the final column.

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	2.755490976
Taurus	24	21.333	2.667	7.112889	0.333421882
Gemini	22	21.333	0.667	0.44889	0.021042048
Cancer	19	21.333	-2.333	5.442889	0.255139408
Leo	21	21.333	-0.333	0.110889	0.005198003
Virgo	18	21.333	-3.333	11.108889	0.520737308
Libra	19	21.333	-2.333	5.442889	0.255139408
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricorn	18	21.333	-3.333	11.108889	0.520737308
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457

Calculate chi square statistic

- Step 8: Add up (sum) all the values in the last column.

Category	Observed	Expected	Residual= (Obs-Exp)	(Obs-Exp)^2	Component = (Obs-Exp)^2 / Exp
Aries	29	21.333	7.667	58.782889	2.755490976
Taurus	24	21.333	2.667	7.112889	0.333421882
Gemini	22	21.333	0.667	0.44889	0.021042048
Cancer	19	21.333	-2.333	5.442889	0.255139408
Leo	21	21.333	-0.333	0.110889	0.005198003
Virgo	18	21.333	-3.333	11.108889	0.520737308
Libra	19	21.333	-2.333	5.442889	0.255139408
Scorpio	20	21.333	-1.333	1.776889	0.083292973
Sagittarius	23	21.333	1.667	2.778889	0.130262457
Capricorn	18	21.333	-3.333	11.108889	0.520737308
Aquarius	20	21.333	-1.333	1.776889	0.083292973
Pisces	23	21.333	1.667	2.778889	0.130262457
					5.094017203

Chi square Table

Critical values of the Chi-square distribution with d degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	d	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

Test a Chi Square Hypothesis

- Test for Independence
  - A chi-square test for independence shows how categorical variables are related. There are a few variations on the statistic; which one you use depends upon how you collected the data.
  - It also depends on how your hypothesis is worded. All of the variations use the same idea; you are comparing the values you expect to get (expected values) with the values you actually collect (observed values).
  - One of the most common forms can be used in a contingency table.

Test a Chi Square Hypothesis

- The chi square hypothesis test is appropriate if you have:
  - Discrete outcomes (categorical.)
  - Dichotomous variables.
  - Ordinal variables.

Test a Chi Square Hypothesis: Steps

- Sample question: Test the chi-square hypothesis with the following characteristics:
  - 11 Degrees of Freedom
  - Chi square test statistic of 5.094
- Note: Degrees of freedom equals the number of categories minus 1.

Test a Chi Square Hypothesis: Steps

- Step 1: Take the chi-square statistic. Find the p-value in the chi-square table. If you are unfamiliar with chi-square tables, the chi square table link also includes a short video on how to read the table.
- The closest value for df=11 and 5.094 is between .900 and .950.
- Note: The chi square table doesn't offer exact values for every single possibility.
- If you use a calculator, you can get an exact value. The exact p value is 0.9265.

Test a Chi Square Hypothesis: Steps

- Step 2: Use the p-value you found in Step 1. Decide whether to support or reject the null hypothesis.
- In general, small p-values (1% to 5%) would cause you to reject the null hypothesis.
- This very large p-value (92.65%) means that the null hypothesis should not be rejected.

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com