S17

Mathematics
Support Centre

Title: Conducting a Manual Chi-Squared test

Target: On completion of this worksheet you should be able to identify key assumptions of, and conduct, a manual Chi-Squared test

Key Information to start:
A Chi-Square goodness of fit test determines whether our sample data matches what we would expect from a population. It shows the relationship between two categorical variables. The Chi-squared statistic is a single number which tells you how much of a difference there is between your observed data and the expected data.

Important information
Chi-square statistic can only be used on numbers, therefore percentages of something need to be converted into a figure to be analysed. A Chi-square with a low value means there is a higher correlation between the two sets of data, meaning they are statistically similiar.

Chi-square score value
There are numerous methods to collect the Chi Square value but all end up using the same/similar formula. We can use a table or enter values straight into the formula. The Chi-square formula is

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where
O_i is our observed values
E_i is our expected values
n is our amount of variables in a category
and i is our position within the data. $i = 1, 2, ..., n$

Using a Table

No.	O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
1	6	12	-6	36	3
2	15	12	3	9	0.75
3	15	12	3	9	0.75
4	7	12	-5	25	2.08
5	15	12	3	9	0.75
6	14	12	2	4	0.33
				$\Sigma\frac{(O-E)^2}{E} =$	7.66

Alternatively we can use a table like above where using our notation:
O are the observed values
E are the expected values
and we use a construction of the forumula for Chi-square in multiple steps to get our Chi-square value which in the case above is 7.66

Hypothesis
Our null hypothesis H_0 is where our sample matches the population, while the alternative hypothesis H_1 is where it doesn't match the population. We use the Chi-square table to find our critical value, which is our reference point. Our degrees of freedom are $v = n - 1$ and we have a significance level α which is generally 95% (or 0.05) depending on where we want to test our data. We compare our critical value with our score value. If our score value is lower than the critical value, we consider this not to be in the rejection region and therefore we do not reject the null hypothesis. However, if it is greater than our critical, it is considered in our rejection region therefore we would reject our null hypothesis.

P.T.O.

Example 1

256 students 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 $= 23$. Use a Chi-square at 95% to test if this can be used to represent a population.

Answer

Creating the table we get

Category	O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Aries	29	21.333	7.667	58.78289	2.755491
Taurus	24	21.333	2.667	7.112889	0.333422
Gemini	22	21.333	0.667	0.444889	0.020854
Cancer	19	21.333	-2.333	5.442889	0.255139
Leo	21	21.333	-0.333	0.110889	0.005198
Virgo	18	21.333	-3.333	11.10889	0.520737
Libra	19	21.333	-2.333	5.442889	0.255139
Scorpio	20	21.333	-1.333	1.776889	0.083293
Sagittarius	23	21.333	1.667	2.778889	0.130262
Capricorn	18	21.333	-3.333	11.10889	0.520737
Aquarius	20	21.333	-1.333	1.776889	0.083293
Pisces	23	21.333	1.667	2.778889	0.130262
				$\Sigma\frac{(O-E)^2}{E} =$	5.09383

Our E is 21.333 because its the mean of the observed values which is where we add up all the observed values then divide by the amount of types of categories, in this case 12. At 95% we check the statistical table (found in the tables book) for the critical value using our degrees of freedom (a particular reference to find our critical value based upon the size of our data/categories) as $12 - 1 = 11$ which gives us our critical value 4.575. Since our score value, 5.09383, is greater than our critical then we have sufficent evidence to reject our null hypothesis, and conclude that this doesn't represent the population.

Exercises

1. Simon wanted to test if cars came past on a certain road on a particular hour everyday of the week would be a good representitative on how many cars pass every day all day long. His results for the hour was: Monday $= 23$. Tuesday $= 18$, Wednesday $= 17$, Thursday $= 19$, Friday $= 23$. Use Chi-square to test its relevance at 5%

2. A psychologist is testing memory on 20 participants after showing them 20 items for a minute to recall when taken away. The psychologist expects to have an average of 12 items remembered. Using chi-square test, see if this is sufficent to be used in a population for this experiment.

3. A clothes company wanted to know what the average height of its shoppers is, to know what size of clothing they should create for length. They took the height of the first customer, or the second if the first customer has already particpated, everyday for 10 days. Their results were, $180cm$, $174cm$, $178cm$, $166cm$, $172cm$, $183cm$, $175cm$, $176cm$, $179cm$, $176cm$. The company wishes its null hypothesis is not rejected at a 99.5% confidence interval. Find an expected height to the nearest whole cm that would not have its hypothesis rejected at 99.5%.
   Preferably using Excel, find an expected range that is not rejected for the clothes company.

Answers

1. Chi-square value is 1.6 and the critical is 9.49. value is lower than critical therefore don't reject null hypothesis.
2. chi-square is 19.25, critical is 30.144, don't reject null hypothesis, this can be used, to a degree, that the population will gain the same result.
3. The range is $173cm$ to $179cm$.