Chi Square ($\chi^2$) Statistical Instructions

EXP 3082L

Jay Gould's Elaboration on Christensen and Evans (1980)

For the Driver Behavior Study, the Chi Square Analysis II is the appropriate analysis below. It also corresponds to a Two-Way Chi Square Analysis in Bruning and Kintz's *Computational Handbook of Statistics (p. 297-304)* and also corresponds to a Two-Way Chi Square Analysis in Linton and Gallo's *The Practical Statistician.* See page 74, as well as page 33 for this analysis, and page 62 for limitations regarding sample size and expected frequencies (e.g. $f_e$ must be $\geq 2$, not 5). If you are interested, see page 88 (along with page 33) of the Linton and Gallo text for information about a Three-Way Chi Square Analysis, which is more complex but would be the most appropriate for the Driver Behavior Study, given the number of variables and levels involved. The information below regarding <u>Chi Square Analysis I</u> is for simpler research designs, and is provided for completeness, but particularly as a learning tool. It should be read first.

## **<u>Chi Square Analysis I (Simple Chi Square Analysis or Goodness of Fit Test)</u>**

Chi Square is a statistic that is used to analyze frequencies and specifically to determine if the frequency of occurrence of two or more events differs significantly from that which would be expected by chance alone. The null hypothesis would be that the frequency of occurrence of the events could be explained by chance and not by some other factor. Consider, <u>for example</u>, the frequency of use of a right and left entrance door to a building. If individuals did not have a preference for the use of one door over the other, then the two doors should be used equally often or should differ in frequency of use to a degree that could be expected by chance alone. If you wanted to find out if individuals demonstrated a preference for one door over the other, your first task would be to count the number of times that individuals use the left versus the right door. Assume you observe 50 people entering the building, and of these 50 people you find that 35 have used the right door and 15 have used the left door. To determine if this preference for the right door is a real preference or instead is within chance variation, you could compute a chi square analysis to help make such an assessment. To calculate the chi square you would follow the steps outlines in the following table, labled Chi Square Analysis I. In using the calculations presented in this table, remember that your observed frequencies of use of the doors are 35 for the right door and 15 for the left door. Also remember that the theoretical expected frequency of use of each door would be 25, since this is the expected frequency of use if individuals did not demonstrate a preference for one of the doors.

Once you have made the calculations illustrated in this table you essentially have computed the chi square since

$$\chi^2 = \Sigma\ (F_o - F_e)^2 / F_e$$

**Chi Square Analysis I**

| Door | $F_o$ | $F_e$ | $F_o - F_e$ | $(F_o - F_e)^2$ | $(F_o - F_e)^2 / F_e$ |
|---|---|---|---|---|---|
| Right | 35 | 25 | +10 | 100 | 100/25=4 |
| Left | 15 | 25 | -10 | 100 | 100/25=4 |
| $\Sigma$ | (50) | (50) | (0) | (200) | $8 = \chi^2$ |

$F_o$ = the observed frequency of use of a given door.

$F_e$ = the theoretical or expected frequency of use of a given door.

$\Sigma$ = sum

Sums are shown at the bottom of the table, with the one at the right being the value of $\chi^2$, i.e. $\chi^2 = 8$. To determine whether the $\chi^2$ value of 8 is significant, the appropriate degrees of freedom (df) must be calculated. The df are calculated by subtracting one from the number of observed groups (k), i.e. df = k–1. Since there were two doors, and therefore two groups, k = 2, df = 2-1 or 1. To determine if the $\chi^2$ of 8 with 1 df is significant turn to the chi square table at the end of the handout. This table is set for two significant $\alpha$ levels, .05 and .01, and for a range of df. You must first decide under which significance level you wish to operate. Assume you choose the .05 level. To determine if the $\chi^2$ value of 8 is significant with 1 df you would look at the table under df

= 1 and $\alpha$ = .05. The *tabulated value* is 3.84. If our *calculated* $\chi^2$ were less than the tabulated value, then the null hypothesis (that there is no difference between the observed and theoretical scores) would be retained. If the calculated value were *larger* than that tabulated, we could reject the null hypothesis and we could state that the observed and theoretical conditions are significantly different at the 95% level of confidence. Since the value we calculated equals 8, we can reject the null hypothesis. Note that in this case had we selected the 1% level of significance, we still would have rejected the null hypothesis.

<u>Chi Square Analysis II, or Complex Chi Square Analysis, or Test of Association, or Two-Way Chi Square Anaysis</u>

The preceding is the simplest form of the chi square statistic and it is suitable only where there is a single dimension of an observed variable, in this case, the location of doors. With a slightly more complicated arrangement, we can determine the dependence or independence of two dimensions of variables. This involves a <u>Complex Chi-Square Analysis</u> as opposed to the preceding <u>Simple Chi-Square Analysis</u>. Assume that you wanted to find out if prior hurricane experience was related to the decision to evacuate in the face of a subsequent hurricane. The two variables would be prior hurricane experience vs. no prior hurricane experience <u>and</u> decision to evacuate vs. decision not to evacuate in face of a subsequent hurricane. Let's say we go into a town where a hurricane was about to hit and divided the population into four categories:

(a) those with prior hurricane experience who decided to evacuate before the current storm arrives;
(b) those with no prior hurricane experience who decided to evacuate;
(c) those with prior hurricane experience who decided not to evacuate; and
(d) those with no prior hurricane experience who decided not to evacuate.

It should be remembered in all chi square determinations that the subjects in the various cells are assumed to be independent and the same subjects cannot be in two cells. The matrix below shows the form and data for our hypothetical research:

# Chi Square Analysis II

| | Decision to evacuate | Decision not to evacuate | Marginal Totals |
|---|---|---|---|
| Prior hurricane experience | a<br>$f_{oA} = 10$<br>$f_{eA} = \frac{60 \text{ X } 40}{105} = 22.85$ | c<br>$f_{oC} = 30$<br>$f_{eC} = \frac{45 \text{ X } 40}{105} = 17.14$ | $f_{oA} + f_{oC} = 40$ |
| No prior hurricane experience | b<br>$f_{oB} = 50$<br>$f_{eB} = \frac{60 \text{ X } 65}{105} = 37.14$ | d<br>$f_{oD} = 15$<br>$f_{eD} = \frac{45 \text{ X } 65}{105} = 27.85$ | $f_{oB} + f_{oD} = 65$ |
| Marginal Totals: | $f_{oA} + f_{ob} = 60$ | $f_{oC} + f_{oD} = 45$ | $f_{oA} + f_{ob} + f_{oC} + f_{oD} = \Sigma = 105$ |

Where

a, b, c, and d = four different cells

$f_o$ = observed frequency

$f_e$ = expected frequency

$f_e$ of a cell = Row's Proportion of observations × Col's Proportion of observations × Total # of observations, that is to say:

$$f_e = \frac{f_o \text{ Row}}{\text{Total } f_o \text{ All cells}} \quad X \quad \frac{f_o \text{ Col}}{\text{Total } f_o \text{ All Cells}} \quad X \quad \text{Total } f_o \text{ All Cells}$$

This reduces down to : $f_e = \frac{f_o \text{ Row} \times f_o \text{ Col}}{\text{Total } f_o \text{ All Cells}}$

In this method of computing the $\chi^2$ statistic, the $f_e$ must be calculated for each cell by, as we have just seen, multiplying the $f_o$ row total by the $f_o$ column total corresponding to a given cell, and then dividing by the sum of the $f_o$ values for all four cells. The value of $f_e$ for cell A would be calculated as follows:

$$f_{eA} = \frac{(f_{oA} + f_{oB}) \quad (f_{oA} + f_{oC})}{f_{oA} + f_{oB} + f_{oC} + f_{oD}} = \frac{60 \times 40}{10 + 50 + 30 + 15} = \frac{2400}{105} = 22.85$$

Once the $f_e$ values have been calculated for all cells, then the additional necessary values for $\chi^2$ can be calculated as follows:

| Cell | $f_o - f_e$ | $(f_o - f_e)^2$ | $\frac{(f_o - f_e)^2}{f_e}$ |
|---|---|---|---|
| a | 10 – 22.85 = –12.85 | 165.12 | 165.12 / 22.85 = 7.22 |
| b | 50 – 37.14 = 12.86 | 165.38 | 165.38 / 37.14 = 4.45 |
| c | 30 – 17.14 = 12.86 | 165.38 | 165.38 / 17.14 = 9.65 |
| d | 15 – 27.85 = –12.85 | 165.12 | 165.12 / 27.85 = 5.93 |

Once these values are calculated you have computed all the necessary values for the determination of the $\chi^2$ value.

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} = (7.22 + 4.45 + 9.65 + 5.93) = 27.25$$

The *df* for this $\chi^2$ is computed by calculating the following formula.

$df = (C - 1) \quad (R - 1)$ C = number of columns, R = number of rows

Since the matrix we formed was a 2 row by 2 column matrix, *df* = (2 – 1) × (2 – 1) = 1.With a $\chi^2$ value of 27.25 and 1 *df*, we again go to the chi square table to determine if the $\chi^2$ is signifcant.

Going to the table of chi square and looking under the 5% significance level ($\alpha$) for *df* = 1, we find the critical value 3.84. Since our calculated value is 27.25, we are able to reject the null hypothesis of no difference and state that there is a significant relationship between the two variable dimensions. Note: this is true at the .01 level as well as the .05 level.

**Table 1. Distribution of $\chi^2$ ***

| n (*df*) | $\alpha$ = .05 | $\alpha$ = .01 |
|---|---|---|
| 1 | 3.84 | 6.64 |
| 2 | 5.99 | 9.21 |
| 3 | 7.82 | 11.34 |
| 4 | 9.49 | 13.28 |
| 5 | 11.07 | 15.09 |
| 6 | 12.59 | 16.81 |
| 7 | 14.07 | 18.48 |
| 8 | 15.51 | 20.09 |
| 9 | 16.92 | 21.67 |
| 10 | 18.31 | 23.21 |
| 11 | 19.68 | 24.72 |
| 12 | 21.03 | 26.22 |
| 13 | 22.36 | 27.69 |
| 14 | 23.68 | 29.14 |
| 15 | 25.00 | 30.58 |
| 16 | 26.30 | 32.00 |
| 17 | 27.59 | 33.41 |
| 18 | 28.87 | 34.80 |
| 19 | 30.14 | 36.10 |
| 20 | 31.41 | 37.57 |
| 21 | 32.67 | 38.93 |
| 22 | 33.92 | 40.29 |
| 23 | 35.17 | 41.64 |
| 24 | 36.42 | 42.98 |
| 25 | 37.65 | 44.31 |
| 26 | 38.88 | 45.64 |
| 27 | 40.11 | 46.96 |
| 28 | 41.34 | 48.28 |
| 29 | 42.56 | 49.59 |
| 30 | 43.77 | 50.89 |

*Table 1 is taken from Table III of Fischer and Yates: Statistical Tables for Biological, Agricultural and Medical Research, published by Longman Group Ltd., London (previously published by Oliver and Boyd, Edinburgh), and by permission of the authors and publishers.