Pearson's Chi-Square Test of Independence -- Analysis of the Relationship between two Qualitative Variables

or **Analysis of k-Between-Group Data with a Qualitative Response Variable**

Application: This statistic has two applications that can appear very different, but are really just two variations of the same question. The first application is to compare the distributions of scores on a qualitative response variable) obtained from 2 or more groups that represent different populations. Thus, it is applied in the same data situation as an ANOVA for independent samples, except that it is used when the response variable is a qualitative variable. The other application is to test for a pattern of relationship between two qualitative variables -- in this respect, it is something like a correlation (though looking for a *pattern of* relationship between qualitative variables, rather than a *linear* relationship between quantitative variables).

There are two versions of the H0:, depending upon whether one characterizes the analysis as a test of whether the populations represented by one of the variables differ in their patterns of response to the response variable (the other variable -- which corresponds to the first application described above) or as a test of whether there is a relationship between the two variables in the single population represented by the sample as a whole (which corresponds to the second application described above). The H0: for each is given below.

H0: The populations represented by the conditions have the same pattern of responses across the categories of the response variable.

To reject H0: is to say that the populations differ in their response patterns to the categories of the response variable.

H0: The variables have no pattern of relationship within the population represented by the sample.

To reject H0: is to say that there is a pattern of relationship between the variables in the population.

The data: The analysis involves the grouping variable **reptdept** (1 = not separate reptile department, 2 = separate reptile department) and the response variable **reptiles** (1 = display only lizards, 2 = display only snakes, 3 = display both lizards and snakes). The data were collected from 90 pet stores in the midwest.

Researcher Hypothesis: The researcher hypothesized that stores without a separate reptile department would be more likely to display only type of reptile one or the other, because of limited shelf space, whereas those stores with a separate reptile department would be more likely to display both types. (The researcher is characterizing this as a comparison across populations represented by the conditions of reptdept).

H0: for this analysis : There is no pattern of relationship between whether or not pet stores have separate reptile departments and whether they display only lizards, only snakes or both.

Step 1 Organize the scores into a contingency table. Since one of these categorical variables (reptdept) has two categories and the other (reptiles) has three, the contingency table will be a 2x3, for a total of 6 cells, as shown below.

		Reptiles	
Reptdebt	Lizards	Snakes	Both
Not separate			
Separate			

Each store's data will be collated into one of the six cells. For example, a store that did not have a separate reptile department and that displayed only lizards would be tallied into the cell in the upper left; a store that had a separate reptile department and displayed both lizards and reptiles would be tallied into the cell in the lower right. Below is the contingency table filled with the responses from the 90 stores . These values are the obtained frequency (**of**) for each of the cells.

Reptdebt	lizards	Reptiles Snakes	Both
Not separate	19	20	8
Separate	5	6	32

Step 2 Compute the row totals (sum across the values on the same row) and column totals (sum across the values on the same column) of the observed frequencies.

Reptdept	Lizards	Reptiles Snakes	Both	row total
Not separate	19	20	8	47
Separate	5	6	32	43
				overall total
column totals	24	26	40	90

Step 3 Compute the overall total (shown in the table above). As a computational check, be sure that the row totals and the column otals sum to the same value for the overall total.

Step 4 Compute the expected frequency (**ef**) of each cell. This expected frequency is computed as the product of the row total and the column total for that cell, divided by the overall total. For example, the upper left-hand cell (stores not having a separate reptile department that diplay only lizards) has an expected frequency of:

$$ef = \frac{\text{row total} * \text{col total}}{\text{overall total}} = \frac{47 * 24}{90} = 12.53$$

Reptdebt	Lizards		Reptiles Snakes		Both	
	of	ef	of	ef	of	ef
Not separate	19	12.53	20	13.58	8	20.89
Separate	5	11.47	6	12.42	32	19.11

Step 5 Compute Chi-Square

$$X^2 = \sum \frac{(of - ef)^2}{ef} = \frac{(19-12.53)^2}{12.53} + \frac{(20-13.58)^2}{13.58} + \frac{(8-20.89)^2}{20.89} + \frac{(5-11.47)^2}{11.47} + \frac{(6-12.42)^2}{12.42} + \frac{(32-19.11)^2}{19.11}$$

$$= 3.34 + 3.04 + 7.95 + 3.65 + 3.32 + 8.69 = 29.99$$

Step 6 Compute the degrees of freedom (df)

$$df = (\text{number of columns} - 1) * (\text{number of rows} - 1) = (3-1) * (2-1) = 2 * 1 = 2$$

Step 7 Use the Table X^2 to determine the critical Chi-square value for df = 2 and p = .05

$$X^2(df=2, p=.05) = 5.991$$

Step 8 Compare the obtained X^2 and critical X^2, and determine whether or not there is a statistically significant relationship between the two categorical variables.

-- if the obtained X^2 is less than the critical X^2, then retain the null hypothesis -- conclude that there is no relationship between subject's values on one categorical variable and their values on the other categorical variable, in the population represented by the sample

-- if the obtained X^2 is greater than the critical X^2, then reject the null hypothesis -- conclude that there is a relationship between the subject's values on one categorical variable and their values on the other categorical variable, in the population represented in the sample.

For the example data, we would decide to reject the null hypothesis, because the obtained Chi-square value of 29.99 is larger than the critical Chi-square value of 5.991.

By the way: This test should only be applied when at least 80% of the cells have expected frequencies (**ef**) of five or larger. Applying the test when there are fewer cells with this minimum expected frequency can lead to inaccurate results.

Step 9 IF you reject the null hypothesis, determine whether the pattern of the data in the contingency table completely supports, partially supports, or does not support the research hypothesis.

-- IF you reject the null hypothesis, AND if the pattern of data in the contingency table agrees exactly with the research hypothesis, then the research hypothesis is completely supported.

-- IF you reject the null hypothesis, AND if part of the pattern of data in the contingency table agrees with the research hypothesis, BUT part of the pattern of data does not, then the research hypothesis partially supported.

-- IF you retain the null hypothesis, OR you reject the null BUT NO PART of the pattern of data in the contingency table agrees with the research hypothesis, then the research hypothesis is not at all supported.

By the way: Usually the researcher hypothesizes that there is a pattern of relationship between the variables. Sometimes, however, the research hypothesis is that there is NO pattern of relationship. If so, the research hypothesis and H0: are the same! When this is the case, retaining H0: provides support for the research hypothesis, whereas rejecting H0: provides evidence that research hypothesis is incorrect.

Consistent with the research hypothesis, those stores without separate reptile departments tended to display either lizards or snakes (but not both), whereas those stores that had separate reptile departments tended to display both types of reptiles.

Step 10 Reporting the results

It is important to describe the univariate data before telling whether or not there is a pattern of relationship between the variables. Report the number (or percentage) that fall into the categories of each variable (showing the contingency table will help the reader). As for the other statistical tests, the report includes the "wordy" part and the statistical values upon which you made your statistical decision. If H0: is rejected, be sure to describe the pattern of the relationship.

For the sample of 90 stores, there was about an equal number that had and did not have separate reptile departments. With regard to the types of reptiles displayed, about an equal number displayed only snakes as displayed only lizards, with somewhat more displaying both types. As hypothesized, the different types of stores tended to display different types of reptiles ($X^2(2)=29.987$, $p=.001$). More of the stores without separate reptile departments displayed only one type of reptile, whereas those with separate departments displayed both types of reptiles.

C^2 Critical values of Chi-Square

df	$\alpha = .05$	$\alpha = .01$
1	3.84	6.63
2	5.99	9.21
3	7.81	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	34.81
19	30.14	36.19
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	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
¥	43.77	50.89