

## Cochran's Q Test -- Analysis of 2-Within-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 statistical question. In one application the same qualitative (binary) variable is measured at two or more different times from the same sample (or from two or more samples that have been matched on one or more important variables). In the other application, two or more comparable qualitative variables are measured from the same sample (usually at the same time). In both applications, the Cochran's Q Test is used to compare the distributions of the two qualitative variables.

There are two specific versions of the  $H_0$ , depending upon whether one characterizes the conditions as representing a single population under two or more different circumstances (e.g., comparing "before vs. after" as in the example below – some consider this a comparison of different populations "with vs. without") or as representing comparable variables measured from a single population (e.g., if the variables produce the same assignment of individuals). Here are versions of the  $H_0$  statement for each of these characterizations.

**$H_0$ :** The frequencies (or proportions) of responses to the categories of the response variable is the same across the conditions, for the population represented by the sample.

**To reject  $H_0$ :** is to say that within this population, subjects respond differently across the conditions.

**$H_0$ :** Members of this population are assigned into the same categories across the response variables.

**To reject  $H_0$ :** is to say that cases are categorized differently using the different response variables).

**The data:** The researcher who had collected the Pet Shop data wanted to examine whether pet stores displayed different types of reptiles during different times of the year. So, the researcher visited each of the 12 stores four times during the next year that were chosen because of their proximity to holidays, Valentine's Day, July 4, Halloween, Christmas. During each visit, the researcher recorded if the shop displayed only snakes or lizards (coded = 1) or both types of reptiles (coded = 2). In this analysis the one variable is the time of year (Valentine's Day, July 4, Halloween, or Christmas) and the response variable is the type of reptile(s) displayed. From our new database we will use four variables, all coded the same (1=snakes or lizards, 2 = snakes and lizards), **valday** (Valentine's Day), **jul4** (July 4<sup>th</sup>), **hal** (Halloween), **cmas** (Christmas). Here are the data from the 12 stores:

1, 1, 1, 2	1, 1, 1, 2	1, 1, 1, 2	2, 2, 2, 2	2, 1, 1, 2	1, 2, 1, 2
2, 1, 1, 2	1, 1, 1, 2	1, 2, 1, 1	1, 1, 1, 1	2, 1, 1, 2	1, 1, 2, 2

**Research Hypothesis:** The researcher hypothesized that pet shops would be more likely to display both types of reptiles prior to Christmas than during the other times of the year.

**$H_0$  for this analysis:** Stores are equally likely to display both types of reptiles during all parts of the year.

**Step 1** Arrange the data into separate columns for each condition, with the scores for each subject in a separate row. For the computation of this statistic, the response variable scores must be coded as 0 and 1 (usually "1" is used to code a "success" if that is applicable to the responsible variable). Select one of the conditions to be represented by each value. We will use 0 to indicate the store displayed either snakes or lizards and use 1 to indicate the store displayed both snakes and lizards. With a little planning, the data can be collected using the 0-1 codes and you can avoid the need for this "translation".

Valentine's Day	July 4 <sup>th</sup>	Halloween	Christmas
0	0	0	1
0	0	0	1
0	0	0	1
1	1	1	1
1	0	0	1
0	1	0	1
1	0	0	1
0	0	0	1
0	1	0	0
0	0	0	0
1	0	0	1
0	0	1	1

**Step 2** Compute the sum for each column (referred to as G).

$$\text{Valentine's Day} \quad G_1 = 0 + 0 + 0 + 1 + 1 + 0 + 1 + 0 + 0 + 1 + 0 + 0 = 4$$

$$\text{July 4}^{\text{th}} \quad G_2 = 0 + 0 + 0 + 1 + 0 + 1 + 0 + 0 + 1 + 0 + 0 + 0 = 3$$

$$\text{Halloween} \quad G_3 = 0 + 0 + 0 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 1 = 2$$

$$\text{Christmas} \quad G_4 = 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 0 + 0 + 1 + 1 = 10$$

**Step 3** Compute the sum for each row (referred to as L), and then compute the square of this sum ( $L^2$ ).

Valentine's Day	July 4 <sup>th</sup>	Halloween	Christmas	L	$L^2$
0	0	0	1	1	1
0	0	0	1	1	1
0	0	0	1	1	1
1	1	1	1	4	16
1	0	0	1	2	4
0	1	0	1	2	4
1	0	0	1	2	4
0	0	0	1	1	1
0	1	0	0	1	1
0	0	0	0	0	0
1	0	0	1	2	4
0	0	1	1	2	4

**Step 4** Compute the sum of L ( $\Sigma L$ ) and  $L^2$  ( $\Sigma L^2$ ).

$$\Sigma L = 1 + 1 + 1 + 4 + 2 + 2 + 2 + 1 + 1 + 0 + 2 + 2 = 19$$

$$\Sigma L^2 = 1 + 1 + 1 + 16 + 4 + 4 + 4 + 1 + 1 + 0 + 4 + 4 = 41$$

**Step 5** Determine k, the number of conditions.

$$k = 4$$

**Step 6** Compute Q, using the following formula.

$$Q = \frac{(k-1) * [ (k * \Sigma G^2) - (\Sigma G)^2 ]}{(k * \Sigma L) - \Sigma L^2} = \frac{(4-1) * [ 4 * (4^2 + 3^2 + 2^2 + 10^2) - (4+3+2+10)^2 ]}{(4 * 19) - 41}$$

$$= \frac{3 * [ 4 * 129 - 361 ]}{76 - 41} = \frac{465}{35} = 13.286$$

**Step 7** Use the Chi-square table to determine the critical  $X^2$  value for  $df = k - 1$  and  $p = .05$

$$X^2(df=3, p=.05) = 7.82$$

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

-- if the obtained Q 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 Q 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 by the sample.

For the example data, we would decide to reject the null hypothesis, because the obtained Q value of 13.286 is larger than the critical Chi-square value of 7.82.

**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  $H_0$  are the same! When this is the case, retaining  $H_0$  provides support for the research hypothesis, whereas rejecting  $H_0$  provides evidence that the research hypothesis is incorrect.

For the example data, we would conclude that there was complete support for the research hypothesis, because the null hypothesis was rejected and because, as hypothesized, there were more stores that displayed both types of reptiles prior to Christmas.

## **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 response variable and the conditions. Report the number (or percentage) that fall into the "success" category for each condition (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  $H_0$  is rejected, be sure to describe the pattern of the relationship.

The percentage of stores that displayed both snakes and lizards was 25% during Valentine's Day, 33% during July 4<sup>th</sup>, 17% during Halloween, and 83% during Christmas ( $Q(3) = 13.286$ ,  $p < .05$ ). As hypothesized, there were more stores displaying both types of reptiles during the Christmas buying season than during the other times of the year.

## C<sup>2</sup> 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