STATISTICAL SUPPORT
SWISS CENTER FOR AFFECTIVE SCIENCES

Generalized chi-square test for dependent data

E-mail distributed on 12-01-2026

Dear all,

A traditional chi-square test is applied to a 2-dimensional frequency table (e.g., A×B) that tabulates the frequency at which each combination of A and B levels occurs. The chi-square test evaluates whether the distribution of these frequencies in the A levels differs between B levels. Essentially, it is an A×B interaction test with the cell frequencies as the outcome variable.[1]

An important assumption of the chi-square test is that observations are independent. That is, each subject should be counted only once, in one cell of the frequency table. This assumption will not hold for study designs where subjects are observed in more than one level of A and/or B simultaneously, such as within-subject designs, or studies where a categorical outcome is either measured repeatedly, or represents a non-exclusive choice. For repeated categorical outcomes I have previously discussed analysis methods for the special case of paired categorical data. However, this requires a square frequency table with identical A and B levels, and an analysis model that ignores the table's diagonal cells.[2] Instead of a specialized model, we would ideally have a chi-square test that generalizes to A×B structures with arbitrary dependencies. In this mailing, I discuss two such options, **(a)** the **Rao-Scott chi-square test**, and **(b)** a model-based test using **multinomial generalized estimating equations (GEE)**.

1. Toy data example

We consider the following example, from a study where participants were exposed to 4 virtual heights to induce fear (0, 5, 25, 50 m). Afterwards, for each depth, they chose a word that best described their feelings from a list of ≈30. Words were then classified into 6 global themes. A cross-tabulation of these data returns the following:

```
     label
depth Joy None Interest Surprise Stress Fear
   0   23   49       11       12      1    0
   5   28   26       39        2     12    2
   30  27    0       33        9     23   18
   50  22    0       26        4     24   33
```

[1] See my workshop on logistic regression for an introduction to the analysis of frequency tables.

[2] In paired categorical data diagonal dependencies tend to be trivially high, due to repeated measures correlation.

Descriptively, it is apparent some word categories were chosen more frequently for specific depths, such as “None” for low depths, and “Fear” for high depths. Others, like “Joy”, seem to have no apparent relationship with depth. A naïve chi-square test, assuming independent observations, returns the following:

```
        Pearson's Chi-squared test
data:  xtabs(~depth + label, data = fear)
X-squared = 197.58, df = 15, p-value < 2.2e-16
```

So highly significant, supporting a relationship between depth and word choice. However, each participant was exposed to all 4 depths, and therefore the cell counts in the frequency table are not independent. When the dependency between repeated measures is strong (e.g., participants tend to prefer the same word category for describing their feelings) this may bias the chi-square test. The **Rao-Scott chi-square test** aims to correct this by first estimating the variance inflation due to repeated measures, and then adjusting the traditional chi-square test by this inflation factor. This test has historically been applied in survey studies, where the repeated measures structure is referred to as the data’s “design,” and subject identifiers such as ID code are called a “cluster variable.”

We can proceed in R with functions from the `survey` package (Lumley, 2004). Crucially, we cannot enter our data directly in its tabular format this time, they must be in the original, subject-level **long-format**, e.g.:

```
    ID depth         word    label
1  HZI     0 Anticipation Interest
2  HXG     5     Aversion     Fear
3  DFN     0      Boredom     None
4  MGX    50    Amusement      Joy
5  BEJ    50       Terror     Fear
6  BLR    30          Awe      Joy
```

Next, we define the repeated measures design and run the adjusted chi-square test:

```
des <- svydesign(id=~ID, data=fear)
svychisq(~depth + label, design=des, statistic="F")

        Pearson's X^2: Rao & Scott adjustment
data:  svychisq(~depth + label, design=des, statistic="F")
F = 9.9625, ndf = 10.099, ddf = 807.955, p-value = 5.334e-16
```

While the test is still significant, this time we get a larger *p*-value and down-corrected degrees of freedom. Despite its name, the Rao-Scott chi-square test appears to return an *F*-test. While this is its most common form, proper chi-square approximations are available, and can be obtained in R by changing the test statistic, with `"adjWald"` performing well for smaller tables. Of further note in this output are the fractional DFs, which should be a familiar feature from other statistical tests that correct against violations of variance assumptions, such as the Welch *t*-test, general heteroscedastic estimators, non-sphericity corrections, and multilevel models.

A completely different approach to the analysis of dependent frequency tables is to consider an explicit model. For repeated categorical outcomes, we typically have the choice between a marginal model, such as **generalized estimating equations (GEE)**, and a hierarchical model, such as generalized linear mixed models (GLMM). The GEE model is best suited to this problem (as explained in an earlier mailing). However, the conventional R package for this, `geepack` (Højsgaard, Halekoh & Yan, 2006), lacks flexibility somewhat due to only allowing binary outcomes at the lowest measurement level. For scenarios where either A or B is a 2-level factor, this would suffice, but it does not work for variables with an arbitrary number of levels. Fortunately, the package `multgee` (Touloumis, 2016) can fit GEE models for repeated multinomial responses. For the virtual depth data, we can proceed as follows:

```
MLN1 <- nomLORgee(label~depth, data=fear, id=ID, LORstr="independence")
MLN0 <- nomLORgee(label~1, data=fear, id=ID, LORstr="independence")
waldts(MLN0,MLN1)

Goodness of Fit based on the Wald test
Model under H_0: label ~ 1
Model under H_1: label ~ depth
Wald Statistic = 65.2131, df = 15, p-value < 0.0001
```

Basically, we conduct a "manual" ANOVA-like test by first fitting a full model and a reduced model, and then comparing both with a traditional Wald test, which produces a chi-square distributed test statistic. The procedure for model fitting should be familiar to R users, such as the use of formulas and the specification of a subject variable.

Less familiar may be the `"LORstr"` argument. Here, one has to supply a so-called working correlation structure for the repeated measures, much in the same as one would specify, e.g., sphericity in repeated measures ANOVA. However, the exact specification is somewhat less important in GEE models, due to their use of robust standard errors for inference, regardless of the correlation structure supplied. For `nomLORgee`, one could try `"independence"` versus `"time.exch"`, and even compare fits by the quasi-information criterion (QIC), although test results may remain largely unaltered.

2. Strengths and weaknesses

The two approaches discussed each have their own strengths and weaknesses. The Rao-Scott test is fast and simple to apply, so practically speaking should probably be the first choice for a non-independent chi-square test. As well, a notable advantage of this test is that it is symmetric. That is, one will obtain the same result for the A×B table as for the B×A table, just as the naïve chi-square test. The GEE approach, by contrast, is "directional," in that it requires an explicit choice of an outcome variable, and will not be symmetric. That is, model B~A may produce a different conclusion than model A~B. This is a consequence of both the nonlinear nature of GLMs, and the presence of the subject clustering variable. In practice, however, one may find that both models yield the same conclusion at least qualitatively.

Another drawback of the GEE approach is that it may not be applicable to simple repeated measures designs, such as ones where the number of repeated measures equals the number of unique levels in the within-subject variable (e.g., as in the paired *t*-test). Generally, it requires a sufficient number of repeated measures to be run, as well as (preferably) large samples.

The major advantage of the GEE approach, on the other hand, is that it allows model-based extensions, such as **(a)** covariates, **(b)** moderators, **(c)** offset terms, and **(d)** case weights. In principle, the GEE approach could extend to frequency tables of higher dimensions (e.g., A×B×C), with the Wald test comparison affording the flexibility to compare any two nested models. This would not be possible with the Rao-Scott test.

References

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). Hoboken, NJ: Wiley.

Højsgaard, S., Halekoh, U. & Yan J. (2006). The R package geepack for deneralized estimating equations. *Journal of Statistical Software*, 15(2), 1--11

Lumley, T. (2004). Analysis of complex survey samples. Journal of Statistical Software. 9(1), 1–19

Touloumis A. (2015). R package multgee: A generalized estimating equations solver for multinomial responses. *Journal of Statistical Software*, 64(8), 1–14.

--

Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79