Testing Independent Relationships

STEFAN T. GRIES

Introduction

In one of the most frequent empirical scenarios in applied linguistics, a researcher's empirical results can be summarized in a two-dimensional frequency table in which

- the rows list the levels of a categorical variable;
- the columns list the levels of another categorical variable;
- the cells in the table defined by these row and column levels provide the frequencies with which combinations of row and column levels were observed in some data.

An example of data from a study of verb–particle constructions (*John picked*$_{\text{VERB}}$ [*the book*]$_{\text{DO}}$ *up*$_{\text{PART}}$ vs. *John picked*$_{\text{VERB}}$ *up*$_{\text{PART}}$ [*the book*]$_{\text{DO}}$) from Peters (2001) is shown in Table 1, which shows the distribution of 397 constructions depending on whether the referent of the direct object (DO) is a discourse given or new.

A researcher may be interested in whether there is a correlation between the DO's givenness—the row variable—and the construction a speaker produced with that DO—the column variable. A first superficial glance suggests that given DOs are used more often in V-Part-DO (100) than in V-DO-Part (85), but an actual statistical test is required to determine (a) whether the distribution of the constructions with the DOs is significantly different from chance and (b), if so, what preferences and dispreferences this data set reflects. The most frequent statistical test to analyze two-dimensional frequency tables such as Table 1 is the chi-squared test for independence.

The Chi-squared Test for Independence

The chi-squared test for independence is introduced here using the open-source statistical language and environment R (compare R Core Team 2022), which can be freely downloaded from <https://cran.r-project.org/> and which runs on all major operating systems.

Entering the Data

The first step in the analysis of two-dimensional frequency tables is to start the R program and enter the frequency table into R. For example, to enter Table 1, the researcher would type the following at the console prompt (where `c` means "combine values into a vector," or sequence, `ncol` specifies the number of columns into which the sequence of numbers should be coerced, <- represents an assignment arrow, and the Pilcrow sign ¶ means "press ENTER"):

```
Tab1 <- matrix(c(85, 65, 100, 147), ncol=2)¶
```

This creates the matrix of the frequencies shown in Table 1 and stores it as an object called `Tab1`. It is also good practice to add row/column names (as an attribute called `dimnames`) like this (comments after # are ignored by R and are added for the human reader's convenience):

The Encyclopedia of Applied Linguistics. Edited by Carol A. Chapelle.

DOI: 10.1002/9781405198431.wbeal1202.pub2

Table 1 Verb–particle constructions and their correlation with the givenness of the direct object

	V-DO-Part	*V-Part-DO*	*Sum*
Given	85	100	185
New	65	147	212
Sum	150	247	397

```
attr(Tab1, "dimnames") <- list( # assign row & column names
   Givenness=c("given", "new"),
   Construction=c("V-DO-Part", "V-Part-DO"))¶
```

If one now tells R to display the object `Tab1`, then the data look exactly like Table 1:

```
Tab1 # show Tab1¶
         Construction
Givenness V-DO-Part V-Part-DO
    given        85       100
    new          65       147
```

The row and column totals can be obtained from the function `addmargins`:

```
addmargins(Tab1) # show Tab1 with row & column sums¶
         Construction
Givenness V-DO-Part V-Part-DO Sum
    given        85       100 185
    new          65       147 212
    Sum         150       247 397
```

Assumptions

The second step involves determining whether the data can in fact be tested with a chi-squared test for independence. This test has three assumptions of two different kinds: The first involves the independence-of-data-points assumption, and the next two involve the frequencies that would be expected if the data were randomly distributed. The three assumptions are the following:

- all observations are independent of one another;
- 80% of the expected frequencies are ≥5;
- all expected frequencies are >1.

The first assumption requires the researcher to consider whether data points—individual occurrences of a construction with a certain DO—are related to each other. This would be the case if, for example, one and the same speaker provided more than one data point to the data set. In such cases, an individual speaker's preference for a particular construction, or a particular construction with a certain kind of object, could bias the statistical evaluation of the data. Another threat to independence could arise if constructional choices were from different speakers but from successive turns in some conversation, because then a constructional choice in turn t might be influenced by the one in turn $t - 1$ because of priming effects. We will assume that this is not the case here because, for instance, each collected construction is from a different speaker in a different conversation.

The other two assumptions of the chi-squared test will be tested with the function to run the test itself in the following section. Testing the assumptions of the chi-squared test is important because when the data violate the assumptions of the test, its results cannot be trusted: If (too many) expected frequencies are too small, the test can become more likely to return a significant result than it should,

and if the data points are not independent of each other, then the computation of the expected frequencies will be biased. Making sure that the test's assumptions are met is therefore paramount.

Computing and Interpreting the Test

The chi-squared test can be computed easily with a function called `chisq.test`, which takes two arguments: (a) the two-dimensional table for which one wants to compute a chi-squared test (here, `Tab1`) and (b) an argument `correct` that is set to `TRUE` or `FALSE`, depending on whether or not one wants to use a so-called continuity correction, which has sometimes been recommended for smaller sample sizes. Since this recommendation is controversial, we will not use it and discuss a better way to deal with small sample sizes below. The researcher can then assign the result of the chi-squared test to an object `Tab1.test` (output is abbreviated):

```
Tab1.test <- chisq.test( # compute a chi-squared test
   Tab1,                 # on the matrix Tab1
   correct=FALSE)        # with no correction for continuity
Tab1.test                # show the result
    Pearson's Chi-squared test
X-squared = 9.8191, df = 1, p-value = 0.001727
```

Here, there is a very significant correlation between the DO's givenness and the constructional choice: p is smaller than the usual critical value of $p = 0.05$. However, we first need to determine whether we can take this result at face value, given the two additional assumptions regarding the chi-squared test's expected frequencies. For this, the researcher can just print the part of the test results that contains the (rounded) expected frequencies:

```
round(Tab1.test$expected, 2) # show the rounded expected frequencies¶
          Construction
Givenness V-DO-Part V-Part-DO
    given      69.9     115.1
    new        80.1     131.9
```

Obviously, all expected frequencies meet both assumptions—the application of the chi-squared test was justified. Therefore, the next step is to determine which of the four cells in `Tab1` is/are most responsible for this effect, and what to focus on most in the interpretation. To identify these cells, one can inspect the so-called Pearson residuals.

```
round(Tab1.test$res, 2) # show the rounded residuals¶
          Construction
Givenness V-DO-Part V-Part-DO
    given      1.81     -1.41
    new       -1.69      1.31
```

If the Pearson residual in a cell is positive, the observed frequency in that cell is greater than the expected frequency in that cell; if the Pearson residual in a cell is negative, the observed frequency is less than the expected frequency; and the more a Pearson residual deviates from 0, the stronger the effect in that cell. Here, the strongest effect is the preference of V-DO-Part with given DOs (observed frequency: 85 and expected frequency: 69.9). Note how this analysis relativizes our superficial assessment above ("given DOs are preferred in V-Part-DO"): The chi-squared test shows that the 100 occurrences of given DOs in V-Part-DO actually instantiate an *under* representation.

Graphical Interpretation and Effect Size

The above kind of interpretation of chi-squared tests can often be facilitated considerably with graphical displays. Figure 1 shows a mosaic plot, which can be created with the following line:

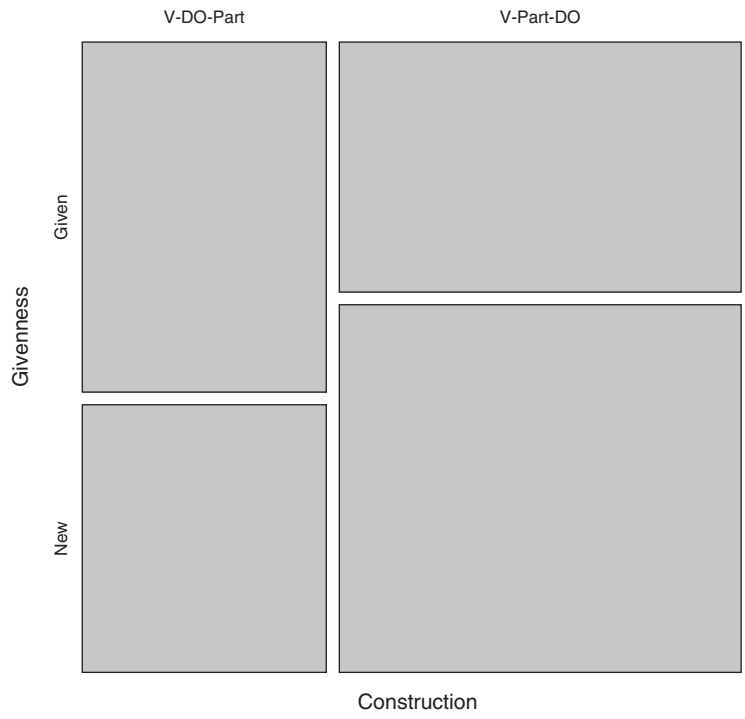

Figure 1 A mosaic plot for the data in Table 1

```
mosaicplot(t(Tab1), main="") # plot a mosaic plot¶
```

In Figure 1, the box sizes are proportional to the cell frequencies and a lack of alignment of the margins between the boxes indicates correlational structure. For example, the fact that the upper left box is longer (vertically) than the upper right box indicates that V-DO-Part is more associated with given DOs than V-Part-DO.

Finally, in order to be able to compare results from different studies, one can compute an effect size, which is independent of the sample size. For two-dimensional tables, a statistic called Cramer's V is often used. It theoretically falls between 0 ("no association") and 1 ("perfect association") and is computed as shown in (1), where min(r, c) means "the minimum of the numbers of rows and columns":

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min(r, c) - 1)}} \tag{1}$$

In this example, the effect size can be computed with the following code, where `sqrt` means "square root," `Tab1.test$statistic` represents the chi-square value of the chi-squared test stored in `Tab1.test`, `sum(Tab1)` represents the sample size n, and `dim(Tab1)` returns the numbers of rows and columns of `Tab1`, of which then the minimum `(min)` is taken. The resulting Cramer's V value is fairly small, certainly much closer to 0 than to 1:

```
 >    sqrt( # compute the square root of this fraction:
 +    Tab1.test$statistic /            # numerator
 +    (sum(Tab1) * (min(dim(Tab1))-1))) # denominator;
X-squared
0.1572683
```

To report the result of a chi-squared test, the researcher should provide the table of observed frequencies, the chi-squared value with its *df*- and *p*-value, and Cramer's *V*. The following section discusses very briefly one way to proceed if the expected frequencies are too small for a chi-squared test.

An Exact Alternative: Fisher–Yates Test

Sometimes, one may have data that result in expected frequencies too small to meet assumptions 2 and 3 of the chi-squared test. For instance, what if Peters had only obtained the data shown here as the hypothetical table `Tab1.hyp`.

```
Tab1.hyp <- floor(Tab1/15)¶
Tab1.hyp¶
          Construction
Givenness V-DO-Part V-Part-DO
    given         5         6
    new           4         9
```

A chi-squared test on `Tab1.hyp` shows that too many of the expected frequencies are smaller than 5:

```
round(chisq.test(Tab1.hyp, correct=FALSE)$expected, 2)¶
          Construction
Givenness V-DO-Part V-Part-DO
    given      4.12      6.88
    new        4.88      8.12
```

In such cases, one could apply the Fisher–Yates exact test, whose application in R is very straightforward (output is abbreviated):

```
fisher.test(Tab1.hyp) # compute Fisher-Yates exact test¶
 Fisher's Exact Test for Count Data
p-value = 0.6752
```

The test shows that the distribution in `Tab1.hyp` is not significantly different from chance: *p* is large, meaning the distribution is too compatible with the null hypothesis of no relation to accept the alternative hypothesis.

An Alternative Approach and Its Extension to Multidimensional Tables

While the chi-squared test is probably the most widely used test for frequency tables, the alternative of *G*-squared is also important. Similar to chi-squared, G^2 is based on comparing observed to expected frequencies; it is computed over the *c* (here, $c = 4$) cells of the table as shown in (2):

$$G^2 = 2 \sum_{1}^{c} \text{observed} \times \log \frac{\text{observed}}{\text{expected}} \tag{2}$$

In R, this can be done like this:

```
2*sum(Tab1 * log(Tab1 / chisq.test(Tab1, correct=FALSE)$expected))¶
[1] 9.835021
```

The value is close to the chi-squared value we computed above (9.8191) and is also tested for significance against the chi-squared distribution:

```
pchisq(9.835021, df=1, lower.tail=FALSE)¶
[1] 0.001712203
```

Obviously, the result is very similar to what we obtained above, but G^2 is still a very important measure to know. This is for two reasons: First, there are some areas of research in applied linguistics—for example, research into collocations or keywords—where G^2 is used much more often than the chi-squared test. Second, it is with G^2 that we can powerfully extend our above testing of two-dimensional tables to higher dimensional tables. For example, we saw above that, in Peters's (2001) data, there was a significant correlation between DO's givenness and constructional choices, but what if we now had a second study testing the same issue which had these results:

```
Tab2 <- matrix(c(143, 66, 53, 141), ncol=2)¶
attr(Tab2, "dimnames") <- attr(Tab1, "dimnames")¶
Tab2¶
          Construction
Givenness V-DO-Part V-Part-DO
    given       143        53
    new          66       141
```

That is, we now have a three-dimensional design: 2 (givenness) × 2 (construction) × 2 (study) and need to determine whether the correlations of givenness and constructional choice we see in `Tab1` and `Tab2` are significantly different. Such questions can be addressed with regression methods (e.g., binary logistic regression and Poisson regression), which use G^2, not chi-squared, as their main significance-testing statistic. Such an analysis would show that the results of `Tab1` and `Tab2` differ significantly from each other ($G^2 = 13.234$, $df = 1$, $p < 0.001$; see Gries 2021, Ch. 5). Thus, recognizing G^2 as an important alternative to chi-squared offers powerful new ways to test for not just simple but also more complex (in)dependent relationships.

SEE ALSO: Comparing Groups With Multiple Independent Variables; Comparing Two Related Samples; Corpus Linguistics: Quantitative Methods; Inference; Probability and Hypothesis Testing; Quantitative Methods; Multiple Regression

References

Gries, St. Th. (2021). *Statistics for linguistics with R: A practical introduction* (3rd rev. & ext. ed.). Mouton de Gruyter.

Peters, J. (2001). Given vs. new information influencing constituent ordering in the VPC. In R. Brend, A. K. Melby, & A. Lommel (Eds.), *LACUS Forum XXVII: Speaking and comprehending* (pp. 133–140). LACUS.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Suggested Readings

Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures* (5th ed.). Chapman & Hall/CRC.

Zar, J. H. (2010). *Biostatistical analysis* (5th ed.). Prentice Hall.