Testing Independent Relationships

STEFAN TH. GRIES

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

- the rows list the levels of a nominal/categorical variable,
- the columns list the levels of another nominal/categorical variable, and
- 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 set.

An example of data from a study of disfluencies in speech is shown in Table 1, which shows the parts of speech of 335 words following three types of disfluencies. Both the part of speech and the disfluency markers represent categorical variables.

Table 1 shows that 30 *uh*'s were followed by a noun, 20 *uhm*'s were followed by a verb, etc. One question a researcher may be interested in exploring is whether there is a correlation between the kind of disfluency produced—the variable in the rows—and the part of speech of the word following the disfluency—the variable in the columns. An exploratory glance at the data suggests that *uh* mostly precedes conjunctions while silences most precede nouns, but an actual statistical test is required to determine (a) whether the distribution of the parts of speech after the disfluencies is in fact significantly different from chance, and (b) what preferences and dispreferences this data set reflects. The most frequent statistical test used to analyze two-dimensional frequency tables such as Table 1 is the chi-square test for independence.

The Chi-Square Test for Independence

The chi-square test for independence is introduced by describing how data analysis is conducted using the open source statistical language and environment R (R Development Core Team, 2010), which can be freely downloaded from http://cran.at.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 the above data in Table 1, the

Table 1 Fictitious data on the correlation of Disfluency and Part of speech 2

	Noun	*Verb*	*Conjunction*	*Totals*
uh	30	70	90	190
uhm	50	20	40	110
silence	20	5	10	35
Totals	100	95	140	335

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

DOI: 10.1002/9781405198431.wbeal1202

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"):

```
x <- matrix(c(30, 50, 20, 70, 20, 5, 90, 40, 10), ncol=3)¶
```

This creates the matrix of the frequencies shown in Table 1 and stores it as an object called `x`. However, since this matrix does not yet have row and column names, it is useful to add such names using the function `list`. This function takes two vectors, first the row names, second the column names:

```
attr(x, "dimnames") <- list(Disfluency=c("uh", "uhm", "silence"),
POS=c("Noun", "Verb", "Conjunction"))¶
```

If one now tells R to display the object `x`, then the data are shown nearly exactly as represented in Table 1:

```
x¶
          POS
Disfluency  Noun  Verb  Conjunction
  uh          30    70           90
  uhm         50    20           40
  silence     20     5           10
```

If one needs the row and column totals of `x`, too, then these can be obtained from the function `addmargins`:

```
addmargins(x)¶
          POS
Disfluency  Noun  Verb  Conjunction  Sum
  uh          30    70           90  190
  uhm         50    20           40  110
  silence     20     5           10   35
  Sum        100    95          140  335
```

Assumptions

The second step involves determining whether the data are such that one can in fact compute a chi-square test for independence. This test has three assumptions of two different kinds: the first two have to do with the frequencies that would be expected if the data were randomly distributed; the third has to do with whether the data points are independent of each other or not. The three assumptions are the following:

- 80% of the expected frequencies are greater than 4;
- all expected frequencies are greater than 1; and
- all observations are independent of each other.

In R the first two assumptions are best tested with the function of the chi-square test itself, but the third assumption requires the researcher to consider whether data points—individual occurrences of a disfluency with the part of speech of the following word—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 a case, an individual speaker's preference for a particular disfluency, or a particular disfluency before some part of speech, could bias the statistical evaluation of the data. Another threat to independence could arise if disfluencies were from different speakers but from successive turns in the same conversation, because then a disfluency 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 disfluency is from a different speaker in a different conversation.

Testing the assumptions of the chi-square 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, then the test becomes anti-conservative and may return a significant result although the null hypothesis is correct, 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 Significance Testing

If the assumptions of the chi-square test are met, it can be computed very easily in R. The function for the chi-square test is called `chisq.test`, and it takes two arguments: the table for which one wants to compute a chi-square test (here: `x`), and 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 is sometimes recommended for sample sizes between 20 and 60 and which will therefore not be needed for this example. The researcher assigns the result of the chi-square test to an object `x.test`:

```
x.test <- chisq.test(x, correct=FALSE)¶
```

Nothing is returned, but the data structure `x.test` now contains all the results. Three things must now be done. First, the researcher should inspect the frequencies that would have been expected by chance—that is, when there is no correlation between the kind of disfluency and the part of speech of the following word—by calling the part of the test results that contains the expected frequencies:

```
x.test$exp¶
          POS
Disfluency      Noun       Verb  Conjunction
  uh        56.71642  53.880597     79.40299
  uhm       32.83582  31.194030     45.97015
  silence   10.44776   9.925373     14.62687
```

(One can also compute each expected frequency of a cell manually by dividing the product of the cell's row and column total by the total of the table, for example, 190·100 ÷ 335 = 56.71642, and so forth.) Obviously, all expected frequencies are greater than 4 so the application of the chi-square test is justified. Therefore, the second step is to determine whether the observed result from Table 1 is significant—that is, different enough from the expected result shown above—by calling the overall result:

```
x.test¶
      Pearson's Chi-squared test
data:  x
X-squared = 45.2273, df = 1, p-value = 3.566e-09
```

In this example, there is a highly significant correlation between the kind of disfluency and the part of speech that follows: p is much smaller than the critical value of $p = 0.05$ that is usually adopted in the social sciences: 3.566e-09 = $3.566 \cdot 10^{-9}$ = 0.000000003566. However, the fact that there is an overall significant result does not reveal which of the nine cells is/are most responsible for this effect, i.e. what to focus on most in the interpretation. To identify these cells, one can inspect the so-called Pearson residuals.

```
x.test$res¶
          POS
Disfluency       Noun       Verb  Conjunction
  uh        −3.547512   2.196002    1.1892280
  uhm        2.995361  −2.004245   −0.8805362
  silence    2.955245  −1.563384   −1.2097935
```

If the Pearson residual in a cell is positive, then the observed frequency in that cell is greater than the expected frequency in that cell, and if the Pearson residual in a cell is negative, then the observed frequency is less than the expected frequency. The more the Pearson residual deviates from 0, the stronger that effect. In this case, therefore, the strongest effect is the dispreference of *uh* before nouns (observed frequency: 30, expected frequency: 56.7), followed by the preferences of *uhm* and silences before nouns (observed frequencies 50 and 20, expected frequencies 32.8 and 10.4 respectively). By contrast, the Pearson residuals for conjunctions are closer to zero, since their observed frequencies are closer to the expected ones.

Graphical Interpretation and Effect Size

The above kind of interpretation of chi-square tests can often be facilitated considerably with graphical displays. Figure 1 shows two frequent plots, which can be created with the following two lines:

```
mosaicplot(t(x))¶
assocplot(t(x))¶
```

The left panel shows a so-called mosaic plot, in which the box sizes are proportional to the cell frequencies and where the lack of alignment of the margins between the boxes indicates correlational structure; for example, compare the small box of *uh*:Noun against the long box for *uh*:Verb. The right panel shows a so-called association plot, in which dark and light gray rectangles indicate observed frequencies greater and less than the expected frequencies respectively, and in which the area of the box is proportional to the difference in observed and expected frequencies. It is therefore easy to see that the significant effect is mostly due to the fact that *uh* is less likely before nouns and more likely before verbs and conjunctions, and that *uhm* is less likely before verbs, and so forth.

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

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

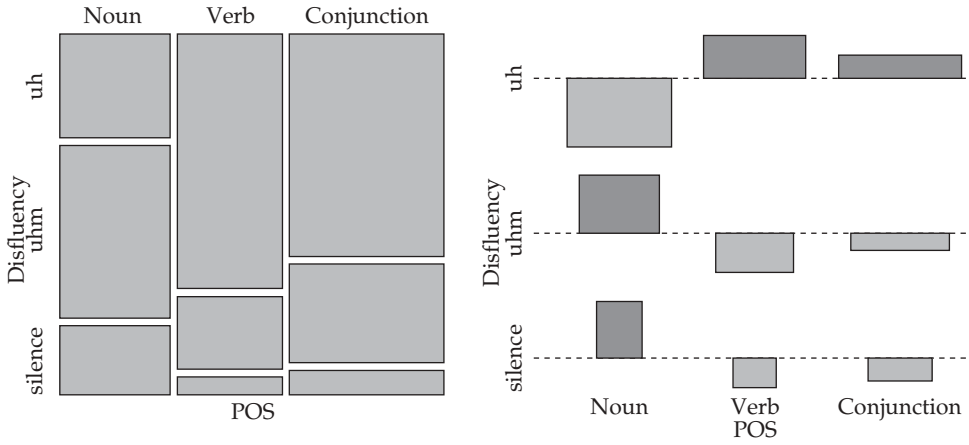

Figure 1 A mosaic plot (left panel) and an association plot (right panel) for the data in Table 1

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

```
sqrt(x.test$stat/(sum(x) * (min(dim(x))-1)))¶
X-squared
0.2598142
```

To report the result of a chi-square test, the researcher should provide the table of observed frequencies, the chi-square value as well as its *df*, its *p*-value, and the effect size. The next section discusses very briefly one way to proceed if the expected frequencies are too small to use the chi-square test, namely an exact test.

An Exact Alternative: An Exact Test for Independence

Sometimes, one may not have data that result in expected frequencies large enough to meet the conditions of the chi-square test. For instance, one may have obtained only 20% of each cell's frequency in Table 1, as in this table `y`.

```
y <- x/5¶
y¶
           POS
Disfluency  Noun  Verb  Conjunction
  uh           6    14           18
  uhm         10     4            8
  silence      4     1            2
```

A chi-square test on `y` shows that more than 20% of the expected frequencies are smaller than 3:

```
chisq.test(y, correct=FALSE)$exp
                POS
      Disfluency       Noun       Verb  Conjunction
        uh        11.343284  10.776119    15.880597
        uhm        6.567164   6.238806     9.194030
        silence    2.089552   1.985075     2.925373
chisq.test(y, correct=FALSE)
            Pearson's Chi-squared test
      data:  y
      X-squared = 9.0455, df = 4, p-value = 0.05997
```

A test that can be applied to such tables is the Fisher–Yates exact test. The application in R is very straightforward:

```
fisher.test(y)¶
        Fisher's Exact Test for Count Data
data:  y
p-value = 0.05338
alternative hypothesis: two.sided
```

The test shows that the distribution is not significant: p is too large. This is interesting for two reasons: First, as can be easily seen, the p-value of the Fisher–Yates exact test of `y` (0.05338) is closer to the standard significance level of 0.05 than the p-value of the chi-square test of `y` (0.05997): The two tests differ especially with small sample sizes. Second, this is interesting since, in percentages for example, the distribution in `y` is of course the same as in `x`: *uhm*:Verb is still four times as frequent as silence:Verb, which shows how sensitive p-values are to sample sizes and why sample-size independent effect sizes should always be provided.

This example from quantitative corpus analysis provides only one of many uses for these two tests of correspondence between categories of nominal or categorical variables. Like many other statistical tests, the chi-square has assumptions about the data that need to be checked. Data failing to meet assumptions can be analyzed in a different way, in this case, using the Fisher–Yates exact test.

SEE ALSO: Corpus Linguistics: Overview; Corpus Linguistics: Quantitative Methods; Describing and Illustrating Quantitative Data; Inference; Probability and Hypothesis Testing

Reference

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Suggested Readings

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, England: Cambridge University Press.

Gries, St. Th. (2009). *Statistics for linguistics with R: A practical introduction*. Berlin, Germany: De Gruyter.

Johnson, K. (2008). *Quantitative methods in linguistics*. Malden, MA: Wiley-Blackwell.