Frequency tables

Tests, effect sizes, and explorations

Stefan Th. Gries
University of California, Santa Barbara

This chapter provides an overview of statistical tests to analyze frequency data. Specifically, it discusses the use, logic, and interpretation of chi-squared tests of two-dimensional frequency tables as well as the computation of effect sizes for such tables, followed by several extensions and follow-up procedures that are not usually discussed (such as the analysis of sub-tables of tables and the Marascuilo procedure). In addition, there is a brief discussion of how Poisson/count regression can be used to analyze frequency data with more than two dimensions.

Keywords: chi-squared test, frequency data, Marascuilo procedure, Poisson regression

1. Introduction

1.1 Discrete vs. continuous data

Usage-based linguistics is essentially a distributional science in the sense that linguists explore the distribution of linguistic elements at every level of linguistic analysis: phonology, morphology, syntax, semantics, pragmatics and text linguistics etc. Corpus linguistics is no exception to this. More specifically, corpus linguists explore:

- the frequencies of occurrence of linguistic elements in corpora, for example, frequency lists;
- the dispersion of linguistic elements in corpora as in, for example, measures of dispersion;
- the frequencies of co-occurrence of linguistic elements in corpora as in, for example, collocation, collocational frameworks, *n*-grams, colligations/collostructions etc.

Very often, the data we study as linguists are discrete in nature. That is, the linguistic elements we study come in different categories and, trivially, if two elements are labeled the same, they belong to the same category, and if they are labeled differently, they belong to different categories. In statistical approaches, this kind of scenario is usually described with the terminology of variables (or factors) and their levels. For example, when direct objects are studied, it may be interesting to describe them in terms of which part of speech the direct object's head is. In other terminology, each direct object studied is then described with regard to the variable PART OF SPEECH by assigning a particular variable level to it; depending on what the direct objects look like, the following levels are conceivable: PART OF SPEECH: *LEXICAL NOUN*, PART OF SPEECH: *PRONOUN*, PART OF SPEECH: *SEMIPRONOUN*, (such as *matters* or *things)*, etc. Trivially, if direct objects are categorized this way, then a direct object whose head is categorized as PART OF SPEECH: *PRONOUN* is, for the purposes of this analysis, identical to another one whose head is categorized as PART OF SPEECH: *PRONOUN* and different from one whose head is categorized as PART OF SPEECH: *LEXICAL NOUN*.

On other occasions, the observed variables are actually not discrete, but continuous, but for the purposes of an analysis they may be grouped into two or more categories such as:

- when the lengths of subjects falling between 1 and 30 syllables, for example, are classified as falling into the categories LENGTH: *SHORT* (such as shorter than the median length) and LENGTH: *LONG* (i.e. longer than the median length);
- when the frequencies of closed-class words falling between 1 and 100,000, for example, are classified into the categories FREQUENCY: *LOW* (such as between 1 and 50 occurrences), FREQUENCY: *INTERMEDIATE* (such as between 51 and 1,000 occurrences), and FREQUENCY: *HIGH* (such as between 1,001 and 100,000 occurrences).

For the kinds of statistical methods to be discussed below, it does not really matter whether the variables involved in a particular study are genuinely discrete or categorical in nature or have just been converted to discrete variables: the methods as well as their results and potential implications are the same.

The analysis of multidimensional frequency tables – i.e., tables reporting observed co-occurrence frequencies of three or more features with regard to which elements have been classified – has a lot to offer to linguists in general and cognitive linguists in particular. For example, frequency effects play an important role in most flavors of Cognitive Linguistics and/or Construction Grammar:

- absolute token frequencies and conditional token probabilities are correlated with (degree of) cognitive entrenchment and unit status (cf. Schmid 2000), age and speed of the acquisition of constructions (cf. Brooks and Tomasello 1999 or

Goldberg, Casenhiser, and Sethuraman 2004), and phonological reduction (cf. Bybee and Scheibman 1999);
– type frequencies are correlated with degrees of productivity and grammaticalization (cf. Bybee 1985);
– conditional probabilities as they can be derived from corpus frequencies are correlated with processing/parsing strategies (cf. Saffran, Aislin, and Newport 1996 or Saffran and Wilson 2003); etc.

But multidimensional frequency tables of course also arise in studies whose target is not frequency effects *per se* but where just the interrelations of several variables is studied on the basis of corpus or experimental data.

The general idea in the analysis of two- or more-dimensional frequency tables is to determine whether the frequencies observed in cells of the table are distributed in a way that is significantly different from a random distribution and, if that is the case, what is (most) responsible for the significant difference and what is not. The entities that are included in an analysis because they are potentially responsible for significant differences will be called predictors, and I use *predictors* here to refer to three different things:

– levels of variables;
– individual variables;
– interactions of *n* variables.

The first two of these three different kinds of predictors are probably obvious from what has been said so far, but the third may not be. An interaction of *n* variables is defined as a non-additive, or unpredictable, joint effect of the n variables (on a dependent variable). Consider a case where the referents subject and object NPs have been coded with regard to a variable Clause (whether they are subjects or objects in a main or a subordinate clause) and their Givenness in discourse (on a scale from 0 to 10). Let us assume:

– referents of subjects are more given than referents of objects;
– referents of subjects and objects in main clauses are more given than referents of subjects and objects in subordinate clauses.

From this, one would expect the referents of subject NPs in main clauses to be most given because they combine the two features – 'given' and 'being in a main clause' that co-occur with high values of Givenness. If they turn out to be *least* given, however, then this would be a two-way interaction between the variables GramRelation (with the levels *subject* and *object*) and Clause (with the levels *main* and *subordinate*).

Before we turn to the actual analysis of frequencies of discrete data, I first need to make a few general remarks that apply to virtually all evaluations of frequency tables, in fact to most statistical methods in general.

1.2 Methodological preliminaries

The most central methodological issue that needs to be discussed briefly is how one of the most fundamental principles of scientific reasoning bears upon statistical analyses of frequency data. This most fundamental principle is *entia non multiplicanda praeter necessitatem*, which is known as Occam's razor, or sometimes also as the principle of parsimony. It prohibits the inclusion of unnecessary explanatory notions into an analysis or, from the reverse perspective, it requires the analyst to show for each explanatory notion he wants to include that it is in fact necessary to include it. For statistical analyses of frequency data, this means that a researcher (i) tries to build a model of the observed data, i.e. a quantitative representation of the potentially relevant relations in the data that contains all predictors under consideration, and then (ii) must successively determine whether the predictors currently included in the model may in fact be included in the model or whether they have to be eliminated from consideration because their influence is too small to be statistically reliable/significant or conceptually noteworthy/substantial. This means that, especially in the area of multifactorial studies, the first statistical analysis is hardly ever the last because once a first statistical model has been built, Occam's razor dictates it be tested for parsimony; in the domain of regression modeling, this process of slimming down predictors is often referred to as *model selection*.

This principle is usually recognized in multifactorial studies (to varying degrees, though), where many researchers now routinely go through a model selection process in which in a stepwise fashion predictors are excluded from consideration until a model consists only of predictors that are significant themselves or that figure in higher-order interactions that are significant. However, for both mono- and multifactorial applications, this principle is not as often recognized for predictors that are neither interactions of variables or variables but variable levels. The above definition of predictors requires that the inclusion of different variable levels should ideally be scrutinized for whether variable levels must be kept apart just as the inclusions of separate variables and interactions should be. Note, though, that a conflation of variable levels must make sense conceptually: it is not useful to create a new combination of variable levels that looks nicer statistically but is conceptually senseless (cf. below for an example) – modeling is usually only a means to an end, not an end in itself. The principle of parsimony is therefore a very important methodological guideline and will surface in different forms below.

2. How to analyze frequency tables

This section constitutes the main part of this chapter. In Section 2.1, I discuss the simpler case of two-dimensional tables, whereas in Section 2.2, I explain the more

complex case of multidimensional tables. The discussion will be based on the open source software R, which can be downloaded from <http://cran.at.r-project.org/>.

2.1 Two-dimensional tables

2.1.1 *2-by-2 tables*

The simplest case of two-dimensional tables are 2-by-2 tables, in which one nominal or categorical variable is cross-tabulated with another nominal or categorical variable. As an example, let us consider the question of whether the disfluencies *uh* and *uhm* are differently frequent directly before nouns and verbs. That is, one variable is DISFLUENCY, with the levels *UH* and *UHM*, and the other variable is PART OF SPEECH of the following word, with the levels *NOUN* vs. *VERB*.

The analysis of such tables is very straightforward. First, the data must be entered into a matrix in R. To that end, the function `matrix` can be used, which requires (i) the observed frequencies in a column-wise fashion `(c(30, 50, 70, 20))` and (ii) the number of columns the table has `(ncol=2)`:

```
x<-matrix(c(30, 50, 70, 20), ncol=2)
```

While this creates the matrix of the frequencies, it is useful to add row and column labels. The function `list` takes two vectors, first the row names, and second, the column names:

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

To see whether the data entry has been successful, the data plus the row and column totals can then be inspected using the function `addmargins`:

```
addmargins(x)
           POS
Disfluency Noun Verb Sum
      uh     30   70 100
      uhm    50   20  70
      Sum    80   90 170
```

Table 1. Fictitious data on the correlation of DISFLUENCY and PART OF SPEECH 1

	Noun	**Verb**	**Totals**
uh	30	70	100
uhm	50	20	70
Totals	80	90	170

Such matrices are typically evaluated using a so-called chi-squared test (exceptions to this will be discussed below). This test requires that all observations are independent of each other and that 80+% of the expected frequencies are larger than 5. If this is the case, one can use the function `chisq.test`, which in the standard form to be discussed here requires the matrix to be tested (`x`) and an argument to be explained below (`correct=FALSE`); the result of the test should be saved into a new data structure, e.g., `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, one should inspect the frequencies that would have been expected by chance – i.e. when there is no correlation by the kind of disfluency and the part of speech of the following word – by calling the part of the test results that contain the expected frequencies:

```
x.test$exp
           POS
Disfluency      Noun     Verb
       uh  47.05882 52.94118
       uhm 32.94118 37.05882
```

(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, e.g., 100·80÷170 = 47.05882, etc.). Obviously, the expected frequencies are all greater than 5. Therefore, the next step is to determine whether the observed result from Table 1 is significant – i.e. different enough from the expected result shown above – by calling the overall result:

```
x.test
   Pearson's Chi-squared test
data:  x
X-squared = 28.3671, df = 1, p-value = 1.004e-07
```

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 $p = 0.05$. However, the fact that there is an overall significant result does not reveal which of the four cells are most responsible for this effect and how. To identify these cells, one should inspect the so-called Pearson residuals, which are computed as in (1).

(1) $$\text{Pearson residuals} = \frac{observed - expected}{\sqrt{expected}}$$

```
x.test$res
           POS
Disfluency       Noun       Verb
       uh  -2.486729  2.344511
       uhm  2.972210 -2.802227
```

First, if the Pearson residual in a cell is positive/negative, then the observed frequency in that cell is greater/less than the expected frequency in that cell. Second, the more the Pearson residual deviates from 0, the stronger that effect. In this case, therefore, the strongest effect is the preference of *uhm* before nouns, followed by the dispreference of *uhm* before verbs.

The final step is to compute an effect size. An effect size quantifies the strength of the observed correlation independently of the sample size. In the case of 2-by-2 tables, one standard effect size is Φ (*phi*), which theoretically ranges from 0 ('no effect') to 1 ('perfect correlation') and is computed as shown in (2). In this case, the correlation is intermediately strong.

(2) $\Phi = \sqrt{\frac{\chi^2}{n}}$

```
sqrt(x.test$stat/sum(x))
X-squared
0.4084912
```

The final question to be addressed is what to do when too many expected frequencies are too small. While sparse data are always problematic in the sense that one does not want to base potentially far-reaching generalizations on small data sets, there is a test that can be used to test such tables for significance, too, which is called the Fisher-Yates exact test. The R function that computes this test is `fisher.test` and its most important argument is just the matrix containing the data:

```
fisher.test(x)
    Fisher's Exact Test for Count Data
data:  x
p-value = 9.66e-08
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.08256479 0.35287555
sample estimates:
odds ratio
 0.1734529
```

In this case, where the sample size and the expected frequencies are unproblematic anyway, the *p*-value provides the same kind of result: the distribution of the two

disfluencies before the two parts of speech is most likely not random, but there is more. The output also provides another kind of effect size for 2-by-2 tables, the so-called odds ratio. The odds ratio is one of several measures that expresses how much the distribution of a binary variable changes in response to another binary variable. In this case, the odds ratio quantifies the ratio of the frequency of *uh* before nouns (30/50) to the frequency of *uh* before verbs (70/20):

```
(30/70)/(50/20)
[1] 0.1714286
```

The result is similar, but not identical, to the one provided by R, which uses a more refined estimation algorithm (and also provides a confidence interval that is not addressed here (cf. Gries 2013: Section 3.1.5 for explanation and exemplification)). The logic, however, is the same: the more the odds ratio differs from 1, the stronger the effect. Sometimes, a scholar might not report odds ratios of 0.5 and 1.5 but logged odds ratios as shown below:

```
log(0.5)
[1] -0.6931472
log(1.5)
[1] 0.4054651
```

One reason for this is that odds ratios are often difficult to compare to each other: A beginner might look at two odds ratios of 0.5 and 1.5 and – erroneously – think they reflect equally strong effects because they are equally far away from 1. This is false as the logs of the odds ratios show: the more a logged odds ratio deviates from 0, the stronger the effect, which is why an odds ratio of 0.5 reflects a stronger effect than an odds ratio of 1.5. (In addition, logged odds ratios are also important in the context of logistic regression, a topic I cannot discuss here; cf. Gries 2013: 203–301 for detailed explanation.)

2.1.2 *Larger two-dimensional r-by-c tables*

Thankfully, the logic of 2-by-2 tables also applies to two-dimensional tables with more than two rows (i.e. $r > 2$) and/or two more columns (i.e. $c > 2$). Consider Table 2 for an extended version of the above disfluencies example.

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

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

The data are entered in the same way as before:

```
x<-matrix(c(30, 50, 20, 70, 20, 5, 90, 40, 10), ncol=3)
attr(x, "dimnames")<-list(Disfluency=c("uh", "uhm",
   "silence"), POS=c("Noun", "Verb", "Conjunction"))
```

And the requirements of the test (in terms of the proportion of expected frequencies being not smaller than 5) analysis is also no different:

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

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

x.test
   Pearson's Chi-squared test
data:  x
X-squared = 45.2273, df = 4, p-value = 3.566e-09

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
```

The expected frequencies are unproblematic: all of them are even larger than 9. Hence, the *p*-value of the chi-squared test can be taken seriously, which points to an association between the kind of disfluency and the part of speech of the following word. The nature of this association then becomes clear from the residuals: *uh* is dispreferred before nouns (negative residual of ≈–3.55) whereas *uh* is preferred before verbs (positive residual of ≈2.2) and before conjunctions (positive residual of ≈1.2).

Two things remain to be done. First, one again needs to compute an effect size, which for *r*-by-*c* tables with $r > 2$ and/or $c > 2$ is called Cramer's *V*. Its formula is shown in (3), where *min(r, c)* means 'take the numbers of rows and columns and pick the smaller of the two'.

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

The effect size is now smaller than before:

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

Second, one should explore whether the data can, or in fact must, be simplified as a consequence of Occam's razor, the principle that requires analysts to adopt the simplest possible model. In this case, for example, the observed data distinguish three parts of speech – nouns, conjunctions, and verbs – but the (signs of the) residuals reveal that verbs and conjunctions behave alike so maybe a two-way distinction – nouns vs. non-nouns – is sufficient.

To test heuristically which distinction to adopt, the data are entered again, but this time the two levels of the part of speech that are suspected to behave the same are conflated:

```
x.2<-matrix(c(30, 50, 20, 70+90, 20+40, 5+10), ncol=2)
attr(x.2, "dimnames")<-list(Disfluency=c("uh", "uhm",
   "silence"), POS=c("Noun", "Not noun"))
addmargins(x.2)
           POS
Disfluency  Noun Not noun Sum
   uh         30      160 190
   uhm        50       60 110
   silence    20       15  35
   Sum       100      235 335
```

Then, the analysis is repeated on the new merged data set:

```
x.2.test<-chisq.test(x.2, correct=FALSE)
x.2.test
   Pearson's Chi-squared test
data:  x.2
X-squared = 43.1801, df = 2, p-value = 4.203e-10
x.2.test$res
           POS
Disfluency       Noun  Not noun
   uh      -3.547512  2.314141
   uhm      2.995361 -1.953958
   silence  2.955245 -1.927790
sqrt(x.2.test$stat/(sum(x.2) * (min(dim(x.2))-1)))
X-squared
0.3590205
```

The chi-square value has hardly changed and the effect size has even gone up considerably. Both of these facts suggest that the real distinction for the disfluencies in this corpus may not be between nouns, verbs, and conjunctions, but just between nouns and non-nouns, but this would have to be tested more rigorously (using, e.g., model comparisons).

2.1.3 *Additional applications*

This section deals with two unfortunately less well-known, but nevertheless very useful methods in the analysis of different kinds of two-dimensional *r*-by-*c* columns.

Testing a subtable of a table

This section is concerned with the question of what to do when one has an *r*-by-*c* subtable, but wishes to evaluate an *s*-by-*d* table with $s \leq r$ and $d \leq c$. As an example, I will use a study of the frequencies with which four emotion metaphors are distributed over four registers (cf. Lohmann 2009 for an example). Lohmann studied the degree to which certain supposedly very pervasive conceptual metaphors are attested in different genres. Consider Table 3 for a (fictitious) example of the kind of data such a study may yield (I will explain the bold-faced figures below.)

If one entered the data …

```
x<-matrix(c(8, 31, 44, 36, 5, 14, 25, 38, 4, 22, 17, 12,
   8, 11, 16, 24), ncol=4)
attr(x, "dimnames")<-list(Register=c("acad", "spoken",
   "fiction", "news"), Metaphor=c("Heated fluid", "Light",
   "NatForce", "Other"))
```

and did a chi-squared test on this table, then one would find a significant result (with $\chi^2 = 19.5151$; $df = 9$; $p = 0.02115$). However, let us assume that one found these data in a study, but that one is also only interested in whether spoken conversation differed from fiction in the use of the metaphors EMOTION IS LIGHT and EMOTION IS A NATURAL FORCE, i.e., the bold-faced figures in Table 3. Contrary to what quite a few

Table 3. Fictitious distribution of emotion metaphors in different genres

Metaphor Register	EMOTION IS A HEATED FLUID IN A CONTAINER	EMOTION IS LIGHT	EMOTION IS A NATURAL FORCE	Other	Totals
Academic writing	8	5	4	8	25
Spoken conv.	31	**14**	**22**	11	78
Fiction	44	**25**	**17**	16	102
News	36	38	12	24	110
Totals	119	82	55	59	315

people seem to think, one cannot simply extract this table from the overall table – i.e., pretend one had done a study oneself with just the variable levels and frequencies one is interested in – and run a chi-squared test on it. Thus, the following (slightly shortened) code and result is *wrong*:

```
subtable<-matrix(c(14, 25, 22, 17), ncol=2)
chisq.test(subtable, correct=FALSE) # WRONG!
        Pearson's Chi-squared test
data:  matrix(c(14, 22, 25, 17), ncol = 2)
X-squared = 3.3016, df = 1, p-value = 0.06921
```

This test is wrong because its chi-square value is based on the marginal totals of the subtable (e.g. 39 vs. 39 for EMOTION IS LIGHT, etc.), but does not take the overall observed frequencies of EMOTION IS LIGHT into consideration (e.g. 82 vs. 55, etc.). The correct test is, unfortunately, slightly more lengthy and involves the following steps (following Bortz, Lienert, and Boehnke 1990: Section 5.4.4).

First, one computes the chi-squared test that compares the observed row sums of the subtable (36 vs. 42) to the ones expected from the proportions of row sums of the whole table (78 vs. 102, i.e., $78/_{180}$ vs. $102/_{180}$):

```
chisq.test(c(36, 42), p=c(78, 102)/180)[c(1,7)]
$statistic
X-squared
0.2526975

$expected
[1] 33.8 44.2
```

Second, one computes the chi-squared test that compares the observed column sums of the subtable (39 vs. 39) to the ones expected from the proportions of column sums of the whole table (82 vs. 55, i.e. $82/_{137}$ vs. $55/_{137}$):

```
chisq.test(c(39, 39), p=c(82, 55)/137)[c(1,7)]
$statistic
X-squared
 3.151996

$expected
[1] 46.68613 31.31387
```

Third, one computes the frequencies that would have been expected in the subtable if the cells were distributed proportional to the expected marginal totals according to the usual two-dimensional chi-square formula mentioned above, by dividing the product of the cell's row and column total by the total of the table, as shown in Table 4.

Table 4. Expected frequencies (when the cells are proportional to the expected marginal totals)

	EMOTION IS LIGHT	EMOTION IS A NATURAL FORCE	Totals
Spoken convers.	(33.8 × 46.69) / 78 ≈ 20.23	(33.8 × 31.31) / 78 ≈ 13.57	33.8
Fiction	(44.2 × 46.69) / 78 ≈ 26.46	(44.2 × 31.31) / 78 ≈ 17.74	44.2
Totals	46.69	31.31	78

Table 5. Contributions to chi-square

	EMOTION IS LIGHT	EMOTION IS A NATURAL FORCE	Totals
Spoken convers.	$(14 - 20.23)^2 / 20.23 \approx 1.92$	$(22 - 13.57)^2 / 13.57 \approx 5.24$	33.8
Fiction	$(25 - 26.46)^2 / 26.46 \approx 0.08$	$(17 - 17.74)^2 / 17.74 \approx 0.03$	44.2
Totals	46.69	31.31	78

As the penultimate step, one computes each table cell's contribution to the chi-square value by dividing the squared difference between the observed and the expected cell frequency by the expected frequency, as shown in Table 5. (By the way, these correspond to the squared Pearson residuals mentioned in (1).)

In R, this can be done much more simply:

```
exp.temp<-matrix(c(20.23, 26.46, 13.57, 17.74), ncol=2)
sum(((subtable-exp.temp)^2)/exp.temp)
[1] 7.266921
```

The final step is then, at last, to compute the difference of this last chi-square value and the sum of the other two. This difference is the required chi-square value and then provides the desired *p*-value:

```
7.266921-(0.2526975+3.151996) # chi-square
[1] 3.862227
pchisq(3.862227, prod(dim(subtable)-1), lower.tail=F) #
   p-value
[1] 0.04938474
```

This chi-square value corresponds – disregarding rounding errors – to what an R function for this method written by the author would provide, as the last row indicates:

```
sub.table(x, 2:3, 2:3) # the data and the rows/columns
   for the sub-table
[…]
```

```
$'Chi-squared tests'
                                  Chi-square Df    p-value
Cells of subtable to whole table 7.2682190  3 0.06382273
Rows (within sub-table)           0.2526975  1 0.61518204
Columns (within sub-table)        3.1519956  1 0.07583417
Contingency (within sub-table)    3.8635259  1 0.04934652
```

As is now obvious, the data in the subtable actually produce a significant result: the two kinds of metaphors are differently frequent in the two registers. Note again that the wrong approach from above – just applying a separate chi-squared test to the subtable – did not return a significant result, which should demonstrate how important it is to apply the correct methods.

The Marascuilo procedure

In order to determine how many different variable levels to retain in either a unidimensional vector of frequencies or percentages, one can use the so-called Marascuilo procedure. Since this procedure can be applied to a simple vector of frequencies or percentages, it can also be used for a 2-by-*c* table, where it tests which of the *c* variable levels of the column variable are better conflated. To explore this procedure, we consider the alternation of particle placement exemplified in (4).

(4) a. He picked up the book.
    b. He picked the book up.

Just like many other constituent order alternations in English, the choice of one order by a speaker is determined by many different factors and usually made unconsciously. One of the factors governing particle placement is the information status of the referent of the direct object (cf. Kruisinga and Erades 1953; Chen 1982; Gries 2003): on the whole, it seems as if new referents prefer to occur after the particle (i.e. as in (4a)) whereas given referents prefer to occur before the particle (i.e. as in (4b)). However, since given vs. new is only the most simplistic classification of information status, one may want to include at least one additional level such as Status: *inferable*, which characterizes referents which have not been mentioned before in the preceding discourse, but which a hearer can infer on the fly from linguistic or contextual knowledge. Consider Table 6 for an example data set.

The first step in applying the Marascuilo procedure is as discussed above, i.e. enter the data and perform a chi-squared test for two-dimensional tables to determine whether there is a correlation between information status and the constituent order:

```
x<-matrix(c(37, 13, 63, 37, 20, 40), ncol=3)
attr(x, "dimnames")<-list("Constituent order"=
   c("Verb-Object-Particle", "Verb-Particle-Object"),
   "Information status"=c("given", "inferable", "new"))
x.test<-chisq.test(x, correct=FALSE)
```

Table 6. Particle placement: Constituent order and Information status

	given	inferable	new	Totals
Verb-Object-Particle	37	63	20	120
Verb-Particle-Object	13	37	40	90
Totals	50	100	60	210

```
x.test$exp
                        Information status
Constituent order          given inferable      new
  Verb-Object-Particle 28.57143  57.14286 34.28571
  Verb-Particle-Object 21.42857  42.85714 25.71429

x.test
        Pearson's Chi-squared test
data:  x
X-squared = 21.0914, df = 2, p-value = 2.631e-05

x.test$res
                        Information status
Constituent order           given  inferable       new
  Verb-Object-Particle  1.576841  0.7748272 -2.439750
  Verb-Particle-Object -1.820780 -0.8946933  2.817181

sqrt(x.test$stat/(sum(x) * (min(dim(x))-1)))
X-squared
0.3169151
```

The result is fairly obvious: there is a not particularly strong, but still highly significant, correlation or interaction between the two variables in the expected direction: given and new referents prefer to occur before and after the particle, respectively. In addition, inferable referents pattern more like given referents – they prefer to occur before the particle – but less strongly so. According to Occam's razor, one should now test whether all three levels of Information status are required especially since in this case a conflation of Information status: *given* and Information status: *inferable* as a counterpart to Information status: *new* would make sense – whereas a conflation of Information status: *given* and Information status: *new* as a counterpart to Information status: *inferable* would not.

The Marascuilo procedure requires three steps. First, one computes the percentages of the variable with two levels, in this case the column variable Constituent order:

```
prop<-prop.table(x, 2) # the 2 means 'column-wise',
   1 would mean 'row-wise'
```

```
prop
                          Information status
Constituent order       given inferable       new
  Verb-Object-Particle   0.74      0.63 0.3333333
  Verb-Particle-Object   0.26      0.37 0.6666667
```

Second, one computes all pairwise differences between the percentages of one constituent order in the three information states:

- 0.74 – 0.63 = 0.11 (*GIVEN – INFERABLE*);
- 0.74 – 0.333 = 0.407 (*GIVEN – NEW*); and
- 0.63 – 0.333 = 0.297 (*INFERABLE – NEW*).

Third, one compares each of the three differences to a threshold value that must be computed with the rather complicated formula shown in (5) (for a significance value of $p = 0.05$):

(5) $$\sqrt{\chi^2_{p = 0.05;\ df = \text{levels} - 1}} \times \sqrt{\frac{perc_1 \times (1 - perc_1)}{\Sigma\, column_1} + \frac{perc_2 \times (1 - perc_2)}{\Sigma\, column_2}}$$

For the comparison (*GIVEN – INFERABLE*), this translates into (6):

(6) $$\sqrt{5.9915} \times \sqrt{\frac{0.74 \times 0.26}{50} + \frac{0.63 \times 0.37}{100}} = 0.1924091$$

Since the observed percentage difference of 0.11 is not larger than the critical percentage difference for $p = 0.05$ at $df = 3 - 1 = 2$ of approximately 0.19, the difference between the percentages of INFORMATION STATUS: *GIVEN* and INFORMATION STATUS: *INFERABLE* is not significant. Again, this procedure is somewhat labor-intensive, but can be computed easily using R. The output of applying such a function (`mar`) to the matrix `x` returns, among other things, the following results for the pairwise comparisons:

```
mar(x)
[…]
$'pairwise comparisons'
      comparisons     diffs crit.ranges decisions
1 given-inferable 0.1100000   0.1924091        ns
2       given-new 0.4066667   0.2127105         *
3   inferable-new 0.2966667   0.1901492         *
```

The results of the Marascuilo procedure at least suggest that one should conflate INFORMATION STATUS: *GIVEN* and INFORMATION STATUS: *INFERABLE* into a new category INFORMATION STATUS: *NON-NEW*, evaluate that matrix, and report and interpret those results:

```
x2<-matrix(c(100, 50, 20, 40), ncol=2)
attr(x2, "dimnames")<-list("Constituent order"=
   c("Verb-Object-Particle", "Verb-Particle-Object"),
   "Information status"=c("non-new", "new"))
x2.test<-chisq.test(x2, correct=FALSE)

x2.test$exp
                      Information status
Constituent order       non-new      new
  Verb-Object-Particle 85.71429 34.28571
  Verb-Particle-Object 64.28571 25.71429

x2.test
        Pearson's Chi-squared test
data:  x2
X-squared = 19.4444, df = 1, p-value = 1.036e-05

x2.test$res
                         Information status
Constituent order         non-new       new
  Verb-Object-Particle  1.543033 -2.439750
  Verb-Particle-Object -1.781742  2.817181

sqrt(x2.test$stat/(sum(x2) * (min(dim(x2))-1)))
X-squared
0.3042903
```

Both the chi-square value and the effect size hardly change as a result of the elimination of one variable level, and given this loss of one *df*, the *p*-value is even much smaller than before. (Note that other statistical approaches may come to different conclusions, which does not, however, obviate the need for some kind of test of whether the three levels of INFORMATION STATUS need to be kept separate and for an explicit discussion of which test was used.)

This is a clear case in which Occam's razor not only makes the results better, but in which it also may lead to new findings: if the researcher had not already expected that *GIVEN* and *INFERABLE* were very similar – unless the researcher wanted to test whether they are the same, he or she should have just coded one non-new information status – then Occam's razor has helped to reveal this patterning.

2.2 Multidimensional tables

Two-dimensional frequency tables have probably the most widespread use of all frequency tables. However, most linguistic choices are not determined by only a single variable, and while the analysis of multidimensional frequency tables is somewhat

more complex, a growing number of linguists have realized that very often only a multifactorial study will reveal the most important generalizations and avoid erroneous interpretations arising from the omission of important predictor variables (cf. the well-known example of Simpson's paradox; cf. Sheskin 2011: 718–720).

Multidimensional frequency tables can be analyzed in many different ways, which often makes it difficult for the beginner to choose one method over another: loglinear models/Poisson regression, binary or multinomial logistic regression, (multiple) correspondence analysis, association rules, … are among the most frequent methods but I cannot discuss them all here. Binary logistic regression is a very widely used method but, as the name suggests, it is restricted to dependent variables with only two levels (cf. Gries 2013: Section 5.3 for in-depth discussion as well as Baayen 2008: Section 6.3.1; Johnson 2008: Section 5.4; and Speelman, this volume). I will therefore discuss one example of a Poisson regression (which could also be investigated with a binary logistic regression). Let me begin, however, with the warning that this chapter cannot discuss all the tricky details of regression model selection so readers are advised to brush up their knowledge in this area and/or study additional materials (especially those readers who do not know linear regressions already); I find Crawley's (2005, 2012) books most instructive, and Faraway (2006) also provides a good, though more technical, introduction.

In this section, I will discuss an example from a recent corpus study published in the *ICAME Journal* (Hommerberg and Tottie 2007). Their study explores two complementation patterns of the verb *try* in British and American English: *try to* vs. *try and*. Their goal is "to show how native speakers of present-day British and American English actually use the two constructions", and they use a data set from the Cobuild Direct Corpus, whose size and composition is summarized in Table 7.

The variables VARIETY and MODE are self-explanatory; the variable TRY refers to whether speakers/writers used *try to* or *try and*, and the variable CLAUSE refers to whether the VP containing try is itself part of a *to*-clause (as in *we're going to try* (*to*/*and*) (Hommerberg and Tottie 2007: 56).

I will assume that Table 7 is in R's workspace as a data frame called `x`. The `str` command summarizes the structure of the table as follows:

```
str(x)
'data.frame': 16 obs. of  5 variables:
$ VARIETY: Factor w/ 2 levels "american","british":
   1 1 1 1 ...
$ MODE   : Factor w/ 2 levels "spoken","written":
   1 1 2 1 1 ...
$ TRY    : Factor w/ 2 levels "and","to":
   1 1 2 2 1 1 2 2 1 ...
```

Table 7. The data studied by Hommerberg and Tottie (2007)

Variety	Mode	Try	Clause	Freq.
american	spoken	and	other	120
american	spoken	and	to	90
american	spoken	to	other	381
american	spoken	to	to	174
american	written	and	other	10
american	written	and	to	26
american	written	to	other	219
american	written	to	to	167
british	spoken	and	other	503
british	spoken	and	to	706
british	spoken	to	other	150
british	spoken	to	to	133
british	written	and	other	49
british	written	and	to	127
british	written	to	other	230
british	written	to	to	144

```
$ CLAUSE : Factor w/ 2 levels "other","to":
   1 2 1 2 1 2 1 2 ...
$ FREQ   : int  120 90 381 174 10 26 219 167 503 706 ...
```

Several things are needed for a Poisson regression. First, the relevant R function is `glm`, which is short for generalized linear model. Second, the function takes two main arguments, the first of which is a formula that specifies which dependent variable – the observed frequencies of occurrence – and which predictors – which independent variables and which of their interactions – to include. Formulae in R are written as "dependent variable ~ predictors/independent variables", where *n* independent variables combined with asterisks mean 'include the independent variables and their interactions' while *n* independent variables combined with colons mean 'include the interaction of these independent variables'. The second argument is `family=poisson`, which instructs R to compute a Poisson regression with a log-link and not a 'normal' linear regression with a Gaussian identity function. In essence, this ensures that the regression cannot predict negative values (which would not make sense since frequencies cannot be negative; cf. Crawley 2012: Section 13.3 for discussion).

Given the discussion of model selection, the first step of the actual analysis consists of fitting a maximal model in which all predictors are included. The following code computes such a model, stores it into a data structure `m1`, and summarizes this data structure (the output here is abbreviated and minimally altered).

```
m1<-glm(FREQ ~ VARIETY*MODE*TRY*CLAUSE, family=poisson)
summary(m1)
[…]
Coefficients:
                 Estimate Std.Error z value Pr(>|z|)
(Intercept)       4.78749  0.09129  52.444  < 2e-16 ***
VARIETYbrit       1.43310  0.10159  14.106  < 2e-16 ***
MODEwrt          -2.48491  0.32914  -7.550 4.36e-14 ***
TRYto             1.15531  0.10468  11.037  < 2e-16 ***
CLAUSEto         -0.28768  0.13944  -2.063  0.03911 *
VARIETYbrit:MODEwrt
                  0.15614  0.36157   0.432  0.66586
VARIETYbrit:TRYto
                 -2.36526  0.14005 -16.889  < 2e-16 ***
MODEwrt:TRYto     1.93118  0.33989   5.682 1.33e-08 ***
VARIETYbrit:CLAUSEto
                  0.62671  0.15116   4.146 3.38e-05 ***
MODEwrt:CLAUSEto  1.24319  0.39737   3.129  0.00176 **
TRYto:CLAUSEto   -0.49606  0.16678  -2.974  0.00294 **
VARIETYbrit:MODEwrt:TRYto
                  0.82503  0.38592   2.138  0.03253 *
VARIETYbrit:MODEwrt:CLAUSEto
                 -0.62985  0.43542  -1.447  0.14803
VARIETYbrit:TRYto:CLAUSEto
                  0.03675  0.21309   0.172  0.86307
MODEwrt:TRYto:CLAUSEto
                 -0.73053  0.42051  -1.737  0.08235 .
VARIETYbrit:MODEwrt:TRYto:CLAUSEto
                 -0.23079  0.48373  -0.477  0.63328
[…]
Null deviance:      2.1620e+03 on 15 degrees of freedom
Residual deviance: -8.7486e-14 on  0 degrees of freedom
[…]
```

The main part of the output above is a table, which lists the included predictors, their coefficient estimates and their significance tests. The most relevant columns are the first (with the name of the predictor), the second headed *Estimate*, and the last with the *p*-value for the predictor. The row for the intercept shows 4.78749 as an estimate, the antilog of which is 120, the observed frequency of the combination of the alphabetically first factor levels: VARIETY: *AMERICAN* MODE: *SPOKEN* TRY: *AND* CLAUSE: *OTHER*.

The estimates that R outputs then for the predictors reflect the difference between the listed predictor and a reference level. For individual variables, the reference level is the alphabetically first, unlisted level. For instance, the value of 1.43310 for VARIETY: *BRITISH* means that the model estimates that, compared to the reference level of VARIETY: *AMERICAN*, VARIETY: *BRITISH* increases (positive sign) the estimated frequencies. For instance, the value of –2.48491 for MODE: *WRITTEN* means that the model estimates that, compared to the reference level of MODE: *SPOKEN*, MODE: *WRITTEN* reduces (negative sign) the estimated frequencies (and more strongly so than VARIETY: *BRITISH* increases them).

Another way to understand the meanings of the coefficients is to compute the predictions of the model, for example, for VARIETY: *BRITISH* MODE: *WRITTEN* TRY: *TO* CLAUSE: *OTHER*, all one needs to do is to add up all coefficients whose predictors are part of this configuration and antilog the sum:

```
exp(4.78749+1.43310-2.48491+1.15531+0.15614-2.36526+
   1.93118+0.82503)
[1] 230.0002 # rounding difference only
```

This predicted frequency corresponds to the observed frequency because this is the maximal model that contains all predictors. According to Occam's razor, insignificant predictors must now be weeded out. Crucially, the elimination of insignificant predictors always begins with the highest-order interactions and, as long as there are still insignificant predictors, proceeds to lower-order interactions and then to individual variables. Crucially, a predictor is *not* removed even if it is significant when it still participates in a higher-order interaction. In this case, there is only one four-way interaction – VARIETY: MODE: TRY: CLAUSE – and it is not significant. Thus, one now updates the first model by removing that interaction:

```
m2<-update(m1, ~. -VARIETY:MODE:TRY:CLAUSE)
```

However, one must now first check whether this simplification of the model was justified. This is how it is done:

```
anova(m1, m2, test="LRT")
[…]
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         0    0.00000
2         1    0.22492  -1 -0.22492  0.6353
```

R evaluates the difference between the two models, and because they do not differ from each significantly ($p = 0.6353$), Occam's razor requires one to adopt the simpler one, `m2`. One can then inspect this simpler model (with `summary(m2)`) and it becomes obvious that the only three-way interaction that is not significant is VARIETY: TRY: CLAUSE. Hence:

```
m3<-update(m2, ~. -VARIETY:TRY:CLAUSE)
anova(m2, m3, test="LRT")
[…]
  Resid. Df Resid. Dev Df   Deviance Pr(>Chi)
1         1    0.22492
2         2    0.22633 -1 -0.0014192  0.9699
```

When this new model `m3` is inspected, it is clear that it cannot be simplified any further: each predictor is significant or participates in a significant interaction (VARIETY: MODE is not significant, but participates in VARIETY: MODE: TRY, which is).

```
summary(m3)
[…]
Coefficients:
                Estimate Std. Error z value  Pr(>|z|)
(Intercept)      4.79421    0.08228  58.268  < 2e-16 ***
VARIETYbrit      1.42477    0.08915  15.981  < 2e-16 ***
MODEwrt         -2.59755    0.23561 -11.025  < 2e-16 ***
TRYto            1.14646    0.09105  12.592  < 2e-16 ***
CLAUSEto        -0.30343    0.10550  -2.876  0.00403 **
VARIETYbrit:MODEwrt
                 0.29069    0.22849   1.272  0.20330
VARIETYbrit:TRYto
                -2.34943    0.10553 -22.263  < 2e-16 ***
MODEwrt:TRYto    2.05052    0.23742   8.637  < 2e-16 ***
VARIETYbrit:CLAUSEto
                 0.64521    0.10658   6.054 1.42e-09 ***
MODEwrt:CLAUSEto
                 1.40279    0.21997   6.377 1.80e-10 ***
TRYto:CLAUSEto -0.47355    0.10385  -4.560 5.11e-06 ***
VARIETYbrit:MODEwrt:TRYto
                 0.67402    0.22825   2.953  0.00315 **
VARIETYbrit:MODEwrt:CLAUSEto
                -0.82045    0.17546  -4.676 2.92e-06 ***
MODEwrt:TRYto:CLAUSEto
                -0.90750    0.20548  -4.417 1.00e-05 ***
[…]
Null deviance:  2161.98713 on 15 degrees of freedom
Residual deviance: 0.22633 on  2 degrees of freedom
[…]
```

Often such data are summarized in the form of a table that, simplifying a bit, conveniently summarizes each independent variable's significance in one *p*-value. This table, a so-called ANOVA table, can be created as follows (cf. Gries 2013: 266, 271, for explanation).

```
library(car)
options(contrasts=c("contr.sum", "contr.poly"))
   Anova(m3, type="III", test="LR") # the results are
     not shown here
options(contrasts=c("contr.treatment", "contr.poly"))
```

The table now combines predictors involving the same variables but different variable levels and confirms more succinctly what was already shown above: each predictor in `m3` but VARIETY: MODE is significant.

How are the results interpreted? They are interpreted as already hinted at above, on the basis of the coefficients. Since I cannot discuss all the findings in detail, some comments must suffice. The data show, trivially, compared to the reference combination of VARIETY: *AMERICAN* MODE: *SPOKEN* TRY: and CLAUSE*: OTHER*, setting VARIETY to *BRITISH* increases the predicted counts (the coefficient for VARIETY: *BRITISH* is positive), compared to the reference combination of VARIETY: *AMERICAN* MODE: *SPOKEN* TRY: *AND* CLAUSE*: OTHER*, setting MODE to *WRITTEN* decreases the predicted counts (the coefficient for MODE: *WRITTEN* is negative), etc.

More interesting, however, are the interactions that qualify these main effects. As just one example, consider the interaction VARIETY: *BRITISH* TRY: *TO*. This strong and highly significant interaction means that, while setting both VARIETY to *BRITISH* and TRY to *TO* increases the estimated counts (by the antilog of 1.42477 + 1.14646 ≈ 2.57123), their joint effect does not boost the counts accordingly, but *de*creases them by nearly the same amount (by the antilog of –2.34943); thus, compared to the predicted frequency for VARIETY: *AMERICAN* MODE: *SPOKEN* TRY: *AND* CLAUSE*: OTHER*, 120.81, the predicted frequency for VARIETY: *BRITISH* MODE: *SPOKEN* TRY: *TO* CLAUSE*: OTHER*, is increased by 24.8%, the antilog of 2.57123 – 2.34943. But then there is also a significant interaction VARIETY: *BRITISH* MODE: *WRITTEN* TRY: *TO* … It is clear that complex interactions like these and the degree to which they are significant or not can hardly be recognized by just eye-balling the data and are usually only comprehensible on the basis of well-designed graphs (e.g. bar plots of predicted frequencies).

This concludes the discussion of this example here. Many issues of Poisson regressions could not be covered for reasons of space (such as testing the assumptions of Poisson regressions or how to handle over-/underdispersion). The method is powerful when applied to complex data sets and definitely worth exploring in more detail.

3. Conclusion

This brief chapter could, of course, not do justice to all the complexities that can and do arise in the study of frequency tables. I hope, however, that the above brief remarks and examples have shown how useful a statistically correct and comprehensive exploration of such data can be and also has hopefully whetted the reader's appetite to explore such techniques in more detail (and also their graphical exploration, which I could not address here at all) both in this volume and in the works referred to throughout this paper.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511801686

Bortz, J., Lienert, G. A., & Boehnke, K. (1990). *Verteilungsfreie Methoden in der Biostatistik* (2nd ed.). Heidelberg: Springer. DOI: 10.1007/978-3-662-22593-6

Brooks, P., & Tomasello, M. (1999). How children constrain their argument structure constructions. *Language*, 75, 720–738. DOI: 10.2307/417731

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam & Philadelphia: John Benjamins. DOI: 10.1075/tsl.9

Bybee, J. L., & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics*, 37, 575–596. DOI: 10.1515/ling.37.4.575

Chen, P. (1982). Discourse and particle movement in English. *Studies in Language*, 10, 79–95. DOI: 10.1075/sl.10.1.05che

Crawley, M. (2005). *Statistics: An introduction using R*. New York: John Wiley. DOI: 10.1002/ 9781119941750

Crawley, M. (2012). *The R book* (2nd ed.). Chichester: John Wiley. DOI: 10.1002/9781118448908

Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Boca Raton, FL: Chapman and Hall.

Goldberg, A. E., Casenhiser, D., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, 14, 289–316.

Gries, St. Th. (2003). *Multifactorial analysis in corpus linguistics: A study of particle placement*. London & New York: Continuum.

Gries, St. Th. (2013). *Statistics for linguistics with R: A practical introduction*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110307474

Hommerberg, C., & Tottie, G. (2007). Try to and try and? Verb complementation in British and American English. *ICAME Journal*, 31, 45–64.

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

Kruisinga, E., & Erades, P. A. (1953). *An English grammar*. Vol. I. Groningen: P. Noordhoff.

Lohmann, A. (2009). The register-specificity of metaphor. Paper presented at the workshop *'Corpus, colligation, register variation' of the 31st DGfS-Tagung*.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928. DOI: 10.1126/science.274.5294.1926

Saffran, J. R., & Wilson, D. P. (2003). From syllabus to syntax: Multilevel statistical learning by 12-month old infants. *Infancy*, 4, 273–284. DOI: 10.1207/S15327078IN0402_07

Sheskin, D. J. (2011). *Handbook of parametric and non-parametric statistical procedures* (5th ed.). Boca Raton, FL, London & New York: CRC Press.

Schmid, H.-J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin & New York: Mouton de Gruyter.