Randomization tests in language typology

DIRK P. JANSSEN, BALTHASAR BICKEL, and FERNANDO ZÚÑIGA

Abstract

Two of the major assumptions that common statistical tests make about random sampling and distribution of the data are not tenable for most typological data. We suggest to use randomization tests, which avoid these assumptions. Randomization is applicable to frequency data, rank data, scalar measurements, and ratings, so most typological data can be analyzed with the same tools. We provided a free computer program, which also includes routines that help determine the degree to which a statistical conclusion is reliable or dependent on a few languages in the sample.

Keywords: *chi-square test, Fisher's Exact test, interval data, randomization, sampling, statistics*

1. The limits of classical statistical tests in typology

Two major assumptions that common statistical tests rely upon are not tenable for most typological data. We suggest to use randomization tests, which avoid these assumptions and which are applicable to a wide range of data types, so most typological data can be analyzed with the same tools. To establish our case for the use of randomization-based tests, we will first discuss the underpinnings of the two types of statistical tests that are most often used by typologists, and by social scientists in general. The two types of tests are the parametric tests (such as the *t*-test) and the classical non-parametric tests (such as the chi-square test). Both types of tests make assumptions which can be problematic for typological data. We will focus on the assumptions of the parametric tests first.

Linguistic Typology 10 (2006), 419–440
DOI 10.1515/LINGTY.2006.013

1430–0532/2006/010-0419
©Walter de Gruyter

1.1. *The importance and difficulty of random sampling*

The parametric tests include common tests such as the *t*-test, the Analysis of Variance (Anova), and others. The statistical aim of parametric tests is to infer a property of the population from the sample. Let us first define some terms: the languages under study form the SAMPLE; all languages we want to draw a conclusion about (e.g., a family, an area, or the entire world) are the POPULATION. What we loosely referred to as "a property of the population" is more formally known as a POPULATION PARAMETER, hence the name of the tests.

To make a statistically valid inference about the population, the sample has to be constructed randomly: every language in the population has to have the same chance of being included in the sample. Random sampling guarantees that there are no biases in our sample which may lead to incorrect inferences about the population.

There are a number of problems with the notion of selecting random languages for a typological sample. Obviously, selecting only languages that are well described is not in accordance with random sampling, yet a necessity for typological research. Secondly, random sampling assumes we have a multitude of languages to choose from. This is not the case for smaller families. Sampling one language from such a family may already be near-exhaustive or, in the case of isolates, actually exhaustive. Exhaustive sampling defies the inference from sample to population that parametric tests are about, because sample and population are (almost) the same.

There are also conceptual problems. If Basque happens to be excluded from a sample of European languages, one will still arrive at STATISTICALLY valid inferences about all languages spoken in Europe. This is so because Basque had the same chance as other languages to be included in the sample. The fact that it happened not to be included does not change inference. A fortiori, if our sample of European languages happens to contain five Germanic and only one Romance language, this does not change the population that we make inferences about. The population would still be all European languages.

Such inferences might be statistically valid, but many typologists will be unhappy with them because one generally wants to be able to tell genealogical factors apart from structural or areal factors. A sample excluding Basque or a sample heavily biased towards Germanic will obscure the factors of interest. If we find a distributional bias, it is unclear whether the bias is due to the family relationships in the sample, or due to a real characteristic of the entire population of languages in Europe.

Thus, genealogical relatedness is an important CONFOUNDING factor (a factor which obscures the relation that we are interested in). The standard statistical method to control for confounding factors is STRATIFICATION. Under a simplistic approach, the languages of Europe (the population) are divided into strata (genealogical units, e.g., genera or families) and a random sample is then taken, with the crucial constraint that the same number of languages is taken from each stratum. This is very similar to the way in which opinion polls ensure that the same number of men and women are included in order to avoid any gender bias in sampling. It guarantees that Basque is included and that the same number of Germanic and Romance languages are included.

Stratification on genealogical factors works well for large and well-described language families. When it comes to smaller families, however, it increases the problem signaled earlier: as the strata get smaller, random sampling becomes more restrictive, to the point where it becomes exhaustive and inference from parametric tests is no longer warranted. In our example, each stratum will contain only one member, because there is one stratum (genus) that has only one member: Basque. And note that in this stratum, the sample will be the entire population.

To complicate matters, stratification on genealogical relatedness might not be sufficient because it leaves out language-contact factors. Few valid inferences can be made from an (admittedly small) sample of European languages which, through random sampling within the Romance and Germanic strata, happens to contain Romansh and Swiss German as the representatives of the Romance and Germanic families. These two languages have undergone much more intense cross-genera contact than other Romance and Germanic languages and existing areal patterns in Europe will be greatly overestimated based on this sample.

Thus, language-contact factors limit our choice of languages from each genus further. In typological practice the choice of languages in a sample is often guided by many additional criteria, depending on the research question (e.g., only languages with case in a study of case exponence). The list of requirements can get increasingly restrictive, so that even sampling from the larger families effectively approaches exhaustive sampling. Because of this, random sampling of languages is normally unwanted in typology, even where it is possible (pace Widmann & Bakker 2006).

1.2. *The level of sampling*

Can the problem of sampling be solved by looking at a different level of analysis? It has independently been argued (Dryer 1989, 2000) that the correct level of analysis in quantitative typology may not be the level of the language, but that of the genus. Genera should be sampled at random and the population would be “all genera” (in the world, in an area, etc.). We are obviously not committed to the choice of the genus: depending on the scope and nature of the research question, other levels might be appropriate, such as the stock, phylum, or subbranch. (For ease of discussion, we will assume a survey in which the genus is chosen as the level of analysis throughout.)

The problem of exhaustive sampling of languages is now avoided. Each genus is one data point, and each genus could be represented by one language of the genus or by a mean, mode, or other aggregate measure taken from the genus. However, the problem of exhaustive sampling now returns at the level of the genera. One will want to include as many genera as possible to make the sample most representative, but this will lead to exhaustive sampling of all genera in the area of interest.

1.3. *Possible solutions to sampling and inference*

We have seen that parametric tests are not valid for (near-)exhaustive samples because random sampling is a requirement. More technically, if the sample and the population are identical, parametric statistical tests such as the *t*-test and the Analysis of Variance do not apply.[1] What are the alternatives to random sampling and classical statistical inference?

One option is to deny the use of statistical testing altogether or take the perspective that because the populations are sampled completely, statistical testing is not necessary. There are two problems with this position. First, if in one region 8 out of 10 genera prefer the verb to occur in sentence-final position and in the other region 7 out of 10 genera prefer this, who can decide whether this a meaningful difference (cf. Cysouw 2005)? Secondly, this approach disregards an inherent problem when selecting languages by hand rather than by chance: If other languages were taken to represent genera, would the difference in word order still be the same or would it reverse or disappear?

A better solution is to use non-parametric statistical tests, which do not aim to make inferences about the population but speak only of the cases included in the sample. Because of this, random sampling is no longer an issue. We will see below that it is still possible to make LOGICAL INFERENCES (but not statistical inferences) to larger populations from the results of these tests.

Randomization tests are non-parametric, as are more commonly known tests such as the chi-square test. We will discuss the drawbacks of the chi-square test and its relatives in Section 2 and discuss the randomization alternative in Section 3.

To summarize, random sampling is crucial for classical statistical inference from a sample to a population, but it is rarely possible and never desirable

[1] This can easily be demonstrated as follows. A parametric statistic answers the question whether a difference observed in the sample is likely to also be present in the population. A significant *p*-value means that it is unlikely that the difference in the sample came about by chance. If the sample is identical to the population, any difference is therefore significant by definition and no testing is necessary or applicable.

in typology, because of genealogical, areal, and other constraints that lead to exhaustive sampling. Non-parametric tests focus on inference on the data at hand and are a better alternative.

1.4. *Possible misunderstandings about random sampling*

Before we continue, two implications of the above are worth mentioning. Random sampling does not require INDEPENDENCE OF CASES in the sense that each and every case in the sample is independent of all the others. Independence of cases can never be a statistically valid requirement of sampling, as the languages in such a sample cannot be representative of the population. Thus, in a truly random sample of African languages, one is very likely to encounter a good number of Bantu languages, which reflects the fact that the Bantu family is very large.

Random sampling requires INDEPENDENCE OF SELECTION of cases. Each case in the population has the same chance of appearing in the sample. We can therefore never make a valid statistical inference to the population of "all possible human languages" or "all human languages that ever existed", as only a tiny fraction of these languages has any chance of being sampled at all – those spoken now or known from written records.

2. The importance and difficulty of data distribution

If a test is non-parametric, it involves statistical inference at the level of the sample. This makes it more attractive for typology, but not necessarily suitable for it because many tests still put requirements on the distribution of the data. We will demonstrate this issue using the chi-square test and its distribution-free alternative, the Fisher Exact test.

2.1. *The chi-square test*

The commonly used chi-square test (officially known as the Pearson Chi-Square Test) is a non-parametric test that considers the numbers in a table of counts. First, expected values for the cells are derived according to Formula 1. Using the numbers in Table 1 as an example, the expected value for the top-left corner is $14 \times 43/48 = 12.5$. The Pearson chi-square statistic is the sum of all squared differences between the expected values (E) and the observed values (O), scaled by the size of the expected values, as shown in Formula 2. For Table 1, the statistic equals 4.5 (note that the term "statistic" refers to the Pearson quantity, and not to its probability or p-value).

How can we decide whether the observed counts are significantly different from the expected counts? Pearson observed that, for large enough numbers that are evenly distributed over the table, the distribution of the chi-square statistic is similar to that of the independently known χ^2 distribution (to avoid

$$\text{Expected}_{r,c} = \frac{\text{Total}_r \cdot \text{Total}_c}{\text{Total}_{\text{grand}}}$$

Formula 1. *Computing the expected value for a cell in row r and column c*

$$\text{Chi-Square} = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Formula 2. *Computing the Pearson Chi-Square statistic*

Table 1. *Number of languages which have possessive classes in the Caucasian and Himalayan Enclaves versus the Rest of Eurasia, Mundari data set (Bickel & Nichols 2003, Nichols & Bickel 2005)*

Area	Possessive classes					
	Observed			Expected		
	No	Yes	Total	No	Yes	Total
A: Enclaves	10	4	14	12.5	1.5	14
B: Rest of Eurasia	33	1	34	30.5	3.5	34
Total	43	5	48	43	5	48

Fisher's Exact test: Columns are not independent of rows.
Chi-square test: Invalid test due to two expected values lower than five.
Randomized Chi-square test (see Section 3.1): Columns are not independent of rows.

confusion, we will use chi-square statistic and χ^2 distribution). Only with help of the χ^2 distribution can a Pearson chi-square statistic be mapped to a probability level (e.g., $p = 0.12$).[2]

This means that the Pearson chi-square is not distribution-free, because the use of the χ^2 distribution rests on the assumption that the numbers are large enough and evenly distributed over the table. Regrettably, there is no strict definition of "large enough numbers" or of "evenly distributed". The rule of

[2] The χ^2 distribution with *k* degrees of freedom is computed by summing *k* squared, independent, standardized normal variables. So $\chi^2(4)$ is distributed as $\chi(Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2)$, with each Z a standardized normal variable. A three-by-three table has 4 degrees of freedom and, by mathematical proof, its Pearson chi-square value will be distributed as $\chi^2(4)$ when the cell counts in the four topmost-leftmost cells can be assumed to behave like normal variables. This is the case when those counts are large enough.

thumb that is usually given is that no cell should have an expected count smaller than five (Howitt & Cramer 2005). An alternative version of this rule states that at least 80 percent of the expected numbers must be larger than or equal to five and all should be larger than one (Cochran 1954).

These are all highly problematic restrictions in typology, because empty cells and cells with small values are particularly interesting. They suggest heavy biases in the data, and yet these tables are intrinsically hard to test with the Pearson chi-square test.

2.2. *Fisher's Exact test*

For the specific case of a two-by-two table of counts, a non-parametric and distribution-free test was proposed by R. A. Fisher in the 1930s. The Fisher Exact test can be run in all major statistics packages and in various other tools, and is preferable to the Pearson chi-square in all cases of two-by-two tables with small expected values.

Distribution-free tests put fewer (if any) requirements on the distribution of the observed data. Examples of such requirements are the need to have data that are normally distributed (Anova), or minimum expected values in each cell (Pearson chi-square).[3]

Similar to the randomization tests we will advocate below, the Fisher Exact test considers all possible ways in which a two-by-two table can be constructed without changing the margin totals. In a standard typological table such as shown in Table 1, the margin totals are the number of languages in the two regions and the total number of languages that have, or do not have, the feature under study. These data are taken from Bickel & Nichols (2003) and Nichols & Bickel (2005).[4]

For the example given in Table 1, the test will consider the many ways in which 14 Enclave languages and 34 languages from the Rest of Europe can divide the five languages which have possessive classes between them. There are usually several hundreds or thousands such tables, depending on the number of data points. Significance is determined by finding out how exceptional the observed data are. If the observed table occurs in the range of possible tables only rarely, it is more exceptional, and it is less likely that the observed table

[3] The term can be considered a bit of misnomer: distribution-free tests do not imply that the data do not have ANY distribution (a statistical impossibility), but they do not rely on a KNOWN distribution.

[4] The datasets contained in the *World Atlas of Language Structures* are not genealogically balanced samples. For these papers, we used genealogically balanced samples based on the *WALS* data (but our samples may include more languages because data collection has continued since submitting the *WALS* chapter). We have made these data sets available at http://www.uni-leipzig.de/~autotyp/available.html (files possclg_eurasia.csv and syng_eurasia.csv). The files are also included in the Trotter package.

has occurred by chance. In the example, there are very few other tables that are like the observed data, and the Fisher Exact test concludes that rows and columns are not independent. That is to say, the number of languages with possessive classes is not randomly distributed, but larger in one area (the Enclaves) than in the other.[5]

No statistical test is perfect, but Fisher's Exact test has been severely criticized by some statisticians. The major point of contention is the assumption of fixed margins, which we will discuss in detail in Section 3.4. If the data is not compatible with this assumption, Fisher's test is quite conservative, i.e., some tables which depart from independence will not result in a significant *p*-value (D'Agostino, Chase, & Belanger 1988). With this precaution in mind, Fisher's Exact test is still the recommended distribution-free test for all two-by-two tables which violate the assumptions for the Pearson chi-square test. If, for a particular data set, there is a strong concern about this test being too conservative, a randomized chi-square test (without Yates continuity correction) can be applied.

3. Randomization tests

The Fisher Exact test cannot solve all statistical problems in typology, because it is limited to two-by-two tables of counts. Randomization tests are also nonparametric AND distribution-free, as they are a generalization of Fisher's Exact test. (Historically, Fisher's Exact test is an easy-to-compute version of what is now known as the class of Fisher-Pitman tests or randomization tests, see Sprent 1998 and Fisher 1935.) Randomization tests can be applied to larger tables of counts and, as we will see below, to measurements different from counts. The randomization approach can therefore provide a unified solution to a number of statistical problems in typology.

3.1. *Randomization of frequency tables larger than two-by-two*

The statistical question that underlies all randomization testing is whether the observed data could have been generated by chance. Consider the frequency tables in Table 2. The first panel (2a) shows the observed table and the other panels (2b, 2c) show two alternative tables with the same margin totals. The alternative tables are constructed by assuming that, if chance has it, the features F1 to F3 are randomly distributed over the three regions. In technical terms, they assume INDEPENDENCE of rows and columns.

[5] To avoid confusion, we have explained the Fisher Exact test in terms of listing all possible tables. In actual fact there is a statistical distribution, the hypergeometric distribution, which summarizes a listing of all cases. Even though Fisher's Exact test is computed with the help of this distribution, it is a distribution-free test because it does not make any assumptions about the distribution of the observed data, like the chi-square test does.

To relate these data to the previous example, one could think of F1 as having zero possessive classes, F2 as having a few, and F3 as having many possessive classes. Clearly, the observed table is unlike either of the two tables based on chance, as F1 is relatively commonly observed in Area A, F2 in Area B, and F3 in Area C. In the middle and rightmost panels, no such strong regularities exist.

The statistical question of interest is how likely it is that the observed table is a product of chance alone. We can re-phrase this question to: How far is the observed data removed from a totally independent table (as shown in the rightmost panel)? One way to measure this is to use the value of the Pearson chi-square statistic (but not the *p*-value connected to it). As explained above, the Pearson statistic compares the observed values with the values expected under independence using Formula 1.

For the three data sets in Table 2, the Pearson chi-square statistics are 10.6, 2.9, and 1 (respectively), which corresponds with the idea that the observed data are furthest removed from independence. Note that chance data sets are not always completely independent: the middle panel shows data in which F2 does not occur at all in area C, although this is entirely coincidental.

In a classical Pearson chi-square test, we determine the probability of the chi-square value of 10.6 (for the observed table) from the χ^2 distribution. We saw in Section 1 that this is unwarranted for the current data, as there are many small expected values. A randomization test using Pearson chi-square will instead consider about 10,000 alternative tables[6] constructed on the basis of chance, similar to those shown in the middle and right panel of Table 2. Next, a Pearson chi-square statistic is computed for each alternative table. If the observed chi-square value is much higher than most of the values that are found in the chance-based tables, the observed table is unlikely to be due to chance. In the example in Table 2, the observed value of 10.6 is indeed higher than the two alternative values (2.9 and 1) and compared to another 9,998 tables, it will turn out to be significantly higher, with only 23 alternative tables meeting or exceeding this value, $p < 0.05$.[7] Figure 1 shows this graphically: the smoothed frequency (density) of finding chi-square values between 0 and 20 is shown by the curve. The 5 % largest values of chi-square are those to the right of the line that is labeled "estimated $p = 0.05$" and are shaded. The observed chi-square value of 10.6 is inside this shaded area and therefore significant.

[6] One may wonder why 10,000 tables are enough and whether more tables are better. Simulations have shown a number of this magnitude to give robust and consistent results. A larger number of tables will only increase the precision of the final *p*-value, something which is usually not required. See Edgington (1995) for discussion.

[7] Significance in this randomization test is defined in a very straightforward way: if the observed outcome is among the 5 percent highest chi-square values, it is significant.

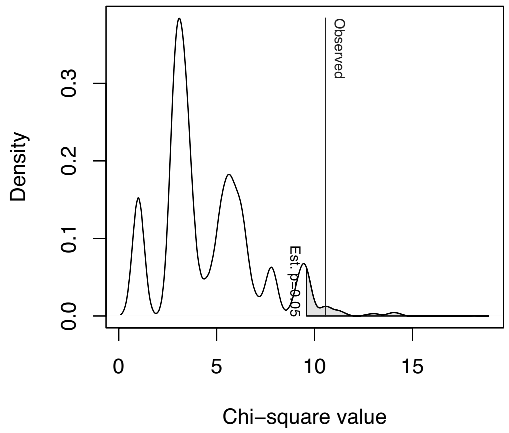

Figure 1. *Simulated values of chi-square, graph produced by the Trotter routine*

If however the observed data had been those shown in the middle panel of Table 2, the randomization test would lead us to conclude that this pattern is very likely to occur by chance because a large number of chance-based table (854 in our simulation) are as far removed from independence as the middle table is, $p > 0.10$.

3.2. *Follow-up tests*

We have now rejected the independence model, which holds that areas and feature values are independent for the observed data in Table 2. In other words, in the observed data, the distribution of features F1 to F3 is not the same for areas A, B, and C. It is usually of interest to examine exactly where the differences are. Follow-up tests can be done on selected columns and rows of the observed

Table 2. *Example of randomization testing: Observed data and two alternatives that will be considered by a randomization test*

A: Observed Table					B: Alternative Table					C: Alternative Table				
	Feature Value					Feature Value					Feature Value			
	F1	F2	F3			F1	F2	F3			F1	F2	F3	
Area A	2	0	1	*3*	Area A	1	1	1	*3*	Area A	1	1	1	*3*
Area B	0	2	0	*2*	Area B	0	1	1	*2*	Area B	1	0	1	*2*
Area C	1	0	3	*4*	Area C	2	0	2	*4*	Area C	1	1	2	*4*
Total	*3*	*2*	*4*	*9*		*3*	*2*	*4*	*9*		*3*	*2*	*4*	*9*

data, and on new columns and rows created by merging existing cells. One possible set of comparisons, using both existing and merged rows, would compare area A versus area B and compare area C versus the sum of area A and B.

In a table with three columns, no more than two follow-up tests should be done. Ideally, these tests should also be independent, that is, the significance of the first follow-up tests should not influence the second one (Agresti 2002). If the tests are not independent, a Bonferroni or Holm correction should be applied to test at a stricter alpha level, here $0.05/2 = 0.025$. Note that the follow-up tests should be planned before the data are collected, too many significant results will be obtained if one simply compares the lowest and the highest scoring areas.

The comparison set introduced above involved area A versus B and area C versus the sum of A and B. This set is statistically independent and complete, that is, the two comparisons together test the same questions as the original randomized chi-square test on the complete table.[8]

3.3. *Comparing classical and randomization tests*

There are a number of advantages of the randomized test over the classical ones. First, the Pearson chi-square is not applicable to the data in Table 2 because its assumptions (expected values all five or larger) are violated. If one does apply the Pearson chi-square, conservative estimates are obtained which will make few comparisons significant.

Fisher's Exact test cannot be applied to tables larger than two-by-two in its canonical form. An extension to larger tables was made by Mehta & Patel (1986; this test is available in *R*) but this extension cannot handle tables with too many columns or rows, and it will run very slowly when one or two cells have high counts in them (not an unusual situation in typology).

Complex multivariate techniques, such as log-linear analysis or logistic regression, are better suited for tables of three or more dimensions. Neither technique can cope well with small cell counts, because even the log-transformed counts or odds behave non-linear in this domain; but see Agresti (2002) for further suggestions.

Randomization tests, on the other hand, can easily deal with this table, larger tables, and with tables of higher dimensionality (for example, a table which distinguishes area, word order, and possessive classes). Although they are not necessarily very fast to compute, they do not take more time when the table becomes more complex.

[8] If likelihood-ratio chi-square values (Wilks' G^2) are used instead of Pearson chi-square values, the follow-up tests will sum exactly to the overall test, see Sprent (1998) or Agresti (2002) for details.

Below, we will discuss applications of randomization to data that are traditionally analyzed with a Mann-Whitney U-test, a Wilcoxon rank-sum test, or a Kruskal-Wallis test. The randomization tests have the advantage that the same principle can be applied to all these data, obviating the need to know about a number of disjunct tests that all have their own narrow field of application.

The randomized test also has drawbacks. The major drawback we see is the current lack of popularity: randomization tests are not yet as widely used as classical tests. Statisticians use randomization (and its relative, bootstrapping) routinely, but, with the notable exception of genetics and a few other sciences, many fields have not caught up yet. Furthermore, not all statistical software packages support these tests. Again, this is changing rapidly (SPSS now provides a related module for the Windows platform), and we have created some easy-to-use scripts for the free statistical software called *R*, which we will describe below.

Non-parametric and distribution-free tests are usually considered less powerful than the parametric tests. This means that the parametric tests are more likely to detect a difference where one exists, where non-parametric tests might occasionally miss an existing difference. Sprent (1998) argues that this received wisdom is only correct if the comparison is made on data to which the parametric tests are applicable. For data that violate the assumptions of parametric tests, non-parametric tests are much more powerful. Still, randomization tests should only be applied to data that do not fully satisfy the assumptions of parametric tests.

We have seen that non-parametric tests focus on statistical inference on the observed cases (the sample). This can be a drawback, as scientists normally want to draw conclusions that go beyond the observed cases. Logical inference can alleviate this problem: if we can argue on typological grounds that the sample is representative of the larger unit (say, the area), we can logically extend the conclusions drawn on the basis of the sample to the area. Logical inference is at the heart of most sciences: if a psychologist performs a test on college students, the results can be extended to the larger population only by logical inference. Statistical inference is not possible, as the members of the larger population had no chance of being included in the sample (see Section 1.4). The premise for the logical inference is that we do not expect college students to be different from the population on dimensions relevant for the test.

Randomization testing requires computing resources. It used to be the case that computing statistics on 10,000 random tables required substantial computer power and much patience. With modern computers, this is no longer an issue. On a machine which is modest by current standards (AMD Athlon at 1.3 GHz), our scripts used less than 20 seconds to compute the 10,000 randomized chi-square values mentioned above.

As a final point, researchers who are new to randomization often notice that the results of the tests cannot be exactly replicated. Because significance is based on a comparison to 10,000 random tables, a slightly different result will be obtained if the same test is run again, on 10,000 different tables. However, the randomization significance level is precise to 0.0001 (1/10,000), which means that a second run would have to be different in 100 tables to make a 0.01 difference in the significance level. In practice, different runs produce significance values well within 0.001 from each other and such variation is harmless. Further discussion of the benefits and drawbacks of randomization tests can be found in Edgington (1995), Manly (1991), and Sprent (1998).

3.4. *The assumption of fixed margins*

Both Fisher's Exact test and the randomization tests using chi-square crucially assume that all alternative tables are constructed so that they have the same margin totals. This assumption of fixed margins has been the subject of a long-standing debate in statistics. Sprent concludes that "its validity in most situations is now widely accepted" (1998: 333); but see Agresti (1992) for discussion.

In the randomization tests used above, assuming fixed margins seems a viable approach to us. The randomization essentially considers what the chances are of finding the observed data given that there are three, two, and four languages in the areas A, B, and C and there are three, two, and four languages with feature values F1, F2, and F3. The assumption of fixed margins allows us to look at the distribution of F1 to F3 over the areas, without making any assumptions about the overall chance that F1 to F3 occur.

If one drops the assumption of fixed margins, the randomization test essentially becomes a bootstrap test. Bootstrap tests also compare the observed data to a range of alternative tables, but the existing cases are sampled WITH replacement (randomization samples without replacement). This sampling means that an alternative table may have more F1 languages than the observed table, if an F1 language was included twice at the expense of an F2 language. Bootstrap tests are a good way to estimate population parameters from non-normal data.

It follows from our discussion above that we deem randomization tests more appropriate to typological data because statistical inferences about the population are not required or are even unwarranted because of exhaustive sampling (cf. Section 1). Also, even the bootstrap cannot reliably estimate population parameters from less than 10 to 15 cases per cell. However, we think it is a distinct advantage of randomization tests over other non-parametric tests that randomization can easily be replaced with its cousin, the bootstrap, if one does not want to assume fixed margins and has large enough numbers.

4. Expanding randomization to other data types

We alluded to the wide applicability of randomization to other types of data than counts above. We now discuss which types of data are traditionally distinguished, and how randomization can deal with these.

4.1. *Types of data*

Three types of data are classically distinguished in science: nominal, ordinal, and interval (Howitt & Cramer 2005). Many classical typologies consider nominal data, such as the absence or presence of a phenomenon (Table 1). A slightly extended version of nominal data is a typology which distinguishes a number of labels (such as in Table 2). As an actual example, Östen Dahl's typology of the words for 'tea' has "derived from *cha*", "derived from *te*", and "other" (Dahl 2005). In the case of nominal data, a statistical analysis considers the frequency with which each label occurs and we suggest to use a randomization test on Pearson chi-square values.

For other typologies the labels given to each language can be ordered in some way, leading to an ordinal or rank measurement. Consider Ian Maddieson's study of consonant inventories (Maddieson 2005), which uses the labels "small", "moderately small", "average", "moderately large", and "large". Running frequency statistics on rank measurements is not incorrect, but such an analysis is insensitive to the fact that "moderately large" and "large" are much closer to each other than "small" and "large" are.

It is important for rank-based analyses that all labels can be ordered. If there is an "other" label that cannot be compared to the others because it is a mixed bag of cases, we can either only use frequency statistics, or we have to exclude all languages of the "other" type. If we include "other", we have what is technically called an INCOMPLETE RANKING, in which some but not all categories can be ordered.

David Gil's typology of "Genitives, adjectives and relative clauses" (Gil 2005) shows another type of incomplete ranking (see Table 3). It is probably impossible to objectively rank labels 2 to 5. To use rank-statistics on these data, one could collapse to a three-way distinction as indicated in the rightmost column. Of course, this would mean loss of precision in the data, as we no longer distinguish between different types of moderately differentiated languages. An ideal analysis would therefore consider both a frequency-based analysis of all data and a rank-based analysis of the collapsed data.

Rank data can be analyzed with classical non-parametric tests like the Mann-Whitney U-test (for two groups) or the Kruskal-Wallis test (for more than two groups). These tests are closely related to each other and to the Wilcoxon rank-sum test, the Jonckheere-Terpstra test, and the Kendall rank correlation test (Sprent 1998, Agresti 2002). All are included in standard statistical packages

Table 3. *Typology of genitives, adjectives, and relative clauses, from Gil (2005)*

	Original label	Ranked label
1:	weakly differentiated	weak
2:	genitives and adjectives collapsed	moderate
3:	genitive and relative clauses collapsed	moderate
4:	adjectives and relative clauses collapsed	moderate
5:	moderately differentiated in other ways	moderate
6:	highly differentiated	strong

and are described in textbooks like Howitt & Cramer (2005). As this is a bewildering array of tests, we recommend to analyze rank data with randomized Anova techniques (see below) if the number of ranks is large enough, and as nominal data otherwise.

A final type of data is that of interval measurements (including ratio measurements, which can be treated as interval measurements for all statistical purposes). The crucial difference between a rank and an interval measurement is that, for interval measurements, the distance between all values is the same. There are currently only few typological measurements of interval type, but there are a number of typologies that come close enough to be analyzed as interval measurements. An example is Bickel & Nichols' (2005a) typology of verbal inflectional synthesis, which we will discuss below. However, one can also use interval statistics on typologies that use a rating scale to express how much each languages exhibits a feature, if the rating scale has at least five (preferably seven) levels and one is willing to make the simplifying assumptions that the points on the scale are approximately equidistant. Maddieson's (2005) consonant classification into five ranks, described above, could also be analyzed as interval data, which brings several benefits in terms of data analysis.

4.2. *Randomized Anova*

We will briefly consider how the randomization method can be used for interval data. Table 4 shows the distribution of inflectional verbal synthesis over two areas, the Caucasian and Himalayan Enclaves versus the Rest of Eurasia (Bickel & Nichols 2003, 2005a). Two things are obvious from looking at these data: the Enclaves seem to have overall higher values for synthesis than the Rest of Eurasia, but the counts for the Enclaves are very small.

A classical Anova (analysis of variance) would be a good test for these data, as it can tell us whether there is a difference between the values of synthesis in each region and, if so, which regions differ from each other. However, the data are not even close to normally distributed and certainly not randomly sampled,

Table 4. *Inflectional Synthesis of the Verb, data for Caucasian and Himalayan Enclaves (Enclaves) versus the Rest of Eurasia (R. Eurasia) (from Bickel & Nichols 2003, 2005a)*

	Observed values of synthesis																		
	0	2	3	4	5	6	7	8	9	10	11	12	13	14	15	18	20	22	25
Enclaves	0	0	0	0	3	3	1	1	0	1	0	5	0	0	1	1	1	1	1
R.Eurasia	1	1	1	3	3	9	5	1	2	4	2	0	2	1	0	0	0	1	0
Total	*1*	*1*	*1*	*3*	*6*	*12*	*6*	*2*	*2*	*5*	*2*	*5*	*2*	*1*	*1*	*1*	*1*	*2*	*1*
	Expected values of synthesis, derived from column and row totals																		
	0	2	3	4	5	6	7	8	9	10	11	12	13	14	15	18	20	22	25
Enclaves	0.3	0.3	0.3	1	2.1	4.1	2.1	0.7	0.7	1.7	0.7	1.7	0.7	0.3	0.3	0.3	0.3	0.7	0.3
R. Eurasia	0.7	0.7	0.7	2	3.9	7.9	3.9	1.3	1.3	3.3	1.3	3.3	1.3	0.7	0.7	0.7	0.7	1.3	0.7
Total	*1*	*1*	*1*	*3*	*6*	*12*	*6*	*2*	*2*	*5*	*2*	*5*	*2*	*1*	*1*	*1*	*1*	*2*	*1*

Descriptive statistics

Enclaves:	19 cases, mean 11.5, median 12, standard deviation 6.1
R. Eurasia:	36 cases, mean 7.5, median 6.5, standard deviation 4.0
Total:	55 cases, mean 8.9, median 7, standard deviation 5.1

which would make any application of a classical Anova flawed. Because this is a complete sample of the Enclaves, the inferences drawn from this test would be meaningless, as argued in Section 1.

Still, if we compare the observed counts with the expected counts, it is clear that the number of Enclave languages with a synthesis value of 12 is much higher than expected (5 observed, 1.7 expected). Also, both mean and median synthesis value for the Enclaves are much higher than for the Rest of Eurasia, which is partly due to the fact that the Rest of Eurasia has consistently fewer languages at the high end of the synthesis scale (0 observed, 0.7 expected) and more languages at the lower end of the scale (1 observed, 0.7 expected). This observation can be made more statistically precise using a randomization method.

A randomized Anova works like the randomized chi-square test described above. For each alternative table, languages are assigned to cells of the table at random, with the constraint that the margin totals stay the same. For each table, an Anova F-value is computed.

The F-value expresses how large the differences BETWEEN the areas are, relative to the differences WITHIN the areas. The assumption is that the F-value for each alternative table might not be fully accurate because of the low numbers in the Enclaves, but the ordering of F-values is correct across all alternative tables. If the observed F-value is much higher than what is routinely found among random alternative tables, the observed table is significant because it is unlikely to be due to chance.

For the comparison between the Enclaves and the rest of Eurasia, the randomized Anova with 10,000 alternatives yields a main effect of Area, which is significant at $p = 0.0054$. Thus, we can conclude that languages in the Enclaves tend to have higher synthesis degrees than those in the Rest of Eurasia.

5. Reliability and misclassification

After any successful statistical analysis has been done, one may wonder what would have happened if a problematic language had been classified or analyzed differently. Additionally, what would have happened if we had chosen other languages from each genus? To some extent, this is an empirical problem that cannot be solved by any statistical procedure. However, what statistics can do is estimate the degree to which our data are sensitive to the issue of misclassification and thereby help us determine how reliable our findings are. We propose a method for this, which involves computing the statistics on all alternative scenarios that are one or more misclassifications away from our observed data and graphing the results. If there are many ways in which one misclassified language changes the significance of the results, we have to be careful when interpreting the data. If one or a few misclassified languages do not make any difference, we can be more certain of our case.

5.1. *Reliability for count data*

The reliability graph is closely related to randomization testing. Recall that randomization testing is based on finding alternative tables with the same margin totals. If, instead, we alter the margin totals in such a way that only the total number of languages (the sample size) remains unaltered, we can explore the issue of misclassification. As an example, we examine the data on possessive classes (POSSCL) from Table 1 again. The two categories are again languages with or without possessive classes. For the two-by-two count table created by looking at this dichotomization for the Enclaves versus the Rest of Eurasia, a reliability landscape is produced as shown in Figure 2. In the figure, the number of positive cases (languages with possessive classes) increases as we move up and to the right.

The bold-faced square near the bottom of the figure symbolizes the observed data point. This point has coordinates (4,1), that is, there are four POSSCL languages in the Enclaves (plotted from left to right) and one in the rest of Eurasia (plotted up-down). The shading of this square signals a significance level of $p < 0.05$. The square to the right of the bold printed one has coordinates (5,1): it is the hypothetical case in which there are five POSSCL languages in the Enclaves and one in the Rest of Eurasia. This square is more significant

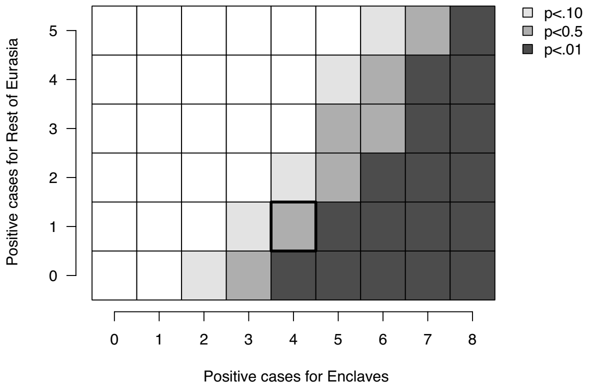

Figure 2. *Reliability Landscape for a two-by-two table, with increasing levels of shading indicating significance at or below 0.10, 0.05 and 0.01 levels. White squares are not significant, p > 0.10.*

(p < 0.01) as our observed data point. In other words, we do not have to fear for having missed one POSSCL language in the Enclaves because if such a language exists, the result of our statistical tests would have been even stronger.

If, on the other hand, we have spuriously classified one Enclave language as POSSCL whereas it is not, we are in less good shape: The (3,1) point in the graph, which shows the hypothetical case in which there are only three POSSCL language in the Enclaves and one in the Rest of Eurasia, is no longer significant ($p < 0.10$). Similarly, if we had missed one POSSCL language from the Rest of Eurasia, this would change our results: the (4,2) square, symbolizing four POSSCL languages in the Enclaves and two in the Rest of Eurasia, is not significant ($p < 0.10$).

We find ourselves in hot water if, for some reason, we suspect that we have overclassified two Enclave languages as POSSCL that are actually not POSSCL (point 2,1), or if we think it is likely that we have missed two POSSCL languages in the Rest of Eurasia (point 4,3). In either case, the resulting square is white, that is, the difference between the regions is not significant at all ($p > 0.10$).

The reliability graph has made it clear that the data in Table 1 are sensitive to the issue of misclassification in that one misclassified language can lead to loss of the observed significant difference between the areas. Table 1 is based on data with Mundari as the representative of the Munda genus in the Rest of Eurasia. A possible alternative is to take Korku in place of Mundari, in which case none of the languages in the Rest of Eurasia show evidence for possessive classes. The reliability graph makes it clear that choosing Korku leads to an increased observed significance (point 4,0, for which $p < 0.01$). If Korku is

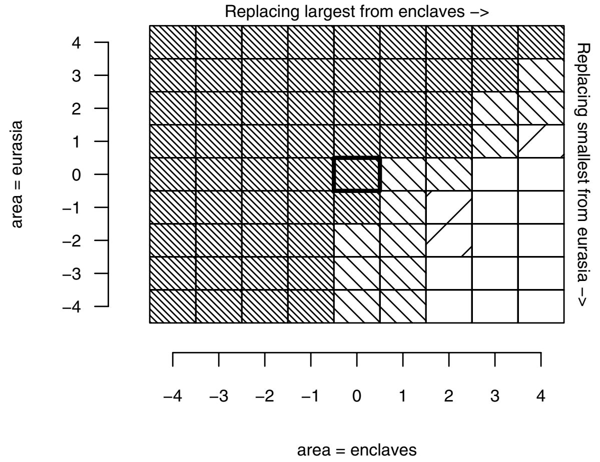

Figure 3. *Anova reliability graph for synthesis data, increasing levels of cross-hatching indicate significance of the $p < 0.10$, $p < 0.03$, and $p < 0.01$ level*

chosen, this also implies that the finding is more reliable because two languages have to be misclassified before the observed significant difference is lost (one has to either find two Eurasian languages with possessive classes, point 4,2; lose two Enclave languages with possessive classes, point 2,0; or one of either, point 3,1).

Thus, the reliability graph allows careful examination of how trustworthy the finding of statistical significance is in a test, and to what degree it depends on coding accuracy and sample selection. This does not of course do away with the empirical problems of uncertainty in typological data-gathering, but it allows us to estimate the impact of this uncertainty beyond mere impressions.

5.2. *Reliability for interval data*

The reliability approach is not limited to count data. We have also developed a similar graphical method for the issue of reliability with interval measures. In this case, the reliability of the statistical tests is most strongly influenced by the presence of OUTLIERS, i.e., languages that fall outside of the body of observations. For example, if Table 4, discussed above, had also included an Enclave language with a synthesis value of 50, this would have been a clear outlier: it is more than twice as high as the next language. With such an outlier, the average for synthesis in the Enclaves would have jumped from its current value 11.5 to 13.4 and the difference with the Rest of Eurasia (average synthesis value 7.5) would have been artificially inflated.

The existing Enclave languages with synthesis values of 20 and higher are not such clear outliers, but one might still wonder what their influence on the randomized Anova analysis is. The Anova reliability graph displays the effect of replacing the smallest and largest value from each area with the grand mean. By replacing these values, the size of the sample stays the same but the influence of the most extreme cases is removed.

This procedure is repeated for replacing the two largest and the two smallest values, and so on for more replacements. The result (shown in Figure 3) is a graph that is very similar to what we have seen for count data above. The observed case is indicated by the bold face square at $(0,0)$. Its shading indicates a probability of $p < 0.01$. The languages of the Enclaves are plotted left to right. When moving right from $(0,0)$, we observe the effect of replacing the largest cases in the Enclaves (i.e., the mean is substituted for the cases with the largest value for synthesis). Because the mean synthesis value for the Enclaves was higher than mean synthesis for the Rest of Eurasia, replacing the largest cases of the Enclaves will reduce the difference. At $(1,0)$, replacing the one largest value indeed leads to a reduced significance level of $p < 0.05$. If the three largest cases from the Enclaves are replaced in (3,0), the difference between the Enclaves and Eurasia becomes non-significant at $p > 0.10$.

Because the Enclaves have the higher mean synthesis value, replacing the smallest cases in the Enclaves (i.e., those with the smallest values for synthesis) with the mean will only enhance the difference between the areas. This is seen when going left from (0,0), where negative numbers indicate the number of smallest cases replaced. Similarly, going up from $(0,0)$ the largest cases in the Rest of Eurasia are replaced. This also enhances the difference between the two areas.

Finally, going down from $(0,0)$ we observe the effect of replacing the lowest cases in Eurasia. This will increase the average synthesis values in the Rest of Eurasia, which reduces the significance level. But even with the four smallest cases replaced at $(0,-4)$, the significance level is still $p < 0.05$.

The critical observations to make are that when the three highest synthesis values in the Enclaves are replaced, the significance level rises to $p > 0.10$ (point 3,0). It does not matter how many of the lowest values from the Rest of Eurasia we replace, as $(0,-4)$ is still at $p < 0.05$. The combined replacement of 2 largest Enclave languages and three smallest Eurasian languages will also lead to loss of significance, $p > 0.10$ at $(2,-3)$.

Again, the reliability graph allows us to form an impression of the degree to which the statistical result depend on coding accuracy and sample selection. The graph does not solve the issue of uncertainty, but it can help focus the discussion on the most important threats to the analysis, like square $(3,0)$ in Figure 3.

6. Conclusion

We think the techniques presented here can greatly facilitate progress in the study of typology. To make the randomization statistic and the reliability graph accessible, we have implemented all the necessary routines for randomization testing and for creating reliability graphs in a free statistical software package called *R* (Ihaka & Gentleman 1996, R Development Core Team 2005, http://www.r-project.org). *R* is very powerful but not particularly user-friendly for occasional users. Our routines implement the tests (which are quite simple) and provide a textual user interface which makes running these test more intuitive, even if one has never used *R* before. The software can be downloaded from the first author's website http://www.kent.ac.uk/psychology/department/people/janssend/trotter/, or from http://www.uni-leipzig.de/~autotyp/ and is further described in Janssen (in preparation).

Received: 16 September 2005
Revised: 19 September 2006

University of Kent
Universität Leipzig
Universität Zürich

Correspondence addresses: (Janssen) Department of Psychology, Keynes College, University of Kent, Canterbury, Kent CT2 7NP, UK; e-mail: d.janssen@kent.ac.uk; (Bickel) Institut für Linguistik, Universität Leipzig, Beethovenstraße 15, 04107 Leipzig, Germany; e-mail: bickel@uni-leipzig.de; (Zúñiga) Seminar für Allgemeine Sprachwissenschaft, Universität Zürich, Seminarstrasse 1103, 8057 Zürich, Switzerland; e-mail: fernando_zuniga@gmx.net

Acknowledgements: The majority of this work was carried out while all three authors worked at the Linguistics Institute in Leipzig, Germany. We would like to thank Michael Cysouw, Dik Bakker, other participants of the *Using WALS Workshop* (Leipzig, December 2004), and three anonymous reviewers for valuable comments. We would also like to thank the many authors of the *R* language for making their invaluable resource freely available.

Our work on randomization started in 2003, building on the seminal work of Manly (1991) and spurred by the first major application of randomization and bootstrapping in social science by Shrout & Bolger (2002). Some of the techniques described here were previously demonstrated in Bickel & Nichols (2003, 2005b) and Janssen, Zúñiga, & Bickel (2004); for related approaches, see also Shosted (2006) and Maddieson (2006).

References

Agresti, Alan (1992). A survey of exact inference for contingency tables. *Statistical Science* 7: 131–177.

— (2002). *Categorical Data Analysis.* 2nd edition. New York: Wiley.

Bickel, Balthasar & Johanna Nichols (2003). Typological enclaves. Paper presented at the 5th Conference of the Association for Linguistic Typology, Cagliari, Italy. Available at http://www.uni-leipzig.de/~autotyp/download

— (2005a). Inflectional synthesis of the verb. In Haspelmath et al. (eds.) 2005, 94–97.

— (2005b). Areal patterns in the *World Atlas of Language Structures*. Paper presented at the 6th Conference of the Association for Linguistic Typology, Padang, Indonesia. Available at http://www.uni-leipzig.de/~autotyp/download

Cochran, William G. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics* 10: 417–451.

Cysouw, Michael (2005). Quantitative methods in typology. In Gabriel Altmann, Reinhard Köhler, & R. Piotrowski (eds.), *Quantitative Linguistics: An International Handbook,* 554–578. Berlin: Mouton de Gruyter.

D'Agostino, Ralph, Warren Chase, & Albert Belanger (1988). The appropriateness of some common procedures for testing the equality of two independent binomial populations. *The American Statistician* 42: 198–202.

Dahl, Östen (2005). Words for 'tea'. In Haspelmath et al. (eds.) 2005, 552–555.

Dryer, Matthew S. (1989). Large linguistic areas and language sampling. *Studies in Language* 13, 257–292.

— (2000). Counting genera vs. counting languages. *Linguistic Typology* 4: 334–350.

Edgington, Eugene (1995). *Randomization tests*. 3rd edition. New York: Dekker.

Fisher, Ronald Aylmer (1935). The logic of inductive inference. *Journal of the Royal Statistical Society Series A* 98: 39–54. Available at http://digital.library.adelaide.edu.au/coll/special/fisher/124.pdf

Gil, David (2005). Genitives, adjectives, and relative clauses. In Haspelmath et al. (eds.) 2005, 246–249.

Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) (2005). *The World Atlas of Language Structures.* Oxford: Oxford University Press.

Howitt, Denis & Duncan Cramer (2005). *Introduction to Statistics in Psychology*. 3rd edition. Harlow: Pearson.

Ihaka, Ross & Robert Gentleman (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5 (3): 299–314.

Janssen, Dirk P. (in preparation). *The Trotter Package for Randomization and Reliability Graphs.*

Janssen, Dirk P., Fernando Zúñiga, & Balthasar Bickel (2004). Statistical explorations of the *WALS* data. Paper presented at the workshop on "Using the World Atlas of Language Structures", Leipzig, Germany, 11 December.

Maddieson, Ian (2005). Consonant inventories. In Haspelmath et al. (eds.) 2005, 10–13.

— (2006). Correlating phonological complexity: Data and validation. *Linguistic Typology* 10: 106–123.

Manly, Bryan (1991). *Randomization and Monte Carlo Methods in Biology*. London: Chapman & Hall.

Mehta, Cyrus R. & Nitin R. Patel (1986). Algorithm 643. FEXACT: A Fortran subroutine for Fisher's exact test on unordered $r \cdot c$ contingency tables. *ACM Transactions on Mathematical Software* 12: 154–161.

Nichols, Johanna & Balthasar Bickel (2005). Possessive classification. In Haspelmath et al. (eds.) 2005, 242–245.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. Wien: R Foundation for Statistical Computing.

Shosted, Ryan (2006). Correlating complexity: A typological approach. *Linguistic Typology* 10: 1–40.

Shrout, Patrick E. & Niall Bolger (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods* 7: 422–445.

Sprent, Peter (1998). *Data Driven Statistical Methods*. London: Chapman-Hall.

Widmann, Thomas & Peter Bakker (2006) Does sampling matter? A test in replicability, concerning numerals. *Linguistic Typology* 10: 83–95.