GUIDELINES FOR STATISTICAL ANALYSIS

Quantitative arguments in dialectology should be supported with statistical analysis. It is important not only to report your observation of a difference or correlation between two groups or measures, but also to determine the **degree of confidence** with which we can reject a **null hypothesis** that the pattern you have observed could have arisen purely by chance, given the quantity of data you are looking at. This degree of confidence about a difference between two groups increases with three factors: the number of observations made; the size of the difference; and the extent to which the members of each group are more similar to one another than to members of the other group (that is, the relative size of between-group vs. within-group differences).

Statistics is a complex science that we cannot hope to do justice to in a linguistics course. If you have not already done so, you are strongly encouraged to take a statistics course at some point in your education. A knowledge of statistics can be extremely useful in many pursuits. These guidelines are not a substitute for proper training in statistics. They provide a simplified introduction to two of the most common and widely applicable procedures for testing the statistical significance of a difference between two groups, something we often have to do in dialectology, as well as one of the standard ways of measuring a correlation between two sets of values.

Tests of differences between groups

Let's begin with tests of differences between groups, e.g. Canadians v. Americans. In comparing groups, the test you use depends on the kind of data you are dealing with. If each participant is represented by one piece of data, such as a response to a single question (e.g., 'yes' or 'no'), or their use of a particular variant on a given occasion (presence or absence of something), then each group of participants will be characterized by the **proportion** (or ratio) of participants in that group who gave the response in question, e.g., 15 out of 36 Canadians (42%) said 'yes', v. 36 out of 43 Americans (84%). This type of difference is tested with a ***chi*-squared test**. On the other hand, if each participant is represented by a series of data (answers to several questions, or multiple observations of a variable), then individual participants will be represented by scalar values, like percents or index scores, and each group of participants is normally characterized by a **mean** (or average) of those values: e.g., the mean percentage of Canadian variants, the mean score on some index you have developed, or the mean frequency of an acoustic formant for the members of that group. Means are calculated by dividing the sum of all the values in a group by the number of values in the group. A difference between group means is evaluated with a ***t*-test**.

Both tests produce a measure of the **probability** (p) that the null hypothesis is confirmed, and that the observed difference could have arisen by chance. In most dialectological research, the probability level for rejecting the null hypothesis is taken to be 0.05, or 5%. This means that you are 95% confident that any difference you have observed is *not* simply the result of random fluctuation in the data, or to the operation of some other factor you haven't considered. If $p \leq 0.05$, the difference can be treated as *significant*; if $p > 0.05$, it must be treated as *not significant*. With a small sample, a probability of up to 0.10 may be taken as marginally significant, and indicative of a trend that might be confirmed by a larger sample in future research.

The *chi*-squared test

1) Construct a table of **observed values** (numbers, not percentages). To take the above example, let's say you have observed that 15 out of 36 Canadians (42%) replied 'yes' to a question, compared with 36 out of 43 Americans (84%). Calculate totals for each column and row and a grand total for the whole table.

	yes	*no*	*Total*
Canadians	15	21	36
Americans	36	7	43
Total	51	28	79

2) Construct a table of **expected values** ('expected' in terms of the *null hypothesis* that there is no difference between the groups, so that they display the same proportions in their data). For each cell in the table of observed values, multiply the row total by the column total and divide the product by the grand total. E.g., for the upper-left cell, this would be: (36 x 51) ÷ 79 = 23.2.

	yes	*no*
females	23.2	12.8
males	27.8	15.2

If nationality had no effect on saying 'yes' or 'no', these are the numbers we would *expect* to observe, given the overall frequency of 'yes' responses (51/79, or 65%) and the proportion of Canadians and Americans in the data (36:43). You now need to measure how far your *observed* values diverge from this expected result.

3) Calculate *chi*-squared (χ^2). For each cell, subtract the expected value from the observed value, and square the difference (multiply it by itself). Then divide the product by the expected value. E.g., for the upper-left cell, this would be: 15 - 23.2 = -8.2; -8.2 x -8.2 = 67.2; 67.2 ÷ 23.2 = 2.9. The sum of all of the values obtained in this way is the *chi*-squared. In the table below, $\chi^2 = 15.1$.

	yes	*no*
females	2.9	5.3
males	2.4	4.5

4) Determine the degrees of freedom (d.f.). This is the number of rows in your basic table (without the totals) minus 1, multiplied by the number of columns minus 1. For a 2x2 table like this one, (2 - 1) x (2 - 1) = 1 x 1 = 1 d.f. If your independent variable had 3 groups instead of 2, you would have 2 d.f. 3 groups with 3 possible responses would be 4 d.f., etc.: you can expand the table with as many columns and rows as you need, following the same procedures to create expected values, compare them to observed values, and calculate *chi*-squared and d.f. Of course, in a larger table, say with 4 region columns, the test will only tell you whether region has a general effect on the variation, not which particular regional distinctions are significant. For that, you need to do 2x2 tests.

5) Look up your *chi*-squared value in a table of *chi*-squared values like the one at the end of this handout, using the row of the table corresponding to your degrees of freedom. In reporting the result of the test, state the probability level, followed by the *chi*-squared value and the degrees of freedom in parentheses, e.g. $p < 0.05$ ($\chi^2 = 4.1$ at 1 d.f.).

A *chi*-squared test is relatively easy to perform with a calculator, or even by hand. If you plan to do several tests, it will be more efficient to set the test up in Excel. Construct a table for observed values, then use formulas and cell addresses to calculate expected values and *chi*-squared as above. This way, each test takes you only a few seconds: just enter the new data into the observed values table and *chi*-squared is calculated instantly.

The *t*-test

The *t*-test requires considerably more calculation than the *chi*-squared test and is most efficiently performed with a computer. In Excel, arrange your data in columns containing all of the individual values on which each group mean is based. To do the test, select TTEST from among the statistical formulas under the INSERT (FUNCTION) menu. Array 1 is the range (column) of observed values for the first group; Array 2 the range for the second group. Select the range using the mouse, or type it in using cell addresses, e.g. A2:A19 for the first range and B2:B19 for the second. You also have to select the number of tails and the type of test to perform. Tails refers to whether you want to test a difference in one direction or both. The distinction is not too crucial at this level, but as a default you can choose two-tailed. Most tests will be Type 1 or 3. Type 1 is used when you have pairs of values relating to the same person, like if you are testing a difference between two conditions, such as the frequency of [in'] in casual speech v. reading style for each person, or between two measures, like the first formant value for /o/ v. /oh/; in this case, one column of numbers is the casual speech scores or the F1 of /o/ and the other is the reading scores or the F1 of /oh/. Type 3 is used when you have two separate groups of individuals, such as men v. women, which have unequal variance (the amount of inter-speaker variation in each group may not be the same). When you enter the TTEST formula (e.g, =TTEST(A2:A19,B2:B19,2,3), Excel returns the exact probability (p) that the difference between the two means is due to chance. This is what you report in your paper: e.g., "the difference between Canadians and Americans was found to be significant at $p = 0.027$," or "the difference between men and women was found to be non-significant ($p > 0.05$)." Make sure you format the cell containing the probability value as a number with 3 decimal places, otherwise you may get an exponential value that is difficult to interpret.

Other tests (multivariate analysis)

T-tests can only compare two means at a time. If you have more than two groups, or more than one dependent variable, you cannot do a *t*-test, except by comparing two groups and one variable at a time. Statisticians do not like to do this too much, since each time you do a *t*-test, there is a chance that the test is inaccurate, so the more tests you do, the greater the chance becomes that one of them will give a false result, suggesting a difference is significant when it really isn't. In this course, you don't need to worry too much about this problem, but if you are comparing several groups and several variables at once, the correct approach would be to use a multivariate test, like an ANOVA (Analysis of Variance), which is much more complicated and has to be performed by a computer. ANOVAs and other multivariate tests can be done with SPSS and other statistical analysis programs, some of which may be available on the computers in the Arts computing lab. Sociolinguists have developed a multivariate analysis program called *Varbrul*, later called *Goldvarb*, which can analyze the effect of an unlimited number of multi-valued independent variables (e.g., age, sex and social class of speaker plus phonological environment of token) on a single, binomial (two-valued) dependent variable (application or non-application of a rule, like /t,d/-deletion, or occurrence of a variant, like 'be like' as a verb of quotation), using stepwise regression. This multiple regression allows you to determine which of your several independent variables has a significant effect on the variation, as well as the quantitative effect of each factor. More recently, many sociolinguists have started using a similar but more flexible multivariate analysis program called *Rbrul*, which overcomes several limitations of *Varbrul*. Both programs can be downloaded for free from the Web and are relatively simple to use, with manuals also available for download. A good book on how to do multivariate analysis in sociolinguistics is *Analysing Sociolinguistic Variation*, by Sali Tagliamonte (Cambridge University Press, 2006). You are not required to do multivariate analysis in this course, but if you wish to try, or if you've had a course in statistics and already know how, you are encouraged to do so.

Tests of correlations between sets of numbers

If you are looking for a **correlation** between two sets of scalar numbers, such as between participants' ages and their index scores, you need to turn to a different set of tests. One easy way to look for a correlation is to construct a graph of the independent variable against the dependent variable and plot the series of paired measures, then look to see whether a pattern appears. You can do this with the Chart Wizard in Excel, plotting one range of values on the X axis and the other on the Y axis of an XY scatter chart. You can then look for a correlation by adding a trend line under the CHART menu: this is a regression line that shows you whether the values on the X axis are correlated with the values on the Y axis. A flat line shows no correlation; a diagonal line shows a correlation; the steeper the slope, the stronger the correlation (e.g., as ages go up, index scores go down, an inverse correlation).

Another way to look for a correlation without creating a graph is to use the **Pearson** product moment correlation coefficient (*r*), which is easily calculated in Excel. This is a value between 1 and -1 that expresses the degree of correlation between two sets of numbers. Select PEARSON from the list of statistical formulas under INSERT (FUNCTION). The two arrays are the two sets of numbers you are comparing: e.g., people's ages in one column, with their dependent measures in the next column. An *r* of 0 shows no correlation at all, indicating that age has no predictive value with respect to the dependent measure. An *r* of 1 is a perfect positive correlation (the values in the first column increase with perfect reliability as the values in the second column increase); an *r* of -1 is a perfect negative correlation (the values in the first column increase as the values in the second column decrease). Most *r* values will be somewhere between 0 and either 1 or -1. The farther away from 0, the stronger the correlation. Roughly speaking, an *r* of less than 0.25 (positive or negative) suggests no correlation; an *r* of 0.33 suggests a weak correlation; an *r* of more than 0.5 suggests a fairly strong correlation; and an *r* of more than 0.66 suggests a very strong correlation/ If you want a more precise determination of whether your *r* value is statistically significant, you can use a table of critical values of the correlation coefficient, which you can find on the Web.

A note on decimal places

Many students mistakenly report all of the decimal places calculated by Excel, which implies a level of precision that is not supported by the data being analyzed. Decimals should be rounded off as appropriate in written reports. As a rule of thumb, probabilities should have 2 or 3 places (e.g., $p = 0.037$; $p < 0.01$); *chi*-squared and *t* values 2 or 1; index scores 1 place; and percentages none (rounded to nearest whole number), unless you are distinguishing values under 5%. Also format the axis labels on your graphs in the same way: change 60.00% to 60%.

Table of values for the *chi*-squared test

TABLE IV

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

How to read the *chi*-squared table

In this table, the columns represent probability levels (p or α), and the rows represent degrees of freedom (d.f.). Note that the *chi*-squared values get bigger from left to right and top to bottom, while the probability levels get smaller from left to right (starting with 0.99, virtual certainty, on the left, and going to 0.01, a 1% chance, on the right; these refer to the likelihood that the null hypothesis is true). Since *chi*-squared is produced by differences between the observed and expected values (per the formula explained above), the larger those differences, the larger the *chi*-squared value, and the more likely the difference is to reach statistical significance. To check the significance level of your *chi*-squared value, first find the row corresponding to your degrees of freedom. Note that larger degrees of freedom (more complex data sets) require larger *chi*-squared values to attain the same level of significance, as you go down each column. Look along the appropriate row until you get to the number in the column that corresponds to $p = 0.05$ (third from the right). If your *chi*-squared value is smaller than this number, your difference is not significant. If it is larger, the difference is significant at $p < 0.05$. To determine the exact level of significance, continue looking to the right until you find a number larger than yours. When you do, go back one column to the left and state your p as being less than the level at the top of this column. For example, at 1 d.f. (for a simple 2x2 table), a *chi*-squared of 3.841 is significant at $p = 0.05$. If your value is larger than this, your difference is significant at $p < 0.05$. If it is larger than 5.024 ($p = 0.025$) but not larger than 6.635 ($p = 0.01$), it is significant at $p < 0.025$. If it is off the chart, use the smallest probability given (here, $p < 0.01$).