vitC.sas: Explanation of code

Goals of code:

- Reading a contingency table in “count” form
- Contingency table tests, proportions, and measures of association
- Reading event data in “individual” form
- Exact tests for general contingency tables (more than 2x2).

Note: There are many different methods to analyze contingency table data. Everything we’ve done previously in the course has had alternative methods. Contingency table data has a lot of alternatives. The answers are usually only slightly different, especially with moderate to large sample sizes. I show how to get large sample results and one of the small sample (“exact”) results.

Reading “count form” data: input row $ col $ n;
When the counts in a contingency table are pre-tabulated, the easiest way to do stats on that table is to enter the table with one row of data for each cell (combination of row and column) in the table. The key piece of information is the number of observations in that cell. Here we store that in the \textit{n} variable. The row and column identifiers can be stored as character or numeric variables. \textit{n} must be numeric.

The table can be stored in a data file or, as here, entered using \texttt{cards}; or \texttt{datalines};.

Contingency table computations: proc freq;
The basic command in proc freq is \texttt{table}. This counts the numbers of observations for each level of the specified variables. \texttt{table var1}; will classify observations by their value of \texttt{var1}. \texttt{table var1*var2}; will classify observations by each combination of \texttt{var1} and \texttt{var2}. You can continue adding \texttt{*var} terms to classify observations by levels of 3 variables (a 3-way contingency table), 4 variables, \ldots. These instructions focus on 2-way contingency tables.

For a 2-way contingency table, the first var specifies the rows; the second specifies the columns. I find it more useful to put “treatments” as rows and “responses” as columns.

For the Vit C data set, we want a 2-way contingency table with cells defined by combinations of \texttt{trt} (what treatment a person was assigned to) and \texttt{cold} (whether or not they got a cold that winter). Because cold is the response, we want that variable to define columns and \texttt{trt} to define the rows. Hence, \texttt{table trt*cold}.

Describing “count” data:weight n;
By default, proc freq will count the number of observations in each cell. That is demonstrated at the end of the example code, where we have “individual” data. When the data are already summarized,
we need to tell SAS that each row represents multiple observations. That is what weight n; does. Tells SAS that the variable n contains the number of individuals associated with that row of data.

Aside: You can multiple rows, each representing multiple individuals. For example, if the vit C study had been done in multiple cities, your data might be four rows (one for each combination of trt and cold) for each city. table trt*cold; weight n; will then report and analyze the total summed over all cities. This form of data would be very useful because you could use it to analyze the total for all cities, or analyze each city separately (e.g., by adding a by city; to the proc freq).

**Controlling output from proc freq:** table statement options
By default, proc freq reports multiple numbers for each group. For a 2-way table, SAS wants to report the count, the percentage of the row total, the percentage of the column total, and the percentage of the overall total. This is too much for me. I usually only want the count.

To suppress the extra output, I add the options norow nocol nopercent. All go after a / that separates the core table statement from its options. You can mix and match these options. norow suppresses the row percentage, nocol suppresses the column percentage, and nopercent suppresses the overall percentage. Later in the code, the row proportion is useful, so I omit norow to see the row percentage but not the other two.

**Chi-square statistic:** chisq; 9option
The chisq option requests the Chi-square test of no association. This is reported in the Statistics for Table of trt by cold piece of output. The Chi-square statistic is the first line. SAS also reports other statistics used in other fields. If you know about the g-statistic, that is called the Likelihood Ratio Chi-square, which is a more appropriate name than g. The Continuity Adj. Chi-square is the Chi-square statistic with a continuity correction (discussed briefly in class). The other statistics are not relevant (for us, now).

When the 2-way table is a 2x2 table, SAS reports the Fisher exact test automatically. The output is in the Fisher’s exact test block of output. The line you want is labelled Two-sided Pr <= P. This is the two-sided p-value that is the customary result from Fisher’s test. The other numbers are intermediate calculations that will be of interest only if you want a non-customary analysis.

**Odds ratio and relative risk statistics:** measures option
When you add the measures option to the table statement, you get two more blocks of output: one without a title and one labeled Estimates of the Relative Risk (Row1/Row2). The untitled block contains quite a few statistics that are commonly used in the social sciences. If you recognize and want one, here is where you find it.

We are interested in the Estimates of the Relative Risk (Row1/Row2) output. The first row is the Odds ratio. Note this is the odds ratio, not the log odds ratio. The odds are calculated as odds of column 1 in row 1 to odds of column 1 in row 2. If you want odds of the event in column 2, you need the reciprocal of the number reported here. The second and third rows are the relative risk, calculated as the risk for row 1 / risk for row 2. The second row calculates the risk of column 1; the third row calculates risk of column 2.
Confidence intervals for all measures: cl option
If you both the cl and measures options, you get confidence intervals for each measure. These are large sample confidence intervals. Again, the only ones needed for this class are the ci’s for the odds ratio and relative risk(s). All three are calculated on the log measure and back transformed.

Proportions for each row and the difference in proportions: riskdiff
This calculation is very frequent in epidemiology, so the proportion gets called “risk”. That is the number of events / sample size. SAS computes the proportions and their difference when you add the riskdiff option to the tables statement.

The output is in two tables. One is labeled “Column 1 risk estimates”; the other “Column 2” · · · For the statistics in the first table, the number of events is the number in the first column. Look at how the counts are printed. For the VitC data set, the first column is Cold = 'No', so the statistics in the first table are for the number of folks who didn’t get a cold divided by the total number in each group (VitC or placebo). The statistics in the second table are for the proportion of folks who did get a cold. You choose which set is more appropriate for your study.

Notice that rows and columns are not interchangeable when we talk about risk. If you write the table statement the other way (e.g. table cold*trt), a row is the number of people getting a cold (or not) and the columns are the treatment. Hence, the Column 1 risk is the number of folks in the placebo treatment who got a cold divided by the number of folks who got a cold. Same numerator, different denominator! Doesn’t affect the Chi-square statistic or the odds ratio. Different number and very different interpretation for proportion or risk.

Each table of results (Column 1 or Column 2) has 6 numbers:

- Risk: The estimated proportion, note discussion above.
- ASE: Asymptotic Standard Error: the se of the proportion, computed using the method discussed in lecture. This is based on theory for large samples, hence the Asymptotic in the name.
- Asymptotic Conf. limits: These are the confidence intervals discussed in lecture. They are calculated as estimate ± Z quantile times se. These endpoints are based on theory for large samples, hence the Asymptotic in the name.
- Exact Conf. limits: CI’s that have better statistical properties for small samples. Calculated in one of many possible ways (details relevant in an Epidemiology course or Stat 457 (counts and proportions), not relevant in 401).

There is a row of results for each group, a row (labeled total) for the total number of events, and a row for the difference. Notice that there is no exact ci for the difference. There is a lot of disagreement on how best to define such a ci, so SAS doesn’t report anything.
The CI’s are 95% by default. You can add \( \alpha = \) to change that. As elsewhere, \( \alpha = 0.10 \) gives you 90% ci’s, and \( \alpha = 0.01 \) gives you 99% ci’s, but you can specify any number between 0 and 1 (e.g., if you want an 87% ci for some reason).

**Confidence intervals for each proportion: table cold / binomial**

If you want more control over what defines each group and what event is reported, a second way to get confidence intervals for each proportion is to analyze the data by group and specify exactly what you want. This is what the `proc sort; proc freq; by trt` block of code does. The table of counts is constructed from one variable, `cold`, but the analysis is done separately for each `trt`. Without the option, this would give the same information we got in the 2-way table (`table trt*cold`).

The `/binomial (level=)` option requests binomial confidence intervals, with the event of interest defined by the `level=` option to the option. Here we are interested in the probability of getting a cold, so `level='Yes'` is the appropriate event.

Note: the comparison can be between numbers (e.g., `level=0`) or character strings (e.g., `level='yes'`). If you use character strings, they must match exactly. Remember that “Yes” is not the same string as “yes” or “ Yes”.

**Contingency tables from individual observations:**

SAS can calculate the counts from individual level data, where each row represents one individual. You need their group and their output. Because each individual is adds one to the total count, you do not need any \( n \) variable. The “madeup” data set is a small data set with two factors producing a 2 x 3 table. The data are read from the SAS program. The `proc freq` code is identical to the previous code with one exception: there is no `weight` statement. SAS will construct and report the table of counts, then do the additional calculations you request.

**Exact tests with tables larger than 2x2: exact chisq**

SAS reports the Fisher exact test by default when the contingency table is 2x2. When the table is larger, the Fisher test isn’t defined, but randomization can still be used to get an exact p-value. The exact test would be more appropriate when sample sizes are small and necessary when sample sizes are really small (e.g., most expected counts < 5).

The exact test is obtained by requesting the desired test on the table statement (`chisq` option) and requesting the exact version using `exact chisq;`. If the problem is large and you still want exact intervals, you can get a Monte-Carlo (sampling) estimate of the p-value using `exact chisq / mc;`.

Note: If you want exact ci’s for the odds ratio, relative risk, or risk difference you can request those by `exact or; exact relrisk; or exact riskdiff;`. 

---

4