Ling 236 Homework #3 Solutions

1. It's often fairly easy to take some data and answer a question like 'Use Fisher's exact test to see if gender and promotion rate appear correlated'. It's often harder to take real problems and decide what if any statistical methods one should make use of on them. In this spirit, here's an email from the corpora mailing list from last week. What would you advise this person?

> ...
>
> Currently, I am comparing the frequency data of a learner corpus and native speaker corpus (they have approximately the same size)and have some statistical queries. For exampe: For the verb KEEP, I have got the freqency of the each verb form in the two corpora as follows:
>
> | Learner corpus | Native speaker corpus |
> |---|---|
> | keep 348 (88.5%) | keep 99 (58.2%) |
> | keeps 15 (3.8%) | keeps 14 (8.2%) |
> | keeping 9 (2.3%) | keeping 32 (18.8%) |
> | kept 21 (5.4%) | kept 25 (14.7%) |
> | Total 392 (100%) | Total 170 (99.9%) |
>
> According to the percentage each form takes in its perspective corpus, I can easily see a large differenc between the use of "keep" in learner corpus and that in native speaker corpus (88.5%:58.2%). But one problem to my interpretation is "Why do you think this difference (88.5%:58.2%) is significant and other differences are not?"
>
> ...

Answer: This question was fairly underspecified. Here are some observations one could make. The use of forms of *keep* in the two corpora is certainly very different ($\chi^2_3 = 75.5, p < 0.0001$). This simply shows that the examples probably aren't drawn from the same distribution. If he just wants to test a particular form, one way to do that is to collapse the rest of the table and just test *keep* versus not *keep*. It's not clear quite what the basis is for thinking the difference in probabilities for *keep* is so significant while the others aren't: in proportion terms, *keep* is less than one-and-a-half times more common than the native proportion, while the usage of *keeping* has sunk by more than 7 times. At any rate, all of these are evaluated as a proportion of the total, so the measurements aren't independent.

Most crucially, no data simply counting forms could *explain* why the native and learner rates are so different. This would require the presence of further explanatory variables, and further examination of the data. *If* the use of tense by learners is correct, one can observe that the percentage of past tense in the native corpus is apparently around 3 times higher than in the learner corpus. This is suggestive of genre/style differences between the corpora. One hypothesis is that the learner corpus is mainly present tense and has more first and second person narratives, and the high usage of *keep* is just correct. An alternative theory is that learners are failing to inflect *keep*. Evaluating these alternatives would require data on the person and number of the subject of verbs, and sentence tense information.

2. There's a small subfield of humanities computing that attempts to use statistics from texts for authorship attribution. Commonly this involves looking at the frequencies of fairly common function words or function word sequences, which are assumed to be fairly consistently used

by an author across texts, but on which another author might have idiosyncratically different usage. This problem considers texts of Jane Austen. In particular, when she died, she left a partly finished novel (*Sanditon I*), which was completed by a fan, attempting to write in her style (*Sanditon II*), and the composite was then published. The table gives the relative frequency of the word *a* preceded by and not preceded by *such* (i.e., the latter is the sum of the counts for all other words *X a*, the word *and* followed by or not followed by *I* and the word *the* preceded by or not preceded by *on* in the two halves of *Sanditon* and two other Jane Austen novels. Was Austen consistent in these habits of style from one work to another? Did her imitator successfully copy these aspects of her style?

Word sequence	*Sense and Sensibility*	*Emma*	*Sanditon I*	*Sanditon II*
such a	14	16	8	2
$\neg$*such a*	133	180	93	81
and I	12	14	12	1
and $\neg I$	241	285	139	153
on the	11	6	8	17
$\neg$*on the*	259	265	221	204

Answer: You could actually get different answers here depending on how you did things. (Remember that line about lies, damned lies, ...) If one examined each word pair separately, then doing 2×3 tables just for the Austen novels, none of the different rates are significant (*such/a*: $\chi^2_2 = 0.267$ ($p = 0.8751$); *and/I*: $\chi^2_2 = 2.43$ ($p = 0.2966$); *on/the*: $\chi^2_2 = 1.553$ ($p = 0.4600$)). If one then extends this to *Sanditon II* and uses 2×4 tables, for each word, then one finds that the rates of use of *and/I* and *on/the* ***are*** significantly different at the $p = 0.05$ level (*such/a*: $\chi^2_3 = 4.067$ ($p = 0.2543$); *and/I*: $\chi^2_3 = 9.44$ ($p = 0.0239$); *on/the*: $\chi^2_3 = 9.564$ ($p = 0.0226$)). Note in particular that the higher total usage rates for these two words serves to make the result significant, even though the ratio of use of *such/a* is higher between *Sanditon II* and other novels than for *on/the*. So, the conclusion on this analysis is that the word usage of *Sanditon II* is somewhat different from the other Austen writings in two respects. This might be evidence of imperfectly copying style (or might be simply the different subject matter of the part of the novel).

If, howevever, one considers all the columns together, one is effectively extracting more statistical power by considering the aggregate data distribution, and looking for the constancy or not across all cells at once. Here, though, this seems to complicate things rather than giving a straightforward result. Doing a 2×6 table, one finds that *Emma* and *Sense and Sensibility* do not differ significantly ($\chi^2_5 = 6.17$ ($p = 0.30$)), but *Sanditon I* differs from them (*and* not being followed by *I* less frequently and *the* not being preceded by *on* more frequently) ($\chi^2_{10} = 23.29$ ($p = 0.01$)). *Sanditon I* and *II* are not consistent ($\chi^2_5 = 17.77$ ($p \leq 0.01$)), largely due to the different incidences of *and* followed by *I*. The conclusion on this analysis is that *Sanditon I* is inconsistent in word usage with *Sanditon II*, but that Austen's own novels also deviate in word usage patterns.

The underlying assumption of this kind of analysis is that patterns of function word usage should stay reasonably constant, and betray the identity of an author by exhibiting their characteristic style. This analysis has sometimes been quite successful (recent e.g.: identifying the author of *Primary Colors*). However, it's not always clear that the assumption is true: if the author can manipulate their own style of writing, or for something like *and/I*, just

whether they change between writing in the first or third person, then these measures could be quite inconsistent for other reasons.

3. In Wasow (1997), in the section on Collocations and HNPS, one finds the following data (graphed in Figure 12 of the paper):

	HNPS	not	
Transparent collocations	90	102	192
Non-collocations	59	329	388

(i) Confirm the figure given in the paper for the chi-square test. (ii) Calculate the odds ratio for heavy NP shift for transparent collocations versus non-collocations. That is, how many times larger are the odds of HNPS with a transparent collocation than with a non-collocation?

Answer: Given the data in this table (which I believe is right given the article), the chi-square value is 67.48. This deviates fractionally from the reported value of 68.08, for reasons that are unclear (the counts were changed slightly after the computation of this chi-square result? or arithmetic error?). The odds for HNPS with a transparent collocation are 90/102 (roughly, 1:1). The odds for HNPS with a non-collocation are 59/329 (roughly, 1:5). Thus the odds ratio is 4.92. That is, the odds of HNPS with a transparent collocation are 5 times greater than the odds of HNPS with a non-collocation.