Ling 236 Homework #3

Due 30 January 2002

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?

   From: "xiaotian guo" ⟨xiaotiang@hotmail.com⟩
   Sender: owner-corpora@lists.uib.no
   To: corpora@hd.uib.no
   Subject: Corpora: statistics in learner English
   Date: Thu, 17 Jan 2002 23:06:10 +0800

   Dear All

   First, let me thank all those who replied to me concerning my request about "overuse and underuse of learner English" a couple weeks ago.

   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?" I would think there is no way to answer this question by means of some statistic help because it really depends on individual circumstances and it will be difficult if not possible to give a demarcation to such kind of comparison. But to make sure about this point, I would like to raise this question to the list members.

   Someone suggested "chi sqare" to me. But after some initial reading, I found it can only review the relationship between the observed frequency and expected frequency and it is based on null hypothesis. It can only tell me whether there is a significant difference as a whole rather than individually concerning the use of the different forms of KEEP in the two corpora. It seems it cannot answer the question I have: why do you think the use of the base form "keep" is significantly different?

Another query is that if I forget about the problem I just raised and try to detect differences in two corpora as a whole, what is the best statistic mothod to use? Oakes pointed out the weakness of Chi-sqare in Statistics in Corpus Linguistics:

The Chi-square test is used for the comparison of frequency data. Kilgariff has shown that this test should be modified when working with corpus data, since the null hypothesis is always rejected when working with high-frequency words.

I wonder whether there is another test which could help with corpora comparison.

With thanks

Guo Xiaotian

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

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?