

## Statistics 406 Problem Set 10

Due Tuesday December 12th

1. In this problem you will use simulation to evaluate the accuracy of the approximate value for the sampling standard deviation of the log odds ratio, given at the top of page 9 in the “Binary response data” notes.

- (a) Construct three  $2 \times 2$  table of the form

$$\begin{array}{cc} n_{11} & n_{10} \\ n_{01} & n_{00} \end{array}$$

One of them should have log odds ratio 0, one should have log odds ratio between 0.5 and 1, and the third table should have log odds ratio between -1.5 and -1.

- (b) Use simulation to assess the coverage probability of the 95% CI for the population log odds ratio based on the approximate value for the sampling standard deviation of the log odds ratio (from page 9 of the notes). Consider all three tables that you constructed in part (a), and vary the total sample size from 20 to 100 in increments of 10 (change the sample size by scaling the tables you found in part (a), rounding to get integer sample sizes).
2. Suppose we have  $n$  timecourses of length  $T$ :  $X_{it}$ ,  $i = 1, \dots, n$  and  $t = 1, \dots, T$  (like the temperature record data). We wish to determine whether any 20 year period was unusual, in having an exceptionally large number of values at the high end of the overall range.

To do this, we will follow these steps (where  $X$  denotes the raw data):

- (a) Lowess smooth each timecourse with a smoothing parameter  $f$ .
- (b) Standardize each of the smoothed timecourses.
- (c) Consider each 20 year period of time, and count the total number of points above 1.5 (this count would cover all 20 timecourses over the 20 year period, so the maximum would be 400). Store the maximum value.
- (d) “Cyclically permute” each raw (not Lowess smoothed) timecourse. That is, pick a random year, shift that year to the beginning, then take the data that got shifted off the left end and stick it back on at the right end. This is easier to see by example. Suppose the data for one timecourse are

A B C D E F G H I J

and we pick “5” as our random year. This would yield the following:

E F G H I J A B C D

Do this independently for each timecourse.

- (e) Lowess smooth (with parameter  $f$ ) and standardize each cyclically permuted timecourse.
- (f) In the permuted data, count the number of points above 1.5 in each 20 year window. Store the maximum value.
- (g) Repeat d-f 1000 times. Calculate the proportion of the time that the result exceeds the result for the actual data calculated in step 3.

You don't need to carry out a full simulation study here (that would take too long to run). You should run the program several times, varying these things:

- As the input data, use either iid standard normal data, or iid standard normal data in which half of the series between year 100 and 110 have been shifted by  $\delta$ . Consider various values of  $\delta$  to get a feel for what level of signal is generally detectable.
- Use different settings for the smoothing parameter  $f$  (e.g. 0, 0.01, 0.1, 0.5).

This is a complicated problem, we will also be discussing it in class.