Department of Statistics

COURSE STATS 330

Assignment 5, 2003

Instructions: Hand in your completed assignment to the Student Resource Centre by 4pm on Thursday 23th October. Your answers to this assignment are **not to be given in the form of a report.** Just answer the questions below. Set you answer out in 4 sections corresponding to the 4 questions.

The data for this assignment relate to a famous literary controversy. In 1861, a series of 10 letters appeared in a New Orleans newspaper, the *New Orleans Daily Crescent.* They were signed with the nom-de-plume "Quintius Curtius Snodgrass". Many literary scholars have attributed the authorship of the letters to the famous American author Mark Twain. We are going to analyse the word length distributions of the letters. Table 1 below classifies the words in the 10 Snodgrass letters according to their length. The letters are also divided into three groups: the first three letters, the next three, and finally the last four.

Table 2 contains similar data from known writings of Mark Twain, taken from letters he is known to have written, and also from two of his books, *Roughing It* and *Following the Equator*. The idea is to compare the word-length distributions of the Snodgrass material with that of the Twain material. If they are very different, then the Twain authorship of the Snodgrass letters becomes more problematic.

Note: The data are supplied in the form of two files, **`twain1.txt`** and **`twain2.txt`**, corresponding to Tables 1 and 2. You may need to do some rearranging of the data to get it suitable for input to the glm function.

Table 1: Words in the Snodgrass letters classified by word length.

	Number of words		
Number of letters in word	*First 3 Snodgrass letters*	*Next 3 Snodgrass letters*	*Last 4 Snodgrass letters*
2	997	831	857
3	1026	828	898
4	856	669	777
5	565	420	446
6	366	326	300
7	318	293	285
8	258	183	197
9	186	150	129
10	96	94	86
11	63	49	40
12	42	30	29
13+	25	25	11

Table 2: Words in known Twain writings classified by word length

	Number of Words				
Length	*Two letters from 1858 and 1861*	*Four letters from 1863*	*Letter from 1867*	*Sample from "Roughing It", 1872*	*Sample from "Following the Equator", 1897*
2	349	1146	496	532	466
3	456	1394	673	741	653
4	374	1177	565	591	517
5	212	661	381	357	343
6	127	442	249	258	207
7	107	367	185	215	152
8	84	231	125	150	103
9	45	181	94	83	92
10	27	109	51	55	45
11	13	50	23	30	18
12	8	24	8	10	12
13+	9	12	8	9	9

Question 1.
For the analysis proposed above to be meaningful, the word count distributions in different writings by Twain must be identical. Compare the word count distributions in the 5 works in Table 2, using both suitable graphs and a well-chosen model. Is it reasonable to assume that Twain's writing can be characterized by a single word length distribution?

Question 2.
On the basis of word length distributions, is it likely that the three groups of Snodgrass letters are written by the same person?

Question 3.
Did Mark Twain write the Snodgrass letters?

Question 4.
(Challenge question) Is it reasonable to assume that the word length distribution has a Poisson distribution? In addressing this question, you will have to get around the problem of zero length words, and the fact that 1-letter words have been excluded. If the Poisson distribution is not correct, can you suggest an alternative? Does the possible "non-Poissonness" of the data invalidate your conclusions in the first part of the assignment?

Hint for question 4:
Try a "truncated Poisson" – this is a distribution that is Poisson, but can't have values 0 or 1. It is rescaled so that all the probabilities add to 1. You can calculate the probability of a count being `y` in R using `dpois(y,mu)/(1- ppois(0:1,mu))`