Homework 7

Software Design **Allen B. Downey**
Fall 2004

Due: Monday 18 October

7.1 Statistical Analysis

1. Download the following files from the class web page:

   ```
   wget http://wb/sd/code/respnd95.asc.gz
   wget http://wb/sd/code/interv95.asc
   wget http://wb/sd/code/birth.py
   ```

   The first file is a database collected by the National Institutes of Health (NIH) as part of their National Survey of Family Growth (NSFG). It contains one (very long) line of data for each of the 10,847 respondents who participated in the study. The second file is an "interval" file that contains one line of data for each of 21,332 pregnancies reported by the respondents.

   Note: this database contains real data collected from people who volunteered to provide an enormous amount of personal information, including information about their reproductive and sexual histories. Some of the information in this database, and some of the data we will be looking at, is both politically sensitive and emotionally charged. I think this work will be interesting, and I hope we will find some interesting things, but it is important to approach this material with appropriate discretion and sensitivity.

2. `birth.py` contains a program that reads the two data files and builds one Respondent object for each line in the respondent file and one Interval object for each line in the interval file. Read the program carefully and make sure you understand what it does and how it works.

3. Design and implement a data structure that will allow you to represent sets and subsets of respondents, and that makes it easy to find the intervals (pregnancies) associated with a given respondent.

4. Before you work with a database like this, it is a good idea to do some consistency checking. First, add code that checks whether the number of intervals for a given respondent matches the `pregnum` attribute. Then add code that checks whether the Interval objects are in the right order, according to the `pregordr` attribute. If either check fails, raise an appropriate exception.

   This kind of checking serves several roles: you are checking the database for missing or incorrect information; you are checking whether you understand the contents of the database; and you are checking whether your program is handling the data correctly.

   Note: for some reason the people who designed the data files only allocated two entries to record the sexes of children that result from a single pregnancy. In cases of triplets and higher-order births, we only know the sex of the first two children. The attribute `nbrnlv` is the number of live births for a given pregnancy.

5. Once you have organized the database into a structure conducive for analysis, we have to decide what questions to investigate. For this assignment, we will investigate how the number of children a woman has affects her intention to have more children, and whether the sex of the children she has affects her intentions.

   For this kind of inquiry, a common approach is to choose a set of variables that we think measure the cause of a phenomenon (so-called independent variables) and a set of variables that we think measure an effect (the dependent variables).

   In this study, the independent variables are the number of male and female children and possibly the birth order of male and female children. The dependent variables are the answers to questions about whether the mother wants more children, whether she intends to have more children, and how many more children she intends to have.

   The database we are working with has a HUGE amount of information about the respondents. I have made a pass through the variable descriptions and chosen some of the variables we might want to use. If you are interested, or if there are additional variables you want to investigate, I can show you the survey documentation.

   The variables I selected are explained in the handout entitled "Section G: Birth Expectations and Desired Family Size."

6. As a warmup exercise, write a program that computes the histogram of family sizes; that is, the number of respondents that have given birth to n children, for each value of n. Remember that we are counting the number of live births for each respondent, which may not be the number of children currently in the household of the respondent, for a variety of reasons.

7. Now that you have an idea where we are going, start to think about some of the functions that you are likely to want, and some of the data structures you will want to build. For example, it seems likely that we will want to find the subset of respondents that satisfy various criteria, and build data structures to represent those subsets. That suggests that we will want functions that evaluate various criteria for respondents (similar to the filters we applied to words in the dictionary) and functions that form subsets of the respondents.

   If you have some ideas about how to design these functions, you should do some design. If you are stumped, you should start writing some code. Start with specific code that checks for specific criteria, and look for ways to generalize it.

8. At the very least, you should write a program that answers this question: "Are respondents who have not had a male child more likely to want additional children than respondents that have had a male child?" Beyond that, I will leave it up to you to think about which groups of respondents you want to investigate, and which dependent variables you want to look at.

   This is an open-ended assignment because I think there are a lot of ways to proceed and I want to give you a chance to explore. If you feel overwhelmed, or have trouble getting started, I would be happy to make suggestions.

   I have explored this dataset a little bit and seen some results that I think are interesting, but I don't have all the answers, or even very many. To the best of my knowledge, no one else has used this database to investigate these issues. If we find something here, we will be the first to find it. Welcome to the frontier of current knowledge.

9. WHAT TO TURN IN: As always, you should make your printed code as presentable as possible before you turn it in.

In addition, I would like you to write a short paper (1-2 pages) that presents the analysis you did and the results you found. You can assume that the reader is familiar with the database (for example, you can refer to variables by name), but you should try to present your results clearly and concisely, using tables and graphs where appropriate.

In some cases, you may see apparent differences between groups that are relatively small, and it may not be clear whether the effect is statistically significant. I don't expect you to do significance testing, but you should be careful not to overstate your confidence in the findings.