

Prof. David Draper  
 Department of  
 Applied Mathematics and Statistics  
 University of California, Santa Cruz

### AMS 7: Homework 4

Due date: Tue 6 Mar 2007 in class [100 total points]

1. (biology and health [20 points]) Millis and Seng (1954) reported results from a study on the relation of birth order to the birth weight of human infants. The table below records frequency distributions they obtained of the birth weights of all first-born and eighth-born male infants of Chinese patients at the Kandang Kerbau Maternity Hospital in Singapore in 1950 and 1951:

Birth weight (lb: oz)	Birth order	
	1	8
3:0—3:7	—	—
3:8—3:15	2	—
4:0—4:7	3	—
4:8—4:15	7	4
5:0—5:7	111	5
5:8—5:15	267	19
6:0—6:7	457	52
6:8—6:15	485	55
7:0—7:7	363	61
7:8—7:15	162	48
8:0—8:7	64	39
8:8—8:15	6	19
9:0—9:7	5	4
9:8—9:15	—	—
10:0—10:7	—	1
10:8—10:15	—	—
	1932	307

In other words, there were 1,932 first-born male infants in the overall sample, of whom 2 had birth weights between 3 pounds 8 ounces and 3 pounds 15 ounces, and there were 307 eighth-born male infants in the sample, of whom 1 weighed between 10 pounds 0 ounces and 10 pounds 7 ounces (and so on). By placing all of the infants in each frequency category at the center of the category (which is not quite right, but it will do for our purposes here), I've worked out that the mean and SD of the birth weights of the first-born infants (expressed in ounces) were approximately 106.1 and 12.2, respectively, and the corresponding values for the eighth-born infants were 114.7 and 15.0.

- (a) Does this difference seem large to you in practical terms? (In addition to making our usual percentage difference calculation, you could, for example, consider this situation

from the point of view of the mothers who gave birth to these babies.) Explain briefly. [5 points]

- (b) Set up a statistical model for this situation, being explicit about the population, sample and imaginary data sets, and use your model (including the usual inferential summary) to build a 95% confidence interval for the population mean difference in birth weight (what *is* the population here, precisely?). Is this difference large in statistical terms? Explain briefly. [15 points]

2. (environmental studies [30 points]) A plant ecologist performs an exhaustive search of an approximately 40-square-mile area, chosen to be representative of the growing region of a particular rare species of tree, and finds 101 trees of this species in her search. She records for each tree whether or not it's rooted in serpentine soils (serpentine is a magnesium-rich silicate mineral) and whether its leaves are pubescent (hairy) or smooth, with the following results: of the 35 trees in serpentine soils, 12 had pubescent leaves, and of the 66 trees in non-serpentine soils, 16 had pubescent leaves.

- (a) Does the difference in percentage of pubescent leaves between the trees in serpentine and non-serpentine soils seem large to you in practical terms? Explain briefly. [5 points]
- (b) Set up a statistical model for this situation, being explicit about the population, sample and imaginary data sets, and use your model (including the usual inferential summary) to build a 95% confidence interval for the difference ( $p_1 - p_2$ ) in population percentages of pubescent leaves (what *is* the population here, precisely?). Is this difference large in statistical terms? Explain briefly. [15 points]
- (c) Your answers in (a) and (b) should have indicated to you that the plant ecologist found a difference that was practically but not statistically significant, which (as usual) means that she didn't get enough data. Suppose that she decides to regard the investigation summarized above as part one of a two-part study: based on what she's found in part one, she'll work out how much more data she really needs and then get the rest of the required data in part two. Here is how she might reason, using the confidence interval approach to sample size determination.

In the overall study (parts one and two combined) she's planning to build a 95% confidence interval for  $(p_1 - p_2)$ , where (say) 1 stands for serpentine soil and 2 for non-serpentine. Using the ideas we developed in class, this confidence interval will be of the form

$$(\hat{p}_1 - \hat{p}_2) \pm 2 \widehat{SE}(\hat{p}_1 - \hat{p}_2) = (\hat{p}_1 - \hat{p}_1) \pm 2 \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}, \quad (1)$$

where  $\hat{p}_1$  and  $\hat{p}_2$  are the sample proportions of pubescent leaves from the serpentine-soil and non-serpentine-soil trees, respectively, and  $n_1$  and  $n_2$  are the numbers of such trees in the overall study. If this interval is to be narrow enough that 0 is just barely not inside it (which is what she would need to demonstrate statistical significance), then it would have to look like the sketch at the top of the next page.

