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

STAT 7: Discussion Section 1

1. A central theme of this course is *quantification*: how to measure things that interest you, and what to do with the measurements once you have them. This problem poses some questions on how to measure variables that arise in biology, environmental science, health, and other fields.

(a) The number of deaths from cancer in the United States has risen steadily over time. In 1985, for example, about 462,000 people died of cancer, up from 331,000 deaths in 1970. In a speech on the floor of the House of Representatives in 1986, a member of Congress said that "these numbers show that no progress has been made in treating cancer." Explain how the number of people dying of cancer could increase even if treatment of the disease were improving, and describe at least one variable that would be a more appropriate measure of the effectiveness of medical treatment for a potentially fatal disease.

(b) Cost is often an important factor in deciding which variable to measure in a study. Sometimes the cost can be determined directly (in dollars, for example), while in other cases it may be expressed in terms of the time required to take the measurement. A crude but inexpensive measure is sometimes preferable to a more precise measure that is too expensive. Blood tests, for example, are not as accurate as exploratory surgery in diagnosing some diseases, but they're faster, cheaper, and pose a lower risk of side-effects.

Foresters need quick measurements of the size of a tree, since they must measure many trees in a woodlot in order to assess its overall condition. List three variables that measure the size of a tree, and arrange them in order from most to least costly.

(c) Various studies have attempted to rank cities in terms of how desirable it is to live and work in each city. Describe five variables that you would measure for each city if you were designing such a study, and briefly explain your choices.

(d) Scientists who study human growth use different measures of the size of an individual; weight, height, and weight divided by height are three of the most common measures. If you were interested in studying the short-term effects of a digestive illness, which (if any) of these three variables would you use? Explain briefly.

(e) The usual method of determining heart rate in humans is to take the person's pulse by counting the number of beats in a given time period. The results are generally reported as beats per minute; for example, if the time period is 15 seconds, the observed count is multiplied by 4. You can either decide ahead of time to specify a particular time period and count how many beats occur, or you can measure how long it takes for the heart to beat a specified number of times.

Here are two measurements on the same person a few minutes apart: (i) 39 beats were observed in 30 seconds; (ii) it took 65 seconds to reach 80 beats. Convert each of these measurements to beats per minute. Which of these two readings do you think would be more accurate? Explain briefly.

2. This problem is about getting some practice with working with formulas involving summation notation. For more practice you can find and work through one of the tutorials on summation notation on the web (try typing a search string like *summation notation tutorial* at `google.com`); for example, some practice questions and explanations about summation notation can be found at

`www.columbia.edu/itc/sipa/math/summation.html`

In what follows n is an arbitrary integer greater than or equal to 1, $(y_1, \ldots, y_n)$ is an arbitrary list of numbers with mean $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, and c is an arbitrary constant.

(a) Expand out and simplify the following expressions.

(i) $\sum_{i=1}^{3} 1$

(ii) $\sum_{i=1}^{n} 1$

(iii) $\sum_{i=1}^{5} i$

(iv) $\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{j=1}^{n} y_j\right)$

(v) $\sum_{i=1}^{n}(y_i + c)$

(b) Show that $\sum_{i=1}^{n} c\, y_i = c \sum_{i=1}^{n} y_i$.

(c) Show that the sum of the deviations (from the mean) has to be zero no matter what the numbers are that go into the calculation:

$$\sum_{i=1}^{n}(y_i - \bar{y}) = 0.$$

3. In each of the following data-gathering situations, visualize the raw data as a table with rows for sample subjects and columns for variables, by doing the following two things: answer the question "In this data set there would be one row for each ______," and identify all of the following data types that apply to each variable: qualitative, quantitative, nominal, ordinal, dichotomous (in this case specify what values of the variable will be coded 1 and 0), discrete (include only variables that are inherently discrete (like number of leaves on a plant), not variables that might be made discrete by the measuring process), continuous, interval-scale, ratio-scale. For example, if I were looking at litter size in foxes, there would be one row for each litter (not one row for each fox), and the variable *litter size* would be quantitative, discrete, and ratio-scale.

(a) In the problem involving litter size in foxes, I decide just to keep track of whether or not each litter has 6 or more foxes in it. (Would this be as informative as recording the actual litter size?)

(b) I measure concentration of phosphates (in millimoles per liter, typically from agricultural run-off) at each of 60 randomly-sampled stream locations in Santa Cruz County, on the same day of the week and time of day in each case.

(c) I have a theory that the rate at which crickets chirp depends on the ambient air temperature, so for 44 nights in a row in the summer at a location where I've observed that crickets sometimes chirp I keep track of the temperature (in $^{\circ}C$) at which their rate of chirping falls below 100 per minute.

(d) I take a sample of vertebrate animals from a forest and record whether each animal is an amphibian, a turtle, a snake, a bird, or a mammal.