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

STAT 7: Homework 1

Tentative due date: by 11.59pm Fri 11 Oct 2019 at `canvas.ucsc.edu` *[100 total points]*
(the **actual** due date will be announced in class and in `canvas`)

Here's a style guide for all of the written work in this class. In figuring out how to write up answers to homework (and quiz, and midterm, and final) problems, pretend the grader is sitting there with you and you're having a brief discussion with her/him on each question — that is, write down in a few sentences what you would say to someone to support your position. It's never enough in this class to just say "yes" or "10.3," even if the right answer is "yes" or "10.3"; you need to say "yes (or 10.3), because ... ." The right answer with no reasoning to support it, or the wrong reasoning, will get about half credit in this course, as will the wrong answer arrived at with a good effort. Leaving a problem or a part of a problem blank will get no credit.

1. *[25 points]* 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) I'm studying the life expectancy of Pacific white-sided dolphins in their natural range from the Gulf of Alaska to Baja California.

(b) I'm examining whether there is a relationship between hair color and gender in humans (for example, is black hair more frequent in men than women?).

(c) In an experiment I'm planning, 10 *Drosophila* eggs will be put in a vial and the positions of the pupae will be noted after pupation. Some will pupate at the margin (or the wall) of the vial (these are called *peripheral*); others will pupate away from the wall (these are called *central*). I'm interested in the proportion of central pupae.

(d) By observation of a sample of 12 nesting pairs of Steller's jays, I make a record of the directions (compass points) relative to the tree trunks in which they've chosen to locate their nests (for example, a nest on a branch pointing due East would have a reading of 90°), to see if they tend to favor South-facing nesting sites that get more sun.

2. *[25 points]* Olson and Miller (1958) measured the interorbital width (how far apart the eyes are) in a sample of 40 domestic pigeons, obtaining the following results (measurements are in mm):

12.2	12.9	11.8	11.9	11.6	11.1	12.3	12.2	11.8	11.8	10.7	11.5	11.3	11.2
11.6	11.9	14.3	11.2	10.5	11.1	12.1	11.9	10.4	10.7	10.8	11.0	11.9	10.2
10.9	11.6	10.8	11.6	10.4	10.7	12.0	12.4	11.7	11.8	11.3	11.1		

We'll usually leave what I'm about to ask you to do to a computer (for example, in STAT 7L), but it's a good idea to do one or two of each thing you ask a computer to do by hand (so that you know what the computer is doing on your behalf), so in that spirit: please group these measurements into a frequency distribution and draw its histogram, making a reasonable choice for how wide the bars should be and where to start the first bar (you may need to look at several choices to find a good one). How would you qualitatively describe the shape of this distribution? Would you identify any of the observations as outliers (if so, which ones, and why)? Explain briefly.

3. *[25 points]* I have sample measurements $(y_1, y_2, \ldots, y_n)$ on a quantitative continuous variable with mean $\bar{y}$, median $\tilde{y}$, and standard deviation (SD) s, and I've drawn a histogram based on these measurements (having made a particular (and reasonable) choice for how wide the bars should be and where to start the first bar).

(a) (i) If I add a positive constant c to all of the data values, what will that do to the mean $\bar{y}$? What will it do to the median $\tilde{y}$? What will it do to the SD s? What will it do to the basic shape of the histogram? Explain briefly, using a combination of intuitive and graphical arguments (an example of the beginning of an intuitive argument: the mean is the balance point of the distribution, so ...). (Extra credit *[5 extra points]* for the mean and SD: demonstrate your answers algebraically using summation notation.) (ii) Repeat (i) if I subtract a positive constant c from all the data values.

(b) (i) If I multiply all of the data values by a positive constant c, what will that do to the mean $\bar{y}$? What will it do to the median $\tilde{y}$? What will it do to the SD s? What will it do to the basic shape of the histogram? Explain briefly, using a combination of intuitive and graphical arguments. (Extra credit *[5 extra points]* for the mean and SD: demonstrate your answers algebraically using summation notation.) (ii) Repeat (i) if I multiply all of the data values by a negative constant c.

If you're not sure of any of your answers, you can approach the problem experimentally: (I) make up a little data set and try adding (say) 6 to all of the numbers to see what will happen; then (II) think about whether there was anything special about your data set and the number 6, or whether your conclusion in (I) would hold in general.

> *(It may surprise you to hear, but this is a general problem-solving strategy in math, which — like physics — is both a theoretical and an experimental science: if a general pattern is not immediately apparent, try just playing around with numbers until the basic relationship starts to come into focus for you.)*

4. *[25 points]* Sokal and Rohlf (1995) present data on the wing lengths (in mm) of a sample of $n = 960$ houseflies; this data set had a mean of 4.6 and an SD of 0.4, and was approximately normally distributed. To get some practice using the normal distribution descriptively, please answer the following questions (explain briefly in each case, and show all of your calculations).

(a) About what percentage of these houseflies had wing lengths between 4.2 and 5.0mm?

(b) If you chose one of these houseflies at random, what (approximately) is the chance that its wing length would be less than 3.6mm?

(c) About how many of the houseflies in this sample had wing lengths greater than 5.4mm?