

# Biostatistics

Math 322 — Spring 2009

Homework 1 — Due 01/23

## Problem 1

- a) R does not have built in commands for calculating the geometric mean or the harmonic mean. Define R functions `geometricmean(v)` and `harmonicmean(v)` that calculates the geometric mean and the harmonic mean of a vector `v`, respectively.

See `090116_examples.R` (on the web page) for examples of how to define new R functions.

- b) Consider the `heights` data set we collected during the first lecture (available as `heights.txt` on the web page).

Calculate the arithmetic, geometric, and harmonic means of the heights, both for the whole sample, and for the samples grouped by sex.

## Problem 2 (Problems 2.8 – 2.11 in the book)

Table 2.12 (page 38 in the book) comes from a paper giving the distribution of astigmatism in 1033 young men, aged 18 – 22, who were accepted for military service in Great Britain. Assume astigmatism is rounded to the nearest 10th of a diopter and each subject in a group has the average astigmatism within that group (e.g., for the group 0.2 – 0.3 diopters, the actual range is from 0.15 to 0.35 diopters, and assume that each man in the group has an astigmatism of  $\frac{1}{2}(0.15 + 0.35) = 0.25$  diopters).

- a) Compute the arithmetic mean.
- b) Compute the standard deviation.
- c) Construct a bar graph to display the distribution of astigmatism.
- d) Suppose we wish to construct an upper bound for astigmatism so that no more than 5% of military recruits exceed this upper bound. What is the smallest possible value for this upper bound?

### Problem 3 (Problems 2.31 – 2.32 in the book)

In Section 2.9 of the book, the data set `LEAD.DAT` is described (`LEAD.DAT` is available on the CD-ROM that came with the book). The data set concerns the effect of lead exposure on neurological and psychological function in children.

- a) Compare the exposed and control groups on age and gender, using appropriate numeric and graphic descriptive measures.
- b) Compare the exposed and control groups on verbal and performance IQ, using appropriate numeric and graphic descriptive measures.

### Problem 4

- a) Use the `airquality` data set built into R (type `?airquality` for a description of the data set). Make a Stem-and-Leaf plot of the ozone levels, and a bar graph of temperatures grouped on months.

Comment on how useful these plots are for summarizing the data?

- b) Consider again the `heights` data set. Make three plots of the data using graphical methods that you choose yourself (although, none of the methods already covered).

Comment on what useful information (if any) that each plot gives you.

Useful R commands: `help.search("plot")`, `apropos("plot")`.