Stat 139: Homework 1

Due: Friday, September 12, before noon

- **Incorporate the relevant R output in the HW write-up**. Choose the output wisely, no need to print-out dozens of pages. No more than **two pages** of output should be displayed for each problem and the relevant parts should be **highlighted**.
- **A hard copy of your solutions is due at noon on Friday, in your TF's drop-box outside the Science Center 300 suite**. The solutions should not be submitted by email unless arrangements are made with a TF prior to the deadline (or unless late, see below).
- **Your code should be attached to the end of your write-up.** Use comments to indicate the code associated with each question.
- See syllabus for details on homework and collaboration policies: **acknowledge your collaborators**; your lowest homework score will be dropped; solutions submitted electronically within 24 hours after deadline will be graded with a penalty; solutions more than 24 hours late will receive no credit.

Related Reading: Ramsey and Schafer, Chapter 1.

Suggested review sources:

- *Introduction to the Practice of Statistics*, Moore, McCabe, [and Craig if 5th ed.], Ch. 1-5.
- Section on Matrix Algebra (Appendix A) in A. Sen and M. Srivastava, "Regression Analysis: Theory, Methods, and Applications" (or any other textbook used in math courses that you took).
- Review Sheet under "Readings" on the course web-site.

1. (Not Graded) Complete the following Conceptual Exercises in Chapter 1 (2nd or 3rd edition of R&S): 2, 7, 9. Feel free to include your answers in the write-up. However, this question will not be graded.
2. (20 points) For each of the following surveys, specify study units, the target population, and the sampling frame. Discuss in 2-3 sentences any possible sources of selection bias, specifically, undercoverage or overcoverage. Finally, specify what type of sampling was used.
   (a) (10 points) To estimate how many books in the library need rebinding, a librarian uses a random number generator to select 100 locations on library shelves. He then walks to each location, looks at the book that resides at that spot, and records whether the book needs rebinding or not.
   (b) (10 points) The Arizona Intrastate Travel Committee commissioned a study to identify in-state travel patterns of residents of major metropolitan cities and to evaluate different sources of vacation planning information. The plan was to conduct phone interviews with Phoenix and Tucson residents. Landline telephone numbers with Phoenix and Tucson area codes were generated randomly so that listed and unlisted telephone numbers could be reached. (Arizona Office of Tourism, 1991.)

3. (5 points) Consider two securities, the first having expected return $\mu_1 = 1$ and standard deviation $\sigma_1 = 0.1$, and the second having $\mu_2 = 0.8$ and $\sigma = 0.12$. Also, let the correlation between securities be $\rho = 0.1$. Suppose you invest $\pi = 0.8$, or 80%, of your money in the first security and 1-π in the second security. What is your expected return and what is the standard deviation of your return?

4. (10 points) Solve the following equations for the matrix X. You can assume that the matrices A and B (when needed) are all invertible n x n matrices.

   (a) $AXB = C$

   (b) $(AX) + B = D$

   (c) Solve equation (b) for X by hand if

$$A = \begin{bmatrix} 2 & 3 & 4 \\ -1 & 5 & -3 \\ 6 & -2 & 8 \end{bmatrix}, B = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, D = \begin{bmatrix} 11 \\ -7 \\ 26 \end{bmatrix}.$$

   Show all steps of your calculations and round the elements of X to the nearest integer.

5. (15 points) The `csdata.txt` data set on the course web-site contains information on 224 computer science students. Use R to perform the following tasks:

   (a) Split the students into two groups with GPA < 3 and GPA ≥ 3 and provide the following numerical summaries of the distributions of SAT Math (`SATM`) scores for each of the two groups: sample mean, sample SD, min, median, max, 1st and 3rd quartiles.

   (b) Plot histograms and box-plots of SATM scores for both groups side-by-side and describe the shapes of their distributions. Are there any visible differences?

   (c) Comment on whether you think the group means are very different or not (without conducting any formal tests).

   (d) Calculate the overall median SATM score in the sample and interpret it.

   (e) SAT is calibrated such that scores in the entire student population are distributed approximately normally with mean 500 and standard deviation 110. Identifying what fraction of student population is expected to score above the sample median found in the previous part? (Hint: use the R command `pnorm()` to find normal probabilities.) Are your findings consistent with an assertion that CS students have stronger math skills?