Problem 1. A total of $r$ keys are to be put, one at a time, in $k$ boxes. Each key is independently put in box $i$ with probability $p_i$, $\sum_{i=1}^{k} p_i = 1$. Each time a key is put in a nonempty box, we say that a collision occurs. Find the expected number of collisions.

Problem 2. The blue M&M candy was introduced in 1995. Before then, the color mix in a bag of plain M&Ms was as follows: 30% brown, 20% yellow, 20% red, 10% green, 10% orange, 10% tan). After the change, it was (24% blue, 20% green, 16% orange, 14% yellow, 13% red, 13% brown). A friend of mine has two bags of M&Ms, and he tells me that one is from 1994 and one from 1996. He won’t tell me which is which, but he gives me one M&M from each bag. One is yellow and one is green. What is the probability that the yellow M&M came from the 1994 bag?

Problem 3. Let $A$, $B$, $C$ be $n \times n$ matrices. Show that their product is associative, that is:

$$(AB)C = A(BC).$$

Now let $x \neq 0$ be a $n \times 1$ column vector. Let

$$D = \frac{xx^T}{x^Tx}.$$  

Use the above associativity property to compute $D^2$.

Problem 4. Let $X_i$, $i = 1, \ldots, n$ be i.i.d. random variables with mean $\mu$ and variance $\sigma^2$. Let $\bar{X}$ denote the sample mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$  

Now define:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$  

Show that $\mathbb{E}[s^2] = \sigma^2$, i.e., that $s^2$ is an unbiased estimate of $\sigma^2$.

Problem 5. Let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ be two random vectors (i.e., vectors of random variables).

Suppose that the covariance between $X_i$ and $Y_j$ is $G_{ij}$. In other words, the covariance matrix between the random vectors $X$ and $Y$ is $G$.

Let $A$ and $B$ be two $n \times n$ matrices. What is the covariance matrix between $AX$ and $BY$?

Obtain the electric.company dataset from the following location:

This data is from an educational experiment performed around 1970 on a set of elementary school classes. The treatment in this experiment was exposure to a new educational television show called The Electric Company. In each of four grades, the classes were split into treated and control groups. At the end of the school year, students in all the classes were given a reading test, and the average test score within each class was recorded. In every class, the treatment group was provided with one of the two options (the column named Supplement in the data): Supplement (S) or Replacement (R). Supplement (S) meant that the Electric Company TV show was supplemented with what the students were taught in class for the treatment group. Replacement (R) meant that the Electric Company TV show was prescribed as a replacement to what students were taught in the class for the treatment group.

Do the following:

a) Plot treated.Posttest vs. control.Posttest for Supplement cases for all grades and cities. Find the sample average of Treated.Posttest - Control.Posttest for Supplement cases.

b) Plot Treated.Posttest vs. Control.Posttest for Replacement cases for all grades and cities. Find the sample average of Treated.Posttest - Control.Posttest for Replacement cases.

c) Make a line $y = x$ through both the above plots. What observations can you make? Do most points lie above the $y = x$ line or below? Can you argue whether the Replacement (R) strategy was better or worse on average than the Supplement (S) strategy for this data?