MATHEMATICS E-156, SPRING 2014
MATHEMATICAL FOUNDATIONS OF STATISTICAL SOFTWARE
Module #1 (Fundamentals of Probability, illustrated in R)

Last modified: January 25, 2014

**Reading from Chihara and Hesterberg**

- Chapter 1. Pay special attention to things like the distinction between an *observational study* and an *experiment*. As a mathematician who is self-taught in statistics, I am weak on the "political" aspects of the subject, but the textbook covers these issues nicely.
- Appendix A, sections A.1 and A.2. On this material I am expert and experienced and will fill in many of the omitted proofs.

**Optional Reading from Haigh**

- Sections 1.1 through 1.5. If you have ever taken any sort of probability course, this will all be review.
- Sections 4.1 and 4.2. This treats the material of appendices A.1 and A.2 in much more detail.

**Proof of the Week**

- The variance $\mathrm{Var}[X]$ of a random variable $X$ is defined as $E[(X - E[X])^2]$.

  Given that $E[a_1X_1 + a_2X_2] = a_1E[X_1] + a_2E[X_2]$ in all cases and that $E[X_1X_2] = E[X_1]E[X_2]$ for independent random variables, prove that
  - $\mathrm{Var}[X] = E[X]^2 - (E[X])^2$.
  - $\mathrm{Var}[aX + b] = a^2\ \mathrm{Var}[X]$.
  - If $X_1$ and $X_2$ are independent, $\mathrm{Var}[X_1 + X_2] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2]$

  (This is all done in section A.2)

**R scripts**

- 0A-StartTutorial.R
  This script includes all the commands from the first of the three tutorials that the textbook authors have posted on their Web site. It is on the "Installing R" page of the course Web site. By using it, you can work through the tutorial without having to type in all the commands.

- 1A-DiceProbs.R This script uses R to deal with a classic probability problem: rolling one, two, or three fair dice. The theme is that you can think of the vectors and vectorized operations of R as corresponding to random variables of a finite probabiliity space and operations on these random variables.

- 1B-FunnyDice.R You will learn how to create a data frame, either by using the R editor or by using the "rep()" and "c()" functions, and how to create and investigate a new random variable made from columns of the data frame.

- 1C-CardPairs.R A probability problem that sometimes baffles beginners is "Show that if you choose a pair of cards at random from a deck of 52 cards, the probability that at least one of them is a spade is 15/34 (not 1/2)." This script creates a "probability data frame" that solves this problem and can solve many others like it. It also addresses some tricky issues involving strings and "factors" which might cause you unexpected trouble if you are not aware of them.

- 1D-CaseStudies.R This script looks briefly at each of the case studies that is described in chapter 1 of the textbook. It raises, but does not solve, a number of interesting statistics problems.

- 1E-Math23.R This script shows how to make a data frame from an existing Excel file, which you will probably want to do for your term project.

- 1P-Proof1.R This script illustrates some of the results about expectation and variance of random variables that are the subject of proof 1.

**Mathematical notes**

1. Finite Probability, done in R

   - The sample space $S$ consists of all the rows in a data frame.
   - An outcome $\omega$ of an "experiment" is an individual row.
   - An "event" $A$ is just a subset of the rows. The R function `which()` returns such a subset.
   - The "sigma field" of events to which probabilities can be assigned is the "power set" $2^S$: the collection of all possible subsets.
   - For a probability function, just assume that each of the $n$ rows has probability $1/n$. Then the probability of event $A$ can be determined from the number of rows in event (subset)$A$.
   - If you want unequal probabilities whose ratio is a rational number like 3/2, include three rows for the first outcome, two for the second.
   - A numeric column in a data frame is a random variable, since it assigns a real number to each row. When the "vectorized" operators in R act on columns of a data frame, they carry out the specified operation on each row and can be viewed as operations on random variables.
   - A constant, since it gets replicated $n$ times when combined with a vector of length $n$, behaves like a constant random variable when it appears in an $R$ expression like `X+2`.
   - A logical column in a data frame specifies an event $A$, the set of rows for which the value is `TRUE`. The complement of event $A$ is the the set of rows for which the value is `FALSE`.
   - Since `TRUE` has the value 1, a logical column can be treated like a random variable; namely, the "indicator function" that equals 1 if event $A$ occurs, 0 if it does not.
   - A "factor" column in a data frame specifies a partition of the sample space into disjoint events (subsets) $B_1, B_2, \cdots B_k$. Although the factors are displayed as strings, they are stored internally as the indices $1 \cdots k$ of the events in the partition.
   - Since an expression like `Card$Rank == "Queen"` has the value 1 or 0, it can be viewed as an indicator function for an event in a partition.
   - Given two sample spaces $S_1$ and $S_2$, the R function `expand.grid()` constructs the product sample space $S_1 \times S_2$.
   - The R function `mean()` computes the expectation of a random variable.
   - Many elementary probability problems can be solved by counting the rows in a data frame that correspond to a specified event.

2. Expectation of a discrete random variable

   If $X$ is a discrete random variable, it can only have values in the finite or countably infinite set $x_i$. Its expectation is defined as

   $$E[X] = \sum_i x_i P(X = x_i), \text{, where the sum may be finite or infinite.}$$

   Strictly speaking, the sum is over the set of possible values, not "over the sample space." However, when the sample space is finite this nicety can be ignored.

   Suppose that there is a second random variable $Y$ which takes values in the set $y_i$. In order to define a probability function on the sample space $S$, we must know the probability of each event of the form $(X = x_i, Y = y_j)$

   If $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$,then $X$ and $Y$ are said to be *independent* random variables.

   (a) Prove that if $X$ and $Y$ are discrete random variables and $Z = X + Y$, then $E[Z] = E[X] + E[Y]$, even if $X$ and $Y$ are not independent. (This is a fussy proof, and you are not responsible for the details.)

(b) Prove that if if $X$ and $Y$ are independent discrete random variables and $Z = XY$, then $E[Z] = E[X]E[Y]$.

(c) The converse is not necessarily true: it is possible to invent "uncorrelated" random variables for which $E[XY] = E[X]E[Y]$, that are not independent.

(d) A similar result, called the "Law of the Unconscious Statistician," is Theorem 4.8 in Haigh.

If $X$ is a discrete random variable then $Y = h(X)$ is also a random variable, and

$$E[Y] = \sum_i P(X = x_i)h(x_i).$$

Given two random variables $X$ and $Y$, we can calculate

$$E[f(X)g(Y)] = \sum_{i.j} P(X = x_i, Y = y_i)f(x_i)g(y_i).$$

$E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ for any pair of functions $f$ and $g$ if and only if $X$ and $Y$ are independent. To prove independence choose a function $f$ that equals 1 only if $X = x_i$ (otherwise $f$ is zero) and a function $g$ that equals 1 only if $Y = y_j$.

3. Proof of the week

   The variance $\mathrm{Var}[X]$ of a random variable $X$ is defined as $E[(X - E[X])^2]$.

   Given that $E[a_1X_1 + a_2X_2] = a_1E[X_1] + a_2E[X_2]$ in all cases and that $E[X_1X_2] = E[X_1]E[X_2]$ for independent random variables, prove that

   (a) $\mathrm{Var}[X] = E[X]^2 - (E[X])^2$.

   (b) $\mathrm{Var}[aX + b] = a^2\ \mathrm{Var}[X]$.

   (c) If $X_1$ and $X_2$ are independent, $\mathrm{Var}[X_1 + X_2] = \mathrm{Var}[X_1]+\mathrm{Var}[X_2]$

**Homework assignment** This assignment should be submitted as a single R script. Include enough comments so that it is clear what you are doing and where each problem begins. You can upload it to the dropbox on the Class 1 page of the Web site.

It is OK to paste and edit lines from the scripts on the course Web site. It is not OK to paste lines from your classmates' solutions!

1. By following the instructions on the Web site, download and install R. You will get credit for this problem automatically by submitting an R script for this assignment.

2. Work through the first R tutorial. Then, in the script that you submit for this assignment, load FlightDelays.csv and write a few lines of R to do the following:

   Determine the median flight delay for each day of the week.
   Determine the mean flight delay for flights on AA on Wednesdays.
   Make a histogram of all the flight lengths (in minutes).
   Make a histogram of all the flight lengths (in minutes) for flights to DEN.

3. Download Dice3.csv from the Extra Datasets page of the course Web site and use it to answer the following question:

   When you roll three fair dice, what is the probability $P1$ that they all show the same number, the probability $P2$ that they show two different numbers, and the probability $P3$ that they show three different numbers? Of course, $P1 + P2 + P3 = 1$. You are welcome to check your answer by approaching this as a counting problem, but the challenge is to do it by having R count rows for each event.

4. This is a variant of the standard example of two random variables that are uncorrelated but not independent.

   You cannot resist the offer of a free two-week online course on the structure of polymeric molybdate oxoanions. Alas, the lecture videos turn out to be in Armenian, and the online textbook is in Amharic. There are two quizzes, each consisting of a single multiple choice question with four responses. You can do nothing but guess, so your probability of getting a score of 100 on a quiz is 1/4, while your probability of getting 0 is 3/4.

   - Make a data frame with columns Q1 and Q2 for your two quiz scores. By using 16 rows, all assumed equally likely, you can make the probabilities come out correct. Do this in three ways:
     QEdit is done using the R Data Editor (see Funny Dice).(your script will not show theresult of your editing)
     QRep is done using the `rep()` function (see Funny Dice).
     QGrid is done using the `expand.grid()` function (see CardPairs).

   - Using QRep or QGrid, which should be identical, create two random variables:
     $X$ is your average quiz score, 100, 50, or 0.
     $Y$ is your improvement rating. This is 0 if you score worse on the second quiz, 1 if you score the same on both quizzes, 2 if you score better on the second quiz.
     Calculate $E[XY] - E[X]E[Y]$. Since the random variables are uncorrelated, this should equal zero.
     Calculate $E[X^2Y^2] - E[X^2]E[Y^2]$. If $X$ and $Y$ were independent this would also equal zero.
     Invent an event $A$ involving $X$ and and an event $B$ involving $Y$ such that $P(A \cap B) \neq P(A)P(B)$, and do the calculation of these probabilities in R by counting rows.

5. Make a data frame by loading the file cereals.csv from your Data subfolder and write a script that does the following. The details are up to you.

   - Make a barplot.
   - Make a histogram.
   - Make a contingency table using two factors.
   - Calculate a mean broken down by factor.
   - Extract a subset of one numeric variable for one factor.