HW#1 - last 3 pages

MATHEMATICS 156/E-156, FALL 2016
MATHEMATICAL FOUNDATIONS OF STATISTICAL SOFTWARE
Module #1 (Fundamentals of Probability, illustrated in R)

Last modified: January 19, 2016

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.
- Theorem 6.2 and Definition 6.3 on pp. 149-150. This theorem explains why the R function var() does not compute what you might have expected.
- 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.

Proofs of the Week

- Proof 1: Calculating variance

  The variance $\text{Var}[X]$ 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

  1. $\text{Var}[X] = E[X^2] - (E[X])^2$.
  2. $\text{Var}[aX + b] = a^2 \text{ Var}[X]$.
  3. If X_1 and X_2 are independent, $\text{Var}[X_1 + X_2] = \text{Var}[X_1]+\text{Var}[X_2]$

- Proof 2: Variance of the mean of n independent random variables

  This is a theorem of probability.

  Suppose that $X_1, X_2, \cdots X_n$ are independent random variables, all with the same expectation μ and variance σ^2. Their mean is

  $$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots X_n).$$

  Prove that

  $$E[\overline{X}] = \mu$$

  and that

  $$\text{Var}[\overline{X}] = \frac{1}{n}\sigma^2.$$

- Proof 3: Variance of the sample mean of n independent random variables

  This is a theorem of statistics.

  Let $X_1, X_2, \cdots X_n$ be independent random variables from a distribution with $\text{Var}[X_i] = \sigma^2 < \infty$.

  We do not know the expectation $\mu = E(X_i)$, although we know from the previous result that the expectation of $\overline{X}$ is equal to μ.

  We also do not know the variance of X_i. We try to estimate it by using the usual formula but, not knowing μ, we can do no better than to use $\overline{X}$ in its place.

  $$\text{Prove that } E[\sum_{i=1}^{n}(X_i - \overline{X})^2] = (n-1)\sigma^2.$$

  It follows that $S^2 = E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2] = \sigma^2$. This is what var() computes.

R scripts

- 1A-FunnyDice.R
  Topic 1 - Creating a data frame for a probability problem
  Topic 2 - Generating samples from a data frame

- 1B-CardPairs.R
  Topic 1 - A probability data frame for a deck of cards
  Topic 2 - Probabilities for events that involve two cards
  Topic 3 - Approximating probabilities by sampling

- 1C-CaseStudies.R
  Topic 1 - Flight delays out of LaGuardia airport
  Topic 2 - A quick look at other datasets

- 1D-Math23.R
  Topic 1 - Old Math 23 grades, broken down by yard and section

- 1P-Proofs1.R
  Topic 1 - Proof 1: calculating variance in a single pass through the data
  Topic 2 - Proof 2: variance of the sum or mean of n random variables
  Topic 3 - Proof 3: variance of the sample mean
  Topic 4 - Testing for independence

Mathematical notes

1. 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, each of which takes on values in a finite subset of $\mathbb{R}$, and $Z = X + Y$, then $E[Z] = E[X] + E[Y]$, even if X and Y are not independent.

(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.
Present an example of non-independent variables X and Y for which $E[Z] = E[X]E[Y]$.

(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

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

It takes a while to realize that this is a theorem that requires proof, since it seems obvious. But the definition of expectation says that

$$E[Y] = \sum_i P(Y = y_j)y_j.$$

Here is an example that is as instructive as a formal proof.

Suppose that X is a random variable that has the following values:

- -2, with probability 0.1.
- -1, with probability 0.2.
- +1, with probability 0.4.
- +2, with probability 0.3.

Calculate $E[Y]$ first by using the definition, then by using the Law of the Unconscious Statistician.

(e) A useful test for independence.

Given two random variables X and Y, we can calculate

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

Prove that $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ for any pair of functions f and g if and only if X and Y are independent.

To prove independence (the "only if" part) 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 $Y = y_j$.

2. Proof 1: Useful rules for calculating variance

   The variance 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) Var$[X] = E[X^2] - (E[X])^2$.

   (b) Var$[aX + b] = a^2$ Var$[X]$.

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

3. Proof 2: Variance of the mean of n independent random variables

   This is a theorem of probability.

   Suppose that $X_1, X_2, \cdots X_n$ are independent random variables, all with the same expectation μ and variance σ^2. Their mean is

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots X_n).$$

   Prove that

$$E[\overline{X}] = \mu$$

   and that

$$\text{Var}[\overline{X}] = \frac{1}{n}\sigma^2.$$

4. Proof 3: Variance of the sample mean of n independent random variables

   This is a theorem of statistics. You can find it on page 149 of the textbook. If you look there, ignore the unfamiliar terminology that was introduced earlier in the chapter.

   Let $X_1, X_2, \cdots X_n$ be independent random variables from a distribution with $\mathrm{Var}[X_i] = \sigma^2 < \infty$.

   We do not know the expectation $\mu = E(x_i)$, although we know from the previous result that the expectation of $\overline{X}$ is equal to μ.

   We also do not know the variance of X_i. We can try to estimate it by using the usual formula but, not knowing μ, we can do no better than to use $\overline{X}$ in its place.

   $$\text{Prove that } E[\sum_{i=1}^{n}(X_i - \overline{X})^2] = (n-1)\sigma^2.$$

   $$\text{It follows that } S^2 = E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2] = \sigma^2.$$

   This *sample variance* S^2 is what `var()` computes.

Problems for class in Week 1 Students who live in or near Cambridge will do these in class at 5:30 PM on February 3.

Distance students will do this in an online section at 8 PM on February 3.

The first two problems are standard probability problems. The challenge is to solve them by brute force in R, creating a data frame and then counting what fraction of the rows correspond to the specified event. The third problem will give you some experience with statistical and graphical functions in R.

The class will divide into groups of three or four students. Each group will tackle one of the three problems and upload its solution to the "Scripts for section problems" page on the Week 1 page of the course Web site, then get to work on the other problems. Starting at 7:15, three groups will explain their uploaded solutions to the entire class.

1. Classic Bayes's Rule

   A flight out of Mogadishu airport has 100 passengers, all male.

   20% of them wear explosive shoes, and within this group 60% have beards. Among the 80% who wear non-explosive shoes, only 5% have beards.

   Create a data frame with one row per passenger and two columns "Shoes" and "Beard." Make a 2x2 contingency table using these columns.

   Then extract the subset of bearded passengers. Use it to calculate the conditional probability that if you inspect a randomly chosen passenger with a beard, he will turn out to have explosive shoes.

2. (This problem would be tedious to do by pencil and paper.)

   The Red Sox and Cardinals are playing another World Series.

   Games 1, 2, 6, and 7 are in Boston, where the Red Sox have a probability of 2/3 of winning. Games 3, 4, and 5 are in St. Louis, where their probability of a Red Sox victory is only 1/2.

   Make a 3-element vector for the Boston games, with a 1 for Red Sox victory. Make a 2-element vector for the St. Louis Games, again with a 1 for Red Sox victory. Use expand.grid() to make a data frame with $3^4 2^3 = 648$ rows that correspond to equally-likely outcomes if the teams play 7 games, and use it to make a histogram of the number of Red Sox victories if the teams play all seven games. There is an R function named rowsum() that looks potentially useful, but you will have to check the online documentation.

   Of course, an actual World Series ends when one team has won four games. Extract the subset that corresponds to an event like "the Red Sox win the series in six games," and figure out the probability of that event. There are eight such events. Different members of the class can analyze different events, and someone should check that the probabilities sum to 1.

3. Load Beerwings.csv. This should already be on your computer if you followed the installation instructions. Then write a few lines of R to do the following:

    - Make a barplot showing beer consumption.

    - Determine the median beer consumption for males and for females.

    - Determine the difference between the mean number of hotwings consumed by males and by females.

    - Count the number of females. Choose a random sample of rows equal in size to the number of females, and calculate the mean number of hotwings consumed by the patrons in those rows of the data frame. Repeat 1000 times and make a histogram of the results. Where do the means for males and for females fall on this histogram?

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 Assignment 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, which you saw in class last week. 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 Week 1 on the Modules page 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, as in script 1B-Funny Dice.R
     QEdit is done using the R Data Editor .(your script will not show the result of your editing)
     QRep is done using the `rep()` function.
     QGrid is done using the `expand.grid()` function.

   - 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. Based on Exercise 1 on page 393.

   A box contains 30 red balls and 20 blue balls.

   (a) If you draw 15 balls at random without replacement, what is the probability that you draw 8 red balls and 7 blue balls?

   (b) If you draw 15 balls at random without replacement, how many red balls to you expect?

   (c) Make a vector in R to represent the 50 balls. Then sample 15 balls without replacement and count the number of red balls. Do this 10000 times, and determine what fraction of the time you get 8 red balls and what is the mean number of red balls. The results should be close to your answers to parts (a) and (b).

   The R function `choose(n,k)` calculates the binomial coefficient $\binom{n}{k}$. By using it you can solve part (a) in one line. You can check your answer against page 406.

6. 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.