Problem Set #2: Probability (revised 9/16/2011)

Economics 435: Quantitative Methods

Fall 2011

1 A simple random variable

Suppose you flip a coin twice. Let:

$$\begin{aligned} x_1 &= I(\text{first coin is heads}) \\ x_2 &= I(\text{second coin is heads}) \\ y &= x_1 + x_2 \end{aligned}$$

where I is the indicator[1] function.

a) Let Ω_x be the sample space for x_1 (it can also be used as the sample space for x_2), and let Ω_y be the sample space for y. Define these spaces (by enumeration or other method).

b) The sample space of (x_1, x_2, y) is $\Omega \equiv \Omega_x \times \Omega_x \times \Omega_y$, where "$\times$" indicates the Cartesian product. Find and report the joint PDF of (x_1, x_2, y) for each combination of values in Ω.

c) Find and report the conditional PDF of y given $x_1 = 1$

d) Find the (marginal) CDF of y. In addition to reporting it in a table, plot it on a graph.

e) Find $E(y|x_1)$.

f) Find $E(x_1|y)$.

g) Find $corr(x_1, y)$

2 The effects of smoking

Suppose that we are interested in determining the effect of cigarette smoking on the probability of lung cancer, and we have the following pieces of information:

- Based on the 2001 Census, Canada has 30 million residents.
- Based on survey data from Health Canada, 20% of Canadians smoke.
- Based on hospital records, 75% of lung cancer patients are smokers.
- Based on hospital records 20,000 Canadians are diagnosed with lung cancer each year.

[1] If you've never seen the indicator function, it takes a statement (like "first coin is heads") as its argument and returns a value of one if the statement is true and zero if the statement is false. For example, $I(\text{Victoria is the capital of B.C.}) = 1$ and $I(5 < 3) = 0$.

Note that these numbers are strictly guesses on my part, especially the 75% one.

a) What is the probability that a randomly selected Canadian will be diagnosed with lung cancer in a given year?

b) What is the probability that a randomly selected Canadian will be diagnosed with lung cancer in a given year, conditional on being a smoker?

c) What is the probability that a randomly selected Canadian will be diagnosed with lung cancer in a given year, conditional on being a nonsmoker?

d) These results imply that a smoker is x times as likely as a nonsmoker to get lung cancer in a given year. Find the value of x.

3 Probability theory

Let A_1 and A_2 be events in the probability space $(\Omega, \Psi, \Pr)$. Prove each of the following theorems, using only results from set theory and the definition of a probability space (once you have proved one result, you can use it in the proof of another result).

a) If $A_1 \subset A_2$, then $\Pr(A_1) \leq \Pr(A_2)$.

b) Let A_1 and A_2 be disjoint events ($A_1 \cap A_2 = \emptyset$), and let A_3 be some other event. Then:

$$\Pr(A_3|A_1 \cup A_2) = \frac{\Pr(A_3|A_1)\Pr(A_1) + \Pr(A_3|A_2)\Pr(A_2)}{\Pr(A_1) + \Pr(A_2)}$$

c) If $\Pr(A_2) > 0$, then:

$$\Pr(A_1|A_2) = \frac{\Pr(A_2|A_1)\Pr(A_1)}{\Pr(A_2)}$$

This result is known as Bayes' Law.

4 Properties of expectations

Let X and Y be random variables with well-defined expected values and variances, and let a, b, and c be constants. Prove the following results using only the definitions of the various terms, the linearity of expectations, and the law of iterated expectations.

a) $var(X) = E(X^2) - [E(X)]^2$

b) $cov(aX + bY, cX + dY) = ac\, var(X) + bd\, var(Y) + (ad + bc)\, cov(X, Y)$

c) If $E(Y|X) = 0$, then $E(g(X)Y) = 0$ for any function $g(.)$.

5 Basic data manipulation in R

Write an R script that does the following:

1. Define a function called `rbern` that takes 3 numerical arguments: (n, k, p) and generates an n-by-k matrix of Bernouilli(p) random variables. You may not use the built-in R function `rbinom` Here's how you should do it:
   (a) Generate a vector of nk random variables from the standard uniform distribution. You will find the function `runif` useful here.

   (b) Convert that vector into a vector of zeros and ones that have the Bernoulli(p) distribution. You will find the function `as.integer` useful here.
   (c) Reshape that vector into an n-by-k matrix. You will find the function `matrix` useful here.

2. Using that function, generate and print out a 3-by-5 matrix of Bernouilli(0.75) random variables.
3. Calculate and print out the average of each column of this matrix. You will find the `apply` and `average` functions useful here.

Send me your R script by WebCT.