BST 675 — Fall 2010 — Dr. Charnigo

Written Assignment 1

Written Assignment 1 is due on Wednesday 08 September at the end of class. You are encouraged to work in groups of two or three, but you may work individually if you prefer.

[10] 1. Give a careful proof of the first distributive law, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

[10] 2. Prove result 5 that $P(A) \leq 1$ without using any subsequent result. (Thus, you are limited to result 4 and axioms 1, 2, and 3.)

[10] 3. Derive a general formula for $P(A \cup B \cup C)$ in terms of probabilities involving A, B, C, and their intersections. (As a first step, put $D := B \cup C$.)

[10] 4. Draw a Venn diagram that provides an intuitive visual interpretation of the formula derived in exercise 3.

[20] 5. Our second motivating case study illustrated Simpson's paradox. Find and describe another example from public health or medicine that illustrates Simpson's paradox.

[20] 6. Suppose that we wish to measure the association between cigarette smoking and lung cancer in Kentucky. A randomized controlled trial is unethical, but we can conduct either a cohort study or a case-control study. A cohort study entails recruiting (say) $5,000$ smokers and $5,000$ non-smokers, whom we then follow forward in time to see how many in each group develop lung cancer. A case-control study entails recruiting (say) 100 people with lung cancer and 100 people without lung cancer, whom we then survey to see how many in each group were smokers. Letting A denote the event that a person develops lung cancer and B the event that a person smokes, describe how you might estimate $P(A|B)$ from a cohort study and how you might estimate $P(A|B)$ from a case-control study under the supposition that $P(A) = 0.01$. (This exercise represents a simplification of reality because we have not made explicit a time horizon for the development of lung cancer. Nonetheless, this exercise reveals why disease risk within an exposure stratum cannot be estimated from a case-control study, unless one makes a supposition about the disease prevalence within the full population. That is why, ordinarily, relative risks are not estimated in a case-control study.)

[20] 7. Suppose that 0.7% (7/10 of 1%, not 70%) of people in the U.S. have HIV and that a diagnostic test is available such that: (i) 90% of people who really have HIV test positive; and, (ii) P% of people who really do not have HIV test negative, where P is some number between 0 and 100. How large must P be so that half of people testing positive for HIV actually have HIV?