CSE 5301 - *Data Analysis & Modeling Techniques*

Homework 1- Spring 2017

Due Date: Feb. 16 2017, 3:30 pm

Probability

1. The probability of a program having a major bug that will result in it crashing is 0.2. There are two testing procedures that can be used to detect such bugs. Each of the testing procedure utilizes 50% of the code and the first one detects 50% of all bugs occurring in its code area. The second testing procedure can detect 30% of the bugs occurring in its code area. Both procedures overlap in their evaluation on 50% of the code that they use during evaluation and can therefore both detect the same bugs if they occur in this area (but do so in different, independent ways). Neither one of the tests will erroneously produce a bug warning and it is assumed that a major bug is equally likely in all code areas (none of these assumptions is assumed to be realistic for real software testing). How high is the probability that a program does not have a major bug if it passes both testing procedures ?

2. In a computer vision application it is important to distinguish between different object put in from of a camera. In particular, it is important to distinguish between a box, a cylinder, and a sphere. To do this, the company developed a simple program that counts the number of corners the program finds in the image. Due to errors, the program sometimes finds too many or too few corners but never more than 5. To evaluate the accuracy of the program, a number of experiments with different objects of the given types are made and the following likelihoods to find $0..5$ corners is determined:

object / #corners found	0	1	2	3	4	5
sphere	0.75	0.1	0.05	0.05	0.025	0025
cylinder	0.05	0.15	0.3	0.35	0.1	0.05
box	0.05	0.05	0.1	0.15	0.3	0.35

   a) Determine the probability of an object being of each of the different types if 3 corners are found and all objects are equally likely to be the object presented.

   b) Determine for each object the likelihood that it is identified correctly using this algorithm.

Statistics of Datasets

3. Given the following data sets. Compute their *mean*, *true variance*, *unbiased variance*, and *median*:

a) $\{1, 5, 2, 7, 3, 1, 8\}$

b) $\{2.2, 1.1, 1.9, 2.4, 5.8\}$

c) $\{7, -3, 1, 4, 0\}$

d) $\{2, 4, 2, 3, 5, 2, 1, 6\}$

Distributions

4. Modeling Experiments
   Particular types of outcomes of experiments can be modeled with a set of standard distributions. For each of the experiments and experimental questions described below derive (and list) the corresponding distribution and answer the particular questions about the experiment. (Note that these descriptions are not intended to be realistic models for the corresponding situations but rather dramatically simplified scenarios.)

   a) Consider a situation where a number of backup copies are made of a hard drive (and all stored in the same location) and by some random event, each bit of each of the backups is randomly corrupted (i.e. randomly set to 0 or 1) with a probability of 0.1. If we have 5 copies, what is the likelihood that we can correctly reconstruct a particular bit using a majority scheme (i.e. if we assume that the most frequently occurring bit is the correct one)? How many copies would we have to make in order to e able to reconstruct a given byte with a probability higher than 99.9%

   b) A set of 1000 data items is stored redundantly in a database with 3 copies existing (therefore there are 3000 entries in the database). During a break-in, 100 random data entries are maliciously modified by inverting the letters and numbers. What is the likelihood that the retrieval attempts for the first 50 data items (each retrieval attempt retrieving all copies of a data item) result in three uncorrupted data copies (i.e. in the situation where all three copies have not been modified) ? What is the likelihood that the same retrieval attempt for 50 items results in all items being correctly restored by a majority vote (i.e. two of the three copies have not been tampered with) ?

   d) A robot uses a special phase difference-based laser sensor to determine the distance from obstacles in his way. This sensor continuously outputs the distance but due to a glitch in the sensor electronics, it randomly returns an incorrect value at an average rate of 3 times per second. What is the likelihood that the sensor does not return any incorrect value within the first minute of operation ?

5. Given two random variables, x, and y, drawn from independent random distributions, X, and Y, with means and variances of μ_X and σ_X^2, and μ_Y and σ_Y^2, respectively, derive the *mean* and the *variance* of the distributions for $x - y$. (Make sure to hand in your entire derivation)