Homework 1: Combinatorics & Empirical Distributions

UC Irvine CS177: Applications of Probability in Computer Science

Due on October 10, 2019 at 11:59pm

Question 1: (20 points)

A system is called "k out of n" if it functions reliably when at least k of its n components are working; in other words, the system uses redundancy to ensure robustness to failure. As an example, consider a *redundant array of inexpensive disks* (RAID) in which one uses n disks to store a collection of data, and as long as at least k are functioning the data can be correctly read. Suppose that disks fail independently, and that the probability of an individual disk failing in a one-year period is p.

a) *Suppose we have a $n = 3$ disk array which can survive one failure ($k = 2$). What is the expected number of disk failures in one year? As a function of p, what is the probability that the whole array will continue to function without any data loss after one year?*

b) *Suppose we have a $n = 5$ disk array which can survive two failures ($k = 3$). What is the expected number of disk failures in one year? As a function of p, what is the probability that the whole array will continue to function without any data loss after one year?*

c) *Suppose $p = 0.05$. Which is more reliable (has greater probability of not losing any data in one year), the RAID from part (a) or part (b)?*

d) *Suppose $p = 0.65$. Which is more reliable, the RAID from part (a) or part (b)?*

Question 2: (20 points)

Consider a social network that allows accounts to be secured with a 6-digit passcode (any sequence of exactly six digits between *0-9* is valid). Assume the network has m users including you, and that all users choose one of the valid 6-digit passcodes uniformly at random. A user's passcode is considered safe if no other user has the same passcode.

a) As a function of m, what is the probability that your own passcode is safe?

b) How many users must there be for there to be a 50% or greater chance that your own passcode is not *safe?*

c) As a function of m, what is the probability that all users have a safe passcode?

d) How many users must there be for there to be a 50% or greater chance that at least one user's passcode is not *safe?*

Question 3: (20 points)

Consider a set of n people who are members of an online social network. Suppose that each pair of people are linked as "friends" independently with probability 1/2. We can think of their relationships as a graph with n nodes (one for each person), and an undirected edge between each pair that are friends. A *clique* is a fully connected subset of the graph, or equivalently a subset of people for which all pairs are friends.

a) *A clique of size 2 is simply a pair of nodes that are linked by an edge. Find the expected number of edges as a function of the number of nodes, n. What is the expected number of friend relationships among $n = 10$ people?*

b) *A clique of size 3 is a triplet of nodes within which all three pairs are linked by an edge. Find the expected number of 3-cliques as a function of the number of nodes, n. What is the expected number of 3-cliques among $n = 10$ people?*

c) *Larger cliques may occur involving groups of nodes of any size k. Derive a general formula for the expected number of cliques of any size $2 \le k \le n$ as a function of the number of nodes, n. What is the expected number of cliques of size $k = 4$ among $n = 10$ people?*

Question 4: (40 points)

We will now analyze some data collected by observing the famous "Old Faithful" geyser in Yellowstone National Park. We define random variable S to be the time an eruption lasts, and random variable T to be the "waiting time" until the next eruption. These are clearly continuous random variables, but we do not precisely know their true distribution. Instead we have a dataset with $n = 272$ independent observations $(s_i, t_i), i = 1, \dots, 272$, of the eruption time s_i and subsequent waiting time t_i. See Figure 1 for a plot of this data.

In the following questions, we compute various quantities using the *empirical distribution* of the data. The empirical distribution of eruption time and waiting time can be represented by a probability mass function $p_{ST}(s,t)$ which places probability $1/n$ on each of the n data points, and probability 0 on the continuous range of other (s,t) values. Under this distribution, the expected values of S and T then take the following simple form:

$$E[S] = \frac{1}{n}\sum_{i=1}^{n} s_i, \qquad E[T] = \frac{1}{n}\sum_{i=1}^{n} t_i.$$

a) *The variance of random variable S equals $Var[S] = E[S^2] - E[S]^2$. Give formulas for computing $Var[S]$ and $Var[T]$ under the empirical distribution. Use Python's* `numpy.sum` *function to write your own code that computes these variances, and report their values.* Hint: Various definitions of the "sample variance" can be found in statistics references, and they are not all equivalent to the variance of the empirical distribution.

b) *The cumulative distribution of S equals $F_S(s) = P(S \le s)$, where the probability is under the empirical distribution. Find eruption times $\bar{s}_1, \bar{s}_2, \bar{s}_3$ such that $F_S(\bar{s}_1) = 0.25$, $F_S(\bar{s}_2) = 0.50$, $F_S(\bar{s}_3) = 0.75$. Using the cumulative distribution of T, also find waiting times $\bar{t}_1, \bar{t}_2, \bar{t}_3$ such that $F_T(\bar{t}_1) = 0.25$, $F_T(\bar{t}_2) = 0.50$, $F_T(\bar{t}_3) = 0.75$.* Hint: One solution would be to use Python's `numpy.argsort` function.

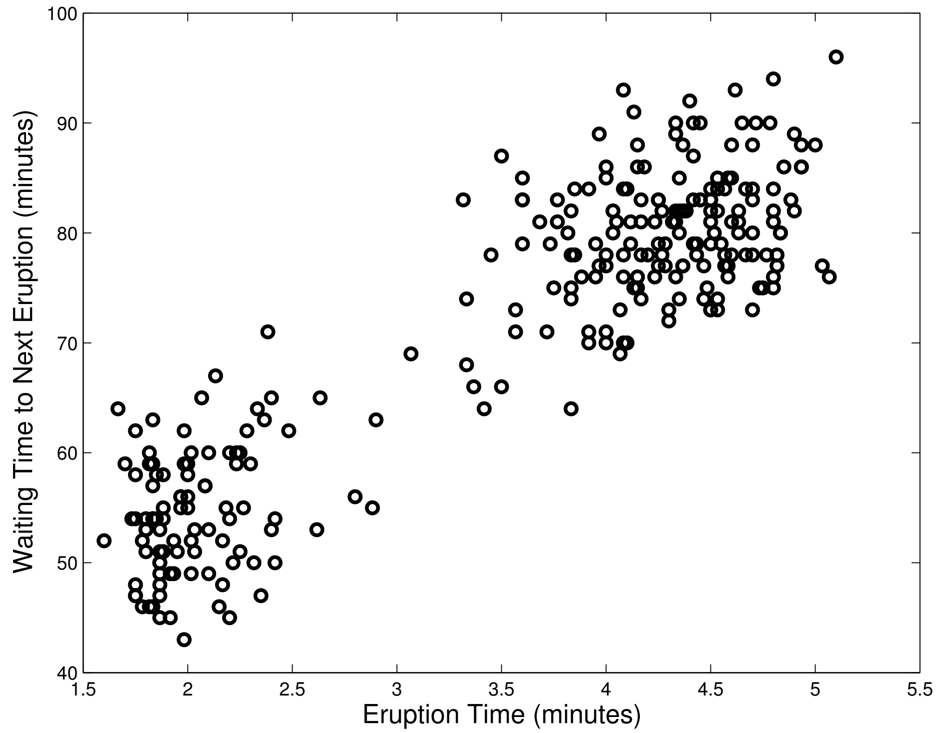

Figure 1: A "scatter plot" of the observations of Old Faithful's eruption time (horizontal axis) and waiting time to the next eruption (vertical axis). Each point is one of the $n = 272$ observations.

Consider two new random variables. Let X indicate whether the eruption time S is "short" or "long": $X = 0$ if $S \leq 3.5$, and $X = 1$ if $S > 3.5$. Let Y indicate whether the waiting time T is "short" or "long": $Y = 0$ if $T \leq 70$, and $Y = 1$ if $T > 70$.

c) *Using the empirical distribution of S and T, determine and report the joint probability mass function $p_{XY}(x, y)$. Also determine and report the marginal probability mass functions $p_X(x)$ and $p_Y(y)$.*

d) *Are the random variables X and Y independent? If not, is the amount of dependence weak or strong? Clearly justify your answer using the probability mass functions from part (c).*