Probability and Statistics – Problem Set 3

Question 1. There are 14 singers at The Voice 2021 getting ready for a public show and each needs to choose an unordered playlist of 3 different songs from 50 total songs. Each singer chooses a playlist of 3 songs; any subset of 3 songs out of 50 is equally likely. The singers choose songs independently of each other, i.e., it is possible to reuse the same song across different singers, but a singer cannot use a song more than once.

a) The music director needs to buy the rights to the songs the singers choose for any show. How many songs should director expect to buy the rights to for the show?
b) The rights to each song cost \$100, and there is an overall processing fee of \$40 to complete all the transactions. What is the expected cost to the music director?

Question 2. Suppose we have two coins. Coin C_1 comes up heads with probability 0.3 and coin C_2 comes up heads with probability 0.9. We repeat this process 3 times:

- Choose a coin with equal probability.
- Flip that coin once.

Suppose X is the number of heads after 3 flips.

a) What is $E[X]$?
b) What is $Var(X)$?
c) Based on the number of heads we get, we earn $Y = \frac{1}{X+1}$, dollars. What is $E[Y]$?

Question 3. To determine whether a community of 1000 people containing Covid19 cases, we have their blood tested. However, rather than testing each individual separately (1000 tests is quite costly), it is decided to use a pooled testing strategy:

- **Phase 1**: First, place 1000 people into groups of 5. The blood samples of the 5 people in each group will be pooled and analyzed together. If the test is positive (at least one person in the pool has Covid19 virus), continue to Phase 2. Otherwise, we can send the group home. Totally, 200 of these pooled tests are performed.

- **Phase 2**: Individually test each of the 5 people in the group. 5 of these individual tests are performed per group in Phase 2. Suppose that the probability that a person has Covid19 is 5% for all people, independently of others, and that the test has a 100% true positive rate and 0% false positive rate (note that this is unrealistic).

Using this strategy, compute the expected total number of blood tests (individual and pooled) that we will have to do across Phases 1 and 2.

Question 4. There are 100 shoelaces in a box. At each stage, you pick two random ends and tie them together. Either this results in a longer shoelace (if the two ends came from different pieces), or it results in a loop (if the two ends came from the same piece). What are the expected

number of steps until everything is in loops, and the expected number of loops after everything is in loops?

Question 5. Suppose we have a hash function $h: \mathcal{U} \to \{0,1, \dots, m-1\}$ which maps from a universe $\mathcal{U}$ of strings (with length < 100) into m buckets, with each string independently and equally likely to be hashed into any bucket. We want to insert n strings $s_1, \dots, s_n$ into our hash table.

a) Let $X_1 = h(s_1)$ be the index of the bucket that string s_1 hashes into. What distribution does X_1 have?
b) What is the probability that two particular strings s_1 and s_2 hash to the same bucket?
c) If Y_1 is the number of strings in the first bucket after inserting all n strings, what distribution does Y_1 have? What is the probability that the first bucket is empty?
d) What is the expected number of empty buckets?

Question 6. Suppose you are working at a technology company ABC, and you are unfortunately on-call for your team the entire year (that means, you are the person that other stakeholders may ping in the middle of the night to debug production issues). There are 5 Software Engineers on your team (including yourself), and each person independently introduces on average 0.1 bugs per work-week (Mon-Fri).

a) What is the probability of having a bug-free work-week?
b) What is the probability of having a bug-free day? What's the relationship between your answer to this part and the previous part?
c) What is the probability that in a (52-week) year, that there are at least 40 bug-free weeks?
d) Suppose it's the first Monday of the year. When would you expect the first day where you had to debug (at least) one issue (in number of work-days from today)?
e) Suppose it's the first Monday of the year. What is the probability that your tenth day of debugging happens in February or later (>20 work-days from now)?