2023 Spring Probability & Statistics for EECS March 19, 2023

Homework 06

Professor: Ziyu Shao & Dingzhu Wen *Due: 23:59 on March 26, 2023*

1. Suppose there are n types of toys, which you are collecting one by one. Each time you collect a toy, it is equally likely to be any of the n types. What is the expected number of distinct toy types that you have after you have collected t toys? (Assume that you will definitely collect t toys, whether or not you obtain a complete set before then.)

2. A coin with probability p of Heads is flipped n times. The sequence of outcomes can be divided into runs (blocks of H's or blocks of T's), *e.g.*, HHHTTHTTTH becomes $\boxed{\text{HHH}}\boxed{\text{TT}}\boxed{\text{H}}\boxed{\text{TTT}}\boxed{\text{H}}$, which has 5 runs. Find the expected number of runs.

3. Elk dwell in a certain forest. There are N elk, of which a simple random sample of size n is captured and tagged (so all $\binom{N}{n}$ sets of n elk are equally likely). The captured elk are returned to the population, and then a new sample is drawn. This is an important method that is widely used in ecology, known as capture-recapture. If the new sample is also a simple random sample, with some fixed size, then the number of tagged elk in the new sample is Hypergeometric.

   For this problem, assume that instead of having a fixed sample size, elk are sampled one by one without replacement until m tagged elk have been recaptured, where m is specified in advance (of course, assume that $1 \le m \le n \le N$). An advantage of this sampling method is that it can be used to avoid ending up with a very small number of tagged elk (maybe even zero), which would be problematic in many applications of capture-recapture. A disadvantage is not knowing how large the sample will be.

   (a) Find the PMFs of the number of untagged elk in the new sample (call this X) and of the total number of elk in the new sample (call this Y).

   (b) Find the expected sample size $E[Y]$ using symmetry, linearity, and indicator r.v.s.

   (c) Suppose that m, n, N are such that $E[Y]$ is an integer. If the sampling is done with a fixed sample size equal to $E[Y]$ rather than sampling until exactly m tagged elk are obtained, find the expected number of tagged elk in the sample. Is it less than m, equal to m, or greater than m (for $n < N$)?

4. People are arriving at a party one at a time. While waiting for more people to arrive they entertain themselves by comparing their birthdays. Let X be the number of people needed to obtain a birthday match, *i.e.*, before person X arrives there are no two people with the same birthday, but when person X arrives there is a match.

Assume for this problem that there are 365 days in a year all equally likely. By the result of the birthday problem form Chapter 1, for 23 people there is a 50.7% chance of a birthday match (and for 22 people there is a less than 50% chance). But this has to do with the *median* of X; we also want to know the *mean* of X, and in this problem we will find it, and see how it compares with 23.

(a) A *median* of a random variable Y is a value m for which $P(Y \leq m) \geq 1/2$ and $P(Y \geq m) \geq 1/2$. Every distribution has a median, but for some distributions it is not unique. Show that 23 is the *unique* median of X.

(b) Show that $X = I_1 + I_2 + \cdots + I_{366}$, where I_j is the indicator random variable for the event $X \geq j$. Then find $E(X)$ in terms of p_j's defined by $p_1 = p_2 = 1$ and for $3 \leq j \leq 366$,

$$p_j = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)\cdots\left(1 - \frac{j-2}{365}\right).$$

(c) Compute $E(X)$ numerically (do NOT submit the code if used).

(d) Find the variance of X, both in terms of the p_j's and numerically (do NOT submit the code if used).

5. Suppose there are 5 boxes (with tags 1, 2, 3, 4, 5) and we are going to put 14 balls into these boxes. It is known that one can at most put 6 balls in a box. How many different ways can you distribute these balls?