Probability & Statistics for EECS: Homework #6 Solution

Problem 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.)

Solution

Let I_j be the indicator of having the jth toy type in your collection after having collected t toys. By symmetry, linearity, and the fundamental bridge, the desired expectation is:

$$n(1 - (\frac{n-1}{n})^t)$$

Problem 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.

Solution

Let I_j be the indicator for the event that position j starts a new run, for $1 \leq j \leq n$. Then $I_1 = 1$ always holds. For $2 \leq j \leq n$, $I_j = 1$ if and only if the jth toss differs from the $(j-1)st$ toss. So for $2 \leq j \leq n$,

$$E(I_j) = P((j-1)st\ toss\ H\ and\ jth\ toss\ T, or\ vice\ versa) = 2p(1-p)$$

Hence, the expected number of runs is $1 + 2(n-1)p(1-p)$.

Problem 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 $1 \leq m \leq n \leq N$). 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 $n < N$)?

Solution

(a) The event $X = k$ says that there are $m-1$ tagged elk and k untagged elk in the first $m+k-1$ elk sampled, and that the $(m+k)th$ elk sampled is tagged. So

$$P(X = k) = \frac{\begin{pmatrix} n \\ m-1 \end{pmatrix}\begin{pmatrix} N-n \\ k \end{pmatrix}}{\begin{pmatrix} N \\ m+k-1 \end{pmatrix}} \cdot \frac{n-m+1}{N-m-k+1}$$

for $k = 0, 1, \ldots, N-n$ (note that $k = 0$ is the case where the first m elk sampled are all tagged, and $k = N-n$ is the case where we have to collect all the untagged elk before recapturing a tagged elk). This is known as the Negative Hypergeometric distribution. The PMF of Y can then be found by noting that $Y = X + m$: for $y = m, m+1, \ldots, N-n+m$

$$P(Y = y) = P(X = y-m) = \frac{\begin{pmatrix} n \\ m-1 \end{pmatrix}\begin{pmatrix} N-n \\ y-m \end{pmatrix}}{\begin{pmatrix} N \\ y-1 \end{pmatrix}} \cdot \frac{n-m+1}{N-y+1}$$

An alternative way to obtain the PMF of X is as follows. First find the probability of a particular way of having $X = k$ occur: getting k untagged elk in a row, followed by m tagged elk in a row. This event has probability

$$\frac{(N-n)(N-n-1)\cdots(N-n-k+1)n(n-1)\cdots(n-m+1)}{N(N-1)\cdots(N-m-k+1)} = \frac{n!(N-m-k)!(N-n)!}{(n-m)!N!(N-n-k)!}$$

Writing 1 for "tagged" and 0 for "untagged", we just found the probability of $00\ldots011\ldots1$, with k 0's and m 1's. But the first $m+k-1$ of these symbols can be in any order without affecting the value of X; moreover, the probability of any such sequence (k 0's and $m-1$ 1's in some order, followed by a 1) is the same as what we just found, since the terms in the numerator remain the same (just in permuted order) and likewise for the denominator. Thus, for $k = 0, 1, \dots, N - n$,

$$P(X = k) = \binom{m+k-1}{m-1} \cdot \frac{n!(N-m-k)!(N-n)!}{(n-m)!N!(N-n-k)!} = \frac{\binom{m+k-1}{m-1}\binom{N-m-k}{n-m}}{\binom{N}{n}}$$

(b) As suggested in the hint, assume that the elk get captured until all N of them have been obtained. This is convenient since then we are just looking at a random permutation of the N elk, and it is valid since what transpires after m tagged elk have been recaptured does not affect the value of X. Define $X_1, \dots, X_m$ as in the hint. Label the untagged elk as $1, 2, \dots, N - n$ and write $X_1 = I_1 + \cdots + I_{N-n}$, where I_j is the indicator of Untagged Elk j being captured before any tagged elk.
By symmetry, $E(I_j) = 1/(n + 1)$ since Untagged Elk j and the n tagged elk are equally likely to be in any order. So $E(X_1) = (N - n)/(n + 1)$. For example, for $N = 10$ elk with $n = 4$ tagged, labeled $7, 8, 9, 10$ and $N - n = 6$ untagged, labeled $1, 2, 3, 4, 5, 6$, and with $m = 3$, the observed evidence (if all elk are collected) could be

$$\underbrace{❺❷❸}_{X_1}\ ⑨\ \underbrace{❻}_{X_2}\ ⑦\ \underbrace{❹❶}_{X_3}\ ⑩⑧.$$

The observed values of I_5, I_2, I_3 are 1 and of I_1, I_4, I_6 are 0. Before the data are collected, $E(I_5) = 1/5$ since Untagged Elk 5 and Tagged Elk $7, 8, 9, 10$ are equally likely to be in any order.
Similarly, $E(X_j) = (N - n)/(n + 1)$ for all $j = 1, \dots, m$ since each untagged elk is equally likely to be positioned anywhere among the n tagged elk. Thus,

$$E(X) = \frac{m(N-n)}{n+1}, E(Y) = m + \frac{m(N-n)}{n+1} = \frac{m(N+1)}{n+1}.$$

(c) With a fixed sample size equal to $E(Y)$, the number of tagged elk in the sample is Hypergeometric with mean

$$\frac{m(N+1)}{n+1} \cdot \frac{n}{N} = m \cdot \frac{1 + \frac{1}{N}}{1 + \frac{1}{n}} < m.$$

If n is small and N is large, then this is a major difference between the two sampling methods; if n is large, then the above expectation is approximately m.

Problem 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).

Solution

(a) Because $P(X \leq m)$ is equal to the probability of $1-p$, where p is the probability of choosing m different dates from 365 days, the order matters. So, $p = A_{365}^m/365^m$, and

$$P(X \leq m) = 1 - \frac{A_{365}^m}{365^m}$$

where $A_m^n = \frac{m!}{(m-n)!}$ is the permutation number.

And we can calculate the following probability by code:

$$\begin{aligned} P(X \leq 22) &\approx 0.48 \quad P(X \geq 22) \approx 0.56 \\ P(X \leq 23) &\approx 0.51 \quad P(X \geq 23) \approx 0.52 \\ P(X \leq 24) &\approx 0.54 \quad P(X \geq 24) \approx 0.49 \end{aligned}$$

Therefore, we can conclude that 23 is the unique median of X by the monotonicity of CDF.

(b) Because by (a), $P(X \leq j) = 1 - A_{365}^j/365^j$, so we have:

$$P(X \geq j) = 1 - P(X \leq j-1) = \frac{A_{365}^{j-1}}{365^{j-1}}$$

Then we can simplify the form of p_j:

$$\begin{aligned} p_j &= (1 - \frac{1}{365})(1 - \frac{2}{365})\ldots(1 - \frac{j-2}{365}) \\ &= 1 \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365-(j-2)}{365} \\ &= \frac{365!/(365-(j-1))!}{365^{j-1}} \\ &= \frac{A_{365}^{j-1}}{365^{j-1}} \end{aligned}$$

Therefore,

$$P(X \geq j) = P(I_j = 1) = E(I_j) = p_j = \frac{A_{365}^{j-1}}{365^{j-1}}.$$

By the definition of I_j: $I_j = 0$ (for $j > X$), $I_j = 1$ (for $j \leq X$), we have

$$X = \sum_{j=1}^{366} I_j.$$

So in conclusion,

$$\begin{aligned} E(X) &= E(\sum_{j=1}^{366} I_j) = \sum_{j=1}^{366} E(I_j) \\ &= \sum_{j=1}^{366} p_j \end{aligned}$$

(c) Calculating by code, we have: $E(X) \approx 24.6166$.

(d) Solution: First note that $I_i^2 = I_i$ (this always true for indicator r.v.s) and that $I_i I_j = I_j$ for $i < j$ (since it is the indicator of $\{X \geq i\} \cap \{X \geq j\}$). Therefore,

$$\begin{aligned} X^2 &= I_1 + \cdots + I_{366} + 2\sum_{j=2}^{366}(j-1)I_j, \\ E\left(X^2\right) &= p_1 + \cdots + p_{366} + 2\sum_{j=2}^{366}(j-1)p_j \\ &= \sum_{j=1}^{366}(2j-1)p_j, \end{aligned}$$

and the variance is

$$\operatorname{Var}(X) = E\left(X^2\right) - (EX)^2 = \sum_{j=1}^{366}(2j-1)p_j - \left(\sum_{j=1}^{366} p_j\right)^2.$$

Entering p <- c$(1, \text{cumprod}(1 - (0:364)/365))$; $\text{sum}((2*(1:366)-1)*p) - (\text{sum(p)})$^2 in R yields $\operatorname{Var}(X) \approx 148.640$.

Problem 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?

Solution

To solve this problem by Generating Function: Because the number of balls in one boxes is in $[0, 6]$, so we can use the function: $G_i(x) = 1 + x + x^2 + x^3 + x^4 + x^5 + x^6$ $(1 \leq i \leq 5)$ to indicate the number of balls in one box.
Therefore, we can have the generating function indicating number of balls in all of the 5 boxes:

$$G(x) = \prod_{i=1}^{6} G_i(x) = (1 + x + x^2 + x^3 + x^4 + x^5 + x^6)^5.$$

Because the total number of balls is 14, so what we need is to find the coefficient of x^{14}:

$$\begin{aligned} G(x) &= (1 + x + x^2 + x^3 + x^4 + x^5 + x^6)^5 = (\frac{1 - x^7}{1 - x})^5 \\ &= (\sum_{n=1}^{5} \binom{5}{n} (-x)^{7n})(\sum_{n=0}^{\infty} x^n)^5 \end{aligned}$$

Since the coefficient of x^i in $(\sum_{n=0}^{\infty} x^n)^k$ is $\binom{i+k-1}{k-1}$, so the coefficient of x^{14} is:

$$\binom{5}{2} \times 1 - \binom{5}{1} \times \binom{7+5-1}{5-1} + \binom{5}{0} \times \binom{14+5-1}{5-1} = 1420$$

In conclusion, the result is 1420.