Probability & Statistics for EECS: Homework #06

Due on March 26 2023 at 23:59

Name: **Zhou Shouchen**
Student ID: 2021533042

Problem 1

Let I_j be the indicator that whether the j-th type of toy is collected. So the number of distinct toys types that we have collected is $X = \sum_{j=1}^{n} I_j$

Let A_j means that the j-th type of toy is collected. Since there are n types of toys, and we have collected t toys.

So $\forall j, P(A_j^c) = P(I_j = 0) = (1 - \frac{1}{n})^t$.

So $P(A_j) = P(I_j = 1) = 1 - P(I_j = 0) = 1 - (1 - \frac{1}{n})^t$.

Thus $E(X) = E(\sum_{j=1}^{n} I_j) = \sum_{j=1}^{n} E(I_j) = \sum_{j=1}^{n} P(A_j) = n \cdot (1 - (1 - \frac{1}{n})^t)$.

So above all, the expectation number of distinct toy types that we hae collected is $n[1 - (1 - \frac{1}{n})^t]$.

Problem 2

Let I_j be the indicator that whether the result of the j-th is different with the $(j+1)$-th$(1 \leq j \leq n-1)$. i.e. whether starts a new run.
Let A_j means that the j-th item of the result sequence is different with the $(j+1)$-th item of the result sequence.
Let X be the number of runs. Then $X = 1 + \sum_{j=1}^{n-1} I_j$ Suppose the result sequence is named S.
So $P(A_j) = P(I_j = 1) = P(S_j = H, S_{j+1} = T) + P(S_j = T, S_{j+1} = H) = 2p(1-p)$.
Thus $E(X) = E(1 + \sum_{j=1}^{n-1} I_j) = 1 + \sum_{j=1}^{n-1} E(I_j) = 1 + \sum_{j=1}^{n-1} P(A_j) = 1 + 2(n-1)p(1-p)$.

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

Problem 3

(a) For $X = k$, we know that the last elk that we captured must be a tagged one. So before the last tagged elk captured, there are totally $m-1$ tagged elk, and k untagged elk been captured.

For the first $k+m-1$ elks, the probability of capturing $m-1$ tagged elk and k untagged elks is $\frac{\binom{n}{m-1}\binom{N-n}{k}}{\binom{N}{m+k-1}}$.

As for the last tagged elk, the probability of capturing it is $\frac{n-(m-1)}{N-(m+k-1)} = \frac{n-m+1}{N-m-k+1}$.

Combine them, we can get that $P(X=k) = \frac{\binom{n}{m-1}\binom{N-n}{k}}{\binom{N}{m+k-1}} \cdot \frac{n-m+1}{N-m-k+1}$.

From the decription, we could know that $Y = X + m$,

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

So above all, the PMF of X and Y is that

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

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

(b) Suppose that we still capture the remaining elks even after m tagged elks are captured. The elks captured after m tagged elks were already captured will not effect what we want to calculate.

And we use the inserting board method, suppose that the n tagged elks are n boards, and the remaining untagged elks can be put into the $n+1$ intervals produced by the n boards.

So with the symmetry, we can know that for each untagged elk, it has equally probability of $\frac{1}{n+1}$ to be in any one of the interval.

Number the untagged elks $1, 2, \cdots, N-n$.

And let $I_{1,j}$ be the indicator that whether the j-th untagged elk is captured before the fisrt tagged elk is captured.

Let X_1 be the number of untagged elks that are captured before the first tagged elk is captured.

So $X_1 = \sum_{j=1}^{N-n} I_{1,j}$.

And for this situation, we can regard if the j-th untagged elk is put into the first interval, then $I_{1,j} = 1$.

So $P(I_{1,j}=1) = \frac{1}{n+1}$.

Thus, $E(X_1) = E(\sum_{j=1}^{N-n} I_{1,j}) = \sum_{j=1}^{N-n} E(I_{1,j}) = \sum_{j=1}^{N-n} P(I_{1,j}=1) = \frac{N-n}{n+1}$.

And for the second situation, let $I_{2,j}$ be the indicator that whether the j-th untagged elk is captured before the second tagged elk is captured, but after the first tagged elk is captured.

We can regard if the j-th untagged elk is put into the second interval, then $I_{2,j} = 1$.

Let X_2 be the number of untagged elks that are captured before the second tagged elk is captured, but after the first tagged elk is captured.

So $P(I_{2,j}=1) = \frac{1}{n+1}$.

Thus, $E(X_2) = E(\sum_{j=1}^{N-n} I_{2,j}) = \sum_{j=1}^{N-n} E(I_{2,j}) = \sum_{j=1}^{N-n} P(I_{2,j}=1) = \frac{N-n}{n+1}$.

.....

Problem 3 continued on next page...

For the i-th situation($i = 1, 2, \cdots, m$), we can regard if the j-th untagged elk is put into the i-th interval, then $I_{i,j} = 1$.

So $P(I_{i,j} = 1) = \dfrac{1}{n+1}$.

Thus, $E(X_i) = E(\sum_{j=1}^{N-n} I_{i,j}) = \sum_{j=1}^{N-n} E(I_{i,j}) = \sum_{j=1}^{N-n} P(I_{i,j} = 1) = \dfrac{N-n}{n+1}$.

And since $X = X_1 + \cdots + X_m$, so with the linearity of expectation,

we can get that $E(X) = E(X_1) + \cdots + E(X_m) = \dfrac{m(N-n)}{n+1}$.

And since $Y = m + X$, so we can get that $E(Y) = E(m + X) = m + E(X) = m + \dfrac{m(N-n)}{n+1}$.

So $E[Y] = \dfrac{m(N+1)}{n+1}$.

So above all, the expected sample size $E[Y] = \dfrac{m(N+1)}{n+1}$.

(c) Let Z be the number of tagged elks that we captured in this method.

And number the tagged elks $1, 2, \cdots, n$, then let I'_j be the indicator that whether the j-th tagged elk is captured.

So $Z = \sum_{j=1}^{n} I'_j$.

Since there are totally $E[Y]$ captured elks, so the probability of the j-th tagged elk is captured is

$P(I'_j = 1) = \dfrac{\binom{N-1}{E[Y]-1}}{\binom{N}{E[Y]}} = \dfrac{(N-1)!}{(E[Y]-1)!(N-E[Y])!} \cdot \dfrac{(E[Y])!(N-E[Y])!}{N!} = \dfrac{E[Y]}{N}$.

So $E(Z) = E(\sum_{j=1}^{n} I'_j) = \sum_{j=1}^{n} E(I'_j) = \sum_{j=1}^{n} P(I'_j = 1) = n \cdot \dfrac{E[Y]}{N} = \dfrac{n}{N} \cdot \dfrac{m(N+1)}{n+1}. = m \cdot \dfrac{1 + \frac{1}{N}}{1 + \frac{1}{n}}$.

Since we are given that $n < N$, so $\dfrac{1 + \frac{1}{N}}{1 + \frac{1}{n}} < 1$, so $E(Z) < m$.

So above all, the expected number of tagged elks that we captured is in this method is less than m.

Problem 4

If there are totally m people, then there are $\binom{m}{2}$ pairs. For each pair, the probability that they have the same birthday is $p = \frac{1}{365}$, which is a small number.
And let A_j means that the j-th paired people has the same birthday.
And I'_j be A_j's indicator.
So $I_j = I(A_j), P(A_j) = p = \frac{1}{365}$.

And let Z be the number of pairs that have the same birthday, so $Z = \sum_{j=1}^{\binom{m}{2}} I'_j$.
Since p is a small number, and A_j are independent. So according to the Poisson approximation, we can get that $Z \sim Pois(\lambda)$, where $\lambda = \sum_{j=1}^{\binom{m}{2}} P(A_j) = \binom{m}{2} \cdot p = \frac{m(m-1)}{2} \cdot \frac{1}{365}$.
So P(no birthday match when m people arrive) $= P(Z = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{\frac{m(m-1)}{2*365}} \approx e^{\frac{m^2}{2*365}}$.

(a) $P(X \geq k) = P(\textit{no birthday match when } k-1 \textit{ people arrive}) = e^{-\frac{(k-1)^2}{2*365}}$.
So $P(X \leq k) = 1 - P(X > k) = 1 - P(X \geq k+1) = 1 - e^{-\frac{k^2}{2*365}}$.
So with calculation, we can get that
$P(X \leq 22) = 1 - 0.515296 = 0.484704 < \frac{1}{2}$, so 22 is not the median of X.
$P(X \leq 23) = 1 - 0.484490 = 0.515510 > \frac{1}{2}$, $P(X \geq 23) = 0.515296 > \frac{1}{2}$. So 23 is the median of X.
$P(X \geq 24) = 0.484490 < \frac{1}{2}$. So 24 is the median of X.
And for $k < 22$, $P(X \leq k) \leq P(X \leq 22) < \frac{1}{2}$,
and for $k > 24$, $P(X \geq k) < P(X \geq 24) < \frac{1}{2}$.
So above all, 23 is the unique median of X.

(b) Suppose that $X = k$, then for $j \leq k, I_j = 1$, and for $j > k, I_j = 0$.
So $\sum_{j=1}^{366} I_j = \sum_{j=1}^{k} I_j + \sum_{j=k+1}^{366} I_j = \sum_{j=1}^{k} 1 + \sum_{j=j+1}^{366} 0 = k = X$.
So $X = I_1 + I_2 + \cdots + I_{366}$ has been proved.

For $P(X \geq j)$, without using Poisson approximation, just use the original method, we can get that
For $j > 2$,$P(X \geq j) = P(\textit{no birthday match when } j-1 \textit{ people arrive}) = \frac{A_{365}^{j-1}}{365^{j-1}} =$
$\frac{365 \cdot 364 \cdot \cdots \cdot (365+2-j)}{365^{j-1}} = 1(1 - \frac{1}{365})(1 - \frac{2}{365}) \cdots (1 - \frac{j-2}{365}) = p_j$.

And for $j = 1, 2$, $P(X \geq j) = P(\textit{no birthday match when } j-1 \textit{ people arrive})$, since $j - 1 = 0, 1$, so there must no have a birthday match. so $P(X \geq j) = 1 = p_j$.
So combine them, we will get that $P(X \geq j) = p_j$.
So $E(X) = E(I_1 + I_2 + \cdots + I_{366}) = E(I_1) + E(I_2) + \cdots + E(I_{366}) = \sum_{j=1}^{366} P(X \geq j) = \sum_{j=1}^{366} p_j$.

So above all, $X = I_1 + I_2 + \cdots + I_{366}$.
And $E(X) = \sum_{j=1}^{366} p_j$.

(c) With the p_j's general term formula, we can calculate the $E(X)$ easily with the help of program and the

formula $E(X) = \sum_{j=1}^{366} p_j$,
and for $3 \le j \le 366, p_j = (1 - \frac{1}{365})(1 - \frac{2}{365}) \cdots (1 - \frac{j-2}{365}), p_1 = p_2 = 1$.
After programming, we can compute that $E(X) = 24.616585894598852$.

So above all, $E(X) = 24.616585894598852$.

(d) Since $X = I_1 + I_2 + \cdots + I_{366}$, so $X^2 = \sum_{j=1}^{366} I_j^2 + 2\sum_{i=1}^{365}\sum_{j=i+1}^{366} I_i I_j$.
From the property of indicator, we can get that $I_j^2 = I_j$.
Since I_i the indicator of whether $X \ge i$, and I_j the indicator of whether $X \ge j$,
from the formula of getting sum, we can find that $j > i$ always hold, so $I_i I_j = I_j$.
and if we swap the order of getting sum, we can get that $\sum_{i=1}^{365}\sum_{j=i+1}^{366} I_i I_j = \sum_{j=2}^{366}\sum_{i=1}^{j-1} I_i I_j = \sum_{j=2}^{366}\sum_{i=1}^{j-1} I_j = \sum_{j=2}^{366}(j-1) \cdot I_j$.
So $X^2 = \sum_{j=1}^{366} I_j + 2\sum_{j=2}^{366}(j-1) \cdot I_j$.
So $E(X^2) = E(\sum_{j=1}^{366} I_j + 2\sum_{j=2}^{366}(j-1) \cdot I_j) = \sum_{j=1}^{366} E(I_j) + 2\sum_{j=2}^{366}(j-1) \cdot E(I_j)$.
As we have proved in (c), that $E(I_j) = P(X \ge j) = p_j$, so $E(X^2) = \sum_{j=1}^{366} p_j + 2\sum_{j=2}^{366}(j-1) \cdot p_j$.

And from the property of variance, we can get that
$Var(X) = E(X^2) - [E(X)]^2 = \sum_{j=1}^{366} p_j + 2\sum_{j=2}^{366}(j-1) \cdot p_j - (\sum_{j=1}^{366} p_j)^2$.
And we can calculate the $Var(X)$ easily with the help of program and the formula.
After programming, we can compute that $Var(X) = 148.64028478843568$.

So above all, $Var(X) = \sum_{j=1}^{366} p_j + 2\sum_{j=2}^{366}(j-1) \cdot p_j - (\sum_{j=1}^{366} p_j)^2 = 148.64028478843568$.

Problem 5

Suppose that the i-th box contains X_i balls, where $0 \leq X_i \leq 6$, and each X_i is with equal probability. i.e. $P(X_i = 0) = P(X_i = 1) = \cdots = P(X_i = 6) = \frac{1}{7}$.
So the total number of balls put into 5 boxes is $X = X_1 + X_2 + \cdots + X_5$. And what we want is to calculate $P(X = 14)$.
And we can use PGF to compute this. The PGF of X_1 is that

$$E[t^{X_1}] = \frac{1}{7}(1 + t + t^2 + \cdots + t^6)$$

Similarly, X_i are i.i.d. So the PGF of X is that

$$E[t^X] = E[t^{X_1 + \cdots + X_5}] = E[t^{X_1}] \cdots E[t^{X_5}] = \frac{1}{7^5}(1 + t + t^2 + \cdots + t^6)^5$$

And we can use the formula of PGF to compute the probability.

$$(1 + t + \cdots + t^6)^5 = [\frac{(1 - t^7)}{1 - t}]^5 = \frac{(1 - t^6)^5}{(1 - t)^5}$$

1.Firstly, for denominator.
Since $|t| < 1$ to make sure that the PGF is convergent, so at this range, use Taylor expansion, we can have

$$\frac{1}{(1 - t)^5} = (1 + t + t^2 + t^3 + \cdots)^5$$

So let a_k be the coefficient of t^k in the Taylor expansion of $\frac{1}{(1 - t)^5}$, and let $x_1 + x_2 + x_3 + x_4 + x_5 = k$, where x_i is the coefficient of t^i in the i-th polynomial, so x_i is a nonnegative integer. From Bose-Einstein that we have learned, a_k is exactly same as the number of solutions of this equaltion. So we can get that

$$a_k = \binom{k - 1 + 5}{4} = \binom{k + 4}{4}$$

2. Secondly, for numerator.
With the Binomial Theorem,

$$(1 - t^7)^5 = \sum_{k=0}^{5} \binom{5}{k} 1^{n-k} (-t^7)^k = \sum_{k=0}^{5} \binom{5}{k} (-1)^k t^{7k}$$

3. Combine the two parts.
We need to compute the coefficient of t^12 of the polynomial

$$(\sum_{k=0}^{5} \binom{5}{k} (-1)^k t^{7k})(\sum_{k=0}^{\infty} a_k t^k)$$

To get the coefficient of t^{14}, there are 3 cases.
(1). The first part take t^0, and the second part take t^{14}.
Then the coefficient is

$$(-1)^0 \binom{5}{0} \cdot a_{14} = (-1)^0 \binom{5}{0} \cdot \binom{14 + 4}{4} = 3060$$

(2). The first part take t^7, and the second part take t^7.
Then the coefficient is

$$(-1)^1 \binom{5}{1} \cdot a_7 = (-1)^1 \binom{5}{1} \cdot \binom{7 + 4}{4} = -1650$$

(3). The first part take t^{14}, and the second part take t^0.
Then the coefficient is

$$(-1)^2\binom{5}{2}\cdot a_0 = (-1)^2\binom{5}{2}\cdot\binom{0+4}{4} = 10$$

So above all, the coefficient of t^{14} is $3060 - 1650 + 10 = 1420$.
And the probability is $P(X = 14) = \dfrac{1420}{7^5}$.

The probability can be also decribe as $P(X = 14) = \dfrac{\#\ 5\ boxes\ contains\ 14\ balls}{\#\ 5\ boxes\ contain\ some\ balls} = \dfrac{\#\ 5\ boxes\ contains\ 14\ balls}{7^5}$
So the number of ways to let 5 boxes contain 14 balls is 1420.
Consider this issue in reverse, the number of the ways to put 14 balls into 5 boxes is the same as letting 5 boxes contain 14 balls, which is 1420.

So above all, the number of differnet ways to distribute these balls is 1420.