Problem 1

Let X have PMF

$$P(X = k) = cp^k/k \text{ for } k = 1, 2, \ldots,$$

where p is a parameter with $0 < p < 1$ and c is a normalizing constant. We have $c = -1/\log(1-p)$, as seen from the Taylor series

$$-\log(1-p) = p + \frac{p^2}{2} + \frac{p^3}{3} + \cdots$$

This distribution is called the Logarithmic distribution (because of the log in the above Taylor series), and has often been used in ecology. Find the mean and variance of X.

Solution

(a)

$$\begin{aligned} E[X] = & \sum_{k=1}^{\infty} k \cdot p(X = k) \\ = & \ c \cdot \sum_{k=1}^{\infty} k \cdot \frac{p^k}{k} \\ = & \ c \cdot \sum_{k=1}^{\infty} p^k \\ = & \ c \cdot \frac{p}{1-p} \\ = & \ -\frac{p}{(1-p)log(1-p)}. \end{aligned}$$

(b)

$$\begin{aligned} E[X^2] = & \sum_{k=1}^{\infty} k^2 \cdot c \cdot \frac{p^k}{k} \\ = & \sum_{k=1}^{\infty} kcp^k. \end{aligned}$$

As $\sum_{k=1}^{\infty} p^k = \frac{p}{1-p}$, we can get $\sum_{k=1}^{\infty} kp^{k-1} = \frac{1}{(1-p)^2}$.

$$\begin{aligned} E[X^2] = & \ cp \sum_{k=1}^{\infty} k \cdot p^{k-1} \\ = & \ \frac{cp}{(1-p)^2}. \end{aligned}$$

$$\begin{aligned} Var[X] = & \ E[X^2] - (E[X])^2 \\ = & \ \frac{cp(1-cp)}{(1-p)^2}. \end{aligned}$$

Problem 2

Let a random variable X satisfies Hypergeometric distribution with parameters w, b, n.

(a) Find $E\left[\begin{pmatrix} X \\ 2 \end{pmatrix}\right]$

(b) Use the result of (a) to find the variance of X.

Solution

(a) Consider an urn with w white balls and b black balls. We draw n balls out of the urn at random without replacement. Let X be the number of white balls in the sample. Then $X \sim \mathrm{HGeom}(w, b, n)$.

Let A_i be the event that the ith chosen ball is white, $i = 1, 2, \ldots, n$.

Let I_i be the indicator of A_i, $i = 1, 2, \ldots, n$.

Since $\binom{X}{2} = \sum_{i<j} I_i I_j$, we have

$$\mathrm{E}\left[\binom{X}{2}\right] = \sum_{i<j} P(A_i \cap A_j)$$

We consider the order of balls, then there are $\frac{(w+b)!}{(w+b-n)!}$ ways to choose n balls from $w + b$ balls, $\frac{w!}{(w-2)!}$ ways to choose 2 from the w white balls, and $\frac{(w+b-2)!}{(w+b-2-(n-2))!}$ ways to choose $n - 2$ balls from the rest $w + b - 2$ balls. Thus we have

$$P(A_i \cap A_j) = \frac{\frac{w!}{(w-2)!} \frac{(w+b-2)!}{(w+b-2-(n-2))!}}{\frac{(w+b)!}{(w+b-n)!}} = \frac{\binom{w}{2}\binom{w+b-2}{n-2}}{\binom{w+b}{n}\binom{n}{2}}.$$

Or by symmetry, $P(A_i \cap A_j) = P(A_1 \cap A_2) = P(A_1) P(A_2 | A_1) = \frac{w}{w+b} \cdot \frac{w-1}{w+b-1}$ for all $i \neq j$.

It follows that

$$\mathrm{E}\left[\binom{X}{2}\right] = \binom{n}{2} \cdot \frac{\binom{w}{2}\binom{w+b-2}{n-2}}{\binom{w+b}{n}\binom{n}{2}} = \frac{\binom{w}{2}\binom{w+b-2}{n-2}}{\binom{w+b}{n}}.$$

(b) We have

$$\mathrm{E}\left[\binom{X}{2}\right] = \mathrm{E}\left(\frac{X(X-1)}{2}\right) = \frac{1}{2}\mathrm{E}(X^2) - \frac{1}{2}\mathrm{E}(X), \quad \mathrm{E}(X) = n \cdot \frac{w}{w+b}$$

By combining the above two equations we have

$$\mathrm{E}(X^2) = 2\frac{\binom{w}{2}\binom{w+b-2}{n-2}}{\binom{w+b}{n}} + \frac{nw}{w+b}$$

Then we have

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = 2\frac{\binom{w}{2}\binom{w+b-2}{n-2}}{\binom{w+b}{n}} + \frac{nw}{w+b} - \left(\frac{nw}{w+b}\right)^2.$$

Problem 3

Suppose there are n types of toys, which you are collecting one by one, with the goal of getting a complete set. When collecting toys, the toy types are random. Assume that each time you collect a toy, it is equally likely to be any of the n types. Let N denote the number of toys needed until you have a complete set. Find $\text{Var}(N)$.

Solution

Consider a state where the collector has already collected m coupons. Let N_m denote the number of coupons does he need to collect to get to $m+1$ type. Then, if the total coupons needed is N, we have:

$$N = \sum_{m=1}^{n} N_m$$

Every coupon collected from here is like a coin toss where with probability $\frac{m}{n}$, the collector hits a coupon he already has and makes no progress. With probability $\frac{n-m}{n}$, he collects a new coupon. So, this becomes a geometric random variable with $p = \frac{n-m}{n}$. We know that a geometric random variable has a mean $\frac{1}{p}$ and variance $\frac{1-p}{p^2}$. Hence,

$$E\left(N_m\right) = \frac{n}{n-m}$$

Further, we have:

$$E(N) = E\left(N_m\right) = \sum_{m=1}^{n} \frac{n}{n-m} = n \sum_{m=1}^{n} \frac{1}{n-m}$$

Substituting $m = n - m$ we get:

$$E(N) = n \sum_{m=1}^{n} \frac{1}{m}$$

Since the random variables N_m are independent, the variance of their sum is equal to the sum of their variances. Therefore, similar to the reasoning for expectation, we have the variance as follows:

$$\text{Var}(N) = n^2 \sum_{i=1}^{n} \frac{1}{i^2} - n \sum_{k=1}^{n} \frac{1}{k}.$$

Problem 4

Suppose there are n types of toys, which you are collecting one by one, with the goal of getting a complete set. When collecting toys, the toy types are random. Assume that each time you collect a toy, independently of past types collected, it is type j with probability p_j, and $\sum_{j=1}^{n} p_j = 1$. Let N denote the number of different types of toys that appear among the first m collected toys. Find $E(N)$ and $\mathrm{Var}(N)$.

Solution

Let A_i be the event that type i coupon appear among the first m collected, $i = 1, 2, ..., n$.
Let I_i be the indicator of A_i, $i = 1, 2, ..., n$.
Let N be the number of different types of coupons that appear among the first m collected.
Then we have $N = I_1 + I_2 + \cdots + I_n$, it follows that

$$\mathrm{E}(N) = \mathrm{E}(I_1 + I_2 + \cdots + I_n) = \sum_{i=1}^{n} \mathrm{E}(I_i) = \sum_{i=1}^{n} P(A_i).$$

Since $P(A_i) = 1 - P(A_i^c) = 1 - (1 - p_i)^m$, we get

$$\mathrm{E}(N) = \sum_{i=1}^{n} \left[1 - (1 - p_i)^m\right].$$

Because

$$\mathrm{E}\left[\binom{N}{2}\right] = \mathrm{E}\left[\sum_{1 \le i < j \le n} I_i I_j\right] = \sum_{1 \le i < j \le n} \mathrm{E}[I_i I_j] = \sum_{1 \le i < j \le n} P(A_i \cap A_j),$$

in which

$$\begin{aligned} P(A_i \cap A_j) &= P(A_i) + P(A_j) - P(A_i \cup A_j) = P(A_i) + P(A_j) - \left(1 - P\left(A_i^c \cap A_j^c\right)\right) \\ &= (1 - (1 - p_i)^m) + (1 - (1 - p_j)^m) - (1 - (1 - p_i - p_j)^m) \\ &= 1 - (1 - p_i)^m - (1 - p_j)^m + (1 - p_i - p_j)^m, \end{aligned}$$

we have

$$\mathrm{E}\left[\binom{N}{2}\right] = \sum_{i=1}^{n} \left[1 - (1 - p_i)^m - (1 - p_j)^m + (1 - p_i - p_j)^m\right].$$

Using $\mathrm{E}\left[\binom{N}{2}\right] = \mathrm{E}\left[\frac{N(N-1)}{2}\right]$, we get

$$\mathrm{E}\left(N^2\right) = 2\mathrm{E}\left[\binom{N}{2}\right] + \mathrm{E}[N],$$

it follows that

$$\begin{aligned} &\mathrm{Var}(N) \\ =&\mathrm{E}\left(N^2\right) - \mathrm{E}^2(N) \\ =&2\mathrm{E}\left[\binom{N}{2}\right] + \mathrm{E}[N] - \mathrm{E}^2(N) \\ =&2\sum_{i=1}^{n} \left[1 - (1 - p_i)^m - (1 - p_j)^m + (1 - p_i - p_j)^m\right] + \sum_{i=1}^{n} \left[1 - (1 - p_i)^m\right] - (\sum_{i=1}^{n} \left[1 - (1 - p_i)^m\right])^2. \end{aligned}$$

Problem 5

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 an r.v. 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 r.v. 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.

(d) Find the variance of X, both in terms of the p_j 's and numerically.

Hint:What is I_i^2, and what is I_iI_j for $i < j$? Use this to simplify the expansion

$$X^2 = I_1^2 + \cdots + I_{366}^2 + 2\sum_{j=2}^{366}\sum_{i=1}^{j-1} I_iI_j$$

Solution

(a) For an arbitrary pair of people, the probability of having the same birthday is $1/365$. It is denoted that the number of birthday match is Z. Since in the corresponding number of samples is relatively large and the probability is small, we have

$$P(\text{At least one birthday match}) = P(Z \geq 1) = 1 - P(Z = 0) \approx 1 - \mathrm{e}^{\lambda}, \tag{1}$$

where $\lambda = \binom{m}{2}p$, m is the number of people, and p is the probability. Therefore, we have

$$P(X \leq 23) \approx 1 - \mathrm{e}^{\lambda} \approx 0.5002 \geq 0.5. \tag{2}$$

On the other hand, we have

$$\begin{aligned} P(X \geq 23) =& P(\text{No match before 23}) \approx \mathrm{e}^{\lambda} \\ =& \mathrm{e}^{\binom{22}{2}\cdot\frac{1}{365}} \approx 0.531 \geq 0. \end{aligned} \tag{3}$$

Thus, 23 is the unique median of X.

(b) For X, it can always be expressed with the sum of binary indicators since it is not decreasing. Then we have

$$\begin{aligned} E(X) =& E(I_1 + I_2 + \cdots + I_{366}) \\ =& E(I_1) + E(I_2) + \cdots + E(I_{366}) \\ =& \sum_{i=1}^{366} p_j. \end{aligned} \tag{4}$$

Problem 5 continued on next page...

(c)

$$\begin{aligned}E(X) =&\sum_{i=1}^{366} p_j \\ =&1+1+\left(1-\frac{1}{365}\right)+\left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)+\cdots+\left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\cdots\left(1-\frac{364}{365}\right) \\ \approx&24.62\end{aligned} \tag{5}$$

(d)

$$\begin{aligned}E(X^2) =&E\left(I_1^2+I_2^2+I_3^2+...+I_{366}^2+2\sum_{j=2}^{366}\sum_{i=1}^{j-1} I_i I_j\right) \\ =&E(I_1^2)+E(I_2^2)+E(I_3^2)+...+E(I_{366}^2)+2E\left(\sum_{j=2}^{366}\sum_{i=1}^{j-1} I_j\right) \\ =&\sum_{j=1}^{366} p_j+2\sum_{j=1}^{366}(j-1)E(I_j) \\ =&\sum_{j=1}^{366}(2j-1)p_j.\end{aligned} \tag{6}$$

Thus, we have

$$D(X)=E(X^2)-[E(X)]^2\approx 148.64. \tag{7}$$

Problem 6

(Optional Challenging Problem) Suppose there are n types of toys, which you are collecting one by one, with the goal of getting a complete set. When collecting toys, the toy types are random. Assume that each time you collect a toy, independently of past types collected, it is type j with probability p_j, and $\sum_{j=1}^{n} p_j = 1$. Let N denote the number of toys needed until you have a complete set. Find $E(N)$ and $\operatorname{Var}(N)$.

Solution

Check the paper: "The mean and variance in coupons required to complete a collection" and Example 5.17 in the book "Introduction to Probability Models".

Calculate the Expectation via Poisson Process

Let us imagine that the collector collects the coupons in accordance to a Poisson process with rate $\lambda = 1$. Furthermore, every coupon that arrives is of type j with probability p_j. Define X_j as the first time a coupon of type j is observed, if the j th coupon arrives in accordance to a Poisson process with rate p_j.

We're interested in the time it takes to collect all coupons, X (for now, eventually, we're interested in the number of coupons to be collected, N). So we get:

$$X = \max_{1 \le j \le n} X_j.$$

Note that if we denote N_j as the number of coupons to be collected before the first coupon of type j is seen, we also have for the number needed to collect all coupons, N :

$$N = \max_{1 \le j \le n} N_j$$

This equation is less useful since the N_j are not independent. It can still be used to get the mean, but trying to get the variance with this approach gets considerably more challenging due to this dependence of the underlying random variables.

But, the incredible fact that the X_j are independent (by the property of Poisson process), allows us to get:

$$F_X(t) = P(X < t) = P\left(X_j < t, \forall j\right) = \prod_{j=1}^{n} \left(1 - e^{-p_j t}\right).$$

Therefore,

$$E(X) = \int_0^\infty \left(1 - \prod_{j=1}^{n} \left(1 - e^{-p_j t}\right)\right) dt = \sum_j \frac{1}{p_j} - \sum_{i<j} \frac{1}{p_i + p_j} + \cdots + (-1)^{n-1} \frac{1}{p_1 + \cdots + p_n},$$

where the equality by 1) the fact that $\int_0^\infty e^{-pt} dt = \frac{1}{p}$ and 2) the inclusion-exclusion principle on the product, i.e., treating product as intersection and summation as union.

It remains to relate $E[X]$, the expected time until one has a complete set, to $E[N]$, the expected number of coupons it takes. This can be done by letting T_i denote the i th inter-arrival time of the Poisson process that counts the number of coupons obtained. Then it is easy to see that

$$X = \sum_{i=1}^{N} T_i.$$

Since the T_i are independent exponentials with rate 1 , and N is independent of the T_i, we see that

$$E[X \mid N] = N E\left[T_i\right] = N$$

Therefore,

$$E[X] = E[N]$$

Problem 6 continued on next page...

Extension to Variance via Survival Function
This approach can easily be extended to find $\operatorname{Var}(N)$, the variance. We can use the following expression to get $E\left(X^2\right)$ via survival function:

$$E\left(X^2\right)=\int_0^{\infty} 2tP(X>t)dt=\int_0^{\infty} 2t\left(1-\prod_{j=1}^{n}\left(1-e^{-p_j t}\right)\right)dt$$

Using the fact that $\int_0^{\infty} te^{-pt}=\frac{1}{p^2}$ and the same algebra as for $E(X)$ we get:

$$\frac{E\left(X^2\right)}{2}=\sum_j \frac{1}{p_j^2}-\sum_{i<j}\frac{1}{\left(p_i+p_j\right)^2}+\cdots+(-1)^{n-1}\frac{1}{\left(p_1+\cdots+p_n\right)^2}$$

Using the law of total variance we get:

$$\operatorname{Var}(X)=E(\operatorname{Var}(X \mid N))+\operatorname{Var}(E(X \mid N))=E(\operatorname{Var}(X \mid N))+\operatorname{Var}(N)$$

Now, we have

$$\operatorname{Var}(X \mid N)=N \operatorname{Var}\left(T_i\right)$$

and since $T_i \sim \operatorname{Exp}(1)$, we have $\operatorname{Var}\left(T_i\right)=1$ meaning, $\operatorname{Var}(X \mid N)=N$. Therefore, we have

$$\operatorname{Var}(X)=E(N)+\operatorname{Var}(N).$$

Accordingly, we have

$$\operatorname{Var}(N)=E\left(X^2\right)-E(X)^2-E(N)=E\left(X^2\right)-E(N)^2-E(N)$$

Finally, we have

$$\begin{aligned}\operatorname{Var}(N)=2&\left(\sum_j \frac{1}{p_j^2}-\sum_{i<j}\frac{1}{\left(p_i+p_j\right)^2}+\cdots+(-1)^{n-1}\frac{1}{\left(p_1+\cdots+p_n\right)^2}\right)-\\&\left(\sum_j \frac{1}{p_j}-\sum_{i<j}\frac{1}{\left(p_i+p_j\right)}+\cdots+(-1)^{n-1}\frac{1}{\left(p_1+\cdots+p_n\right)}\right)^2-\\&\left(\sum_j \frac{1}{p_j}-\sum_{i<j}\frac{1}{\left(p_i+p_j\right)}+\cdots+(-1)^{n-1}\frac{1}{\left(p_1+\cdots+p_n\right)}\right).\end{aligned}$$