Yale University
S&DS 551, Spring 2023
Homework 1

Chang Feng (Felix) Zhou cz397

Problem 1.

(1) Mean

Let $X_1, \dots, X_{n-k}$ be random variables denoting the value of each of the $n-k$ remaining balls and $X = \sum_{i=1}^{n-k} X_i$ be their sum. By the linearity of expectation,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{n-k} \mathbb{E}[X_i] \\
&= \boxed{(n-k)\frac{n+1}{2}}.
\end{aligned}$$

(2) Variance

We compute the second moment in similar fashion. Indeed,

$$\begin{aligned}
\mathbb{E}[X^2] &= \sum_{i,j=1}^{n-k} \mathbb{E}[X_i X_j] \\
&= \sum_{\ell=1}^{n-k} \mathbb{E}[X_\ell^2] + 2 \sum_{1 \le i < j \le n-k} \mathbb{E}[X_i X_j].
\end{aligned}$$

For any $\ell \in [n-k]$,

$$\begin{aligned}
\mathbb{E}[X_\ell^2] &= \frac{1}{n} \sum_{i=1}^{n} i^2 \\
&= \frac{(n+1)(2n+1)}{6} \\
\sum_{\ell=1}^{n-k} \mathbb{E}[X_\ell^2] &= \frac{(n-k)(n+1)(2n+1)}{6}.
\end{aligned}$$

Now, there are $\binom{n}{2}$ possible pairs $1 \le i < j \le n$ and $\binom{n-k}{2}$ pairs of indices from $[n-k]$. It follows that

$$
\begin{aligned}
2 \sum_{1 \le i < j \le n-k} \mathbb{E}[X_i X_j] &= \frac{\binom{n-k}{2}}{\binom{n}{2}} \sum_{1 \le i, j \le n : i \ne j} ij \\
&= \frac{(n-k)(n-k-1)}{n(n-1)} \left(\sum_{1 \le i, j \le n} ij - \sum_{\ell=1}^{n} \ell^2 \right) \\
&= \frac{(n-k)(n-k-1)}{n(n-1)} \left(\left[\sum_{i=1} i \right]^2 - \sum_{j=1}^{n} j^2 \right) \\
&= \frac{(n-k)(n-k-1)}{n(n-1)} \left(\frac{n^2(n+1)^2}{4} - \frac{n(n+1)(2n+1)}{6} \right) \\
&= \frac{(n-k)(n-k-1)}{n(n-1)} \left(\frac{n(n+1)[3n(n+1) - 2(2n+1)]}{12} \right) \\
&= \frac{(n-k)(n-k-1)}{n(n-1)} \left(\frac{n(n+1)(3n^2 - n - 2)}{12} \right).
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \frac{(n-k)(n+1)(2n+1)}{6} + \frac{(n-k)(n-k-1)}{n(n-1)} \left(\frac{n(n+1)(3n^2 - n - 2)}{12} \right) - (n-k)^2 \frac{(n+1)^2}{4} \\
&= \boxed{\frac{k(n+1)(n-k)}{12}}.
\end{aligned}
$$

Problem 2.

First consider the special case of $m = n$. Clearly, we must have

$$\mathbb{E}[S_m/S_n] = 1.$$

Note we use the assumption here that $\mathbb{E}[X_1^{-1}]$ exists.

By the linearity of expectation,

$$\mathbb{E}[S_m/S_n] = \sum_{i=1}^{m} \mathbb{E}[X_i/S_n].$$

However, since each X_i is iid, we conclude that

$$\mathbb{E}[X_i/S_n] = \frac{1}{n}$$

so that

$$\mathbb{E}[S_m/S_n] = \frac{m}{n}$$

as desired.

Problem 3.

We remark that

$$\begin{aligned}\mathbb{P}\{N > n\} &= \mathbb{P}\left\{\sum_{i=1}^{n} X_i \le x\right\} \\ &=: p_{n,x}.\end{aligned}$$

We argue by induction that $p_{n,x} = \frac{x^n}{n!}$. It is clear that $p_{0,x} = 1$ since $x > 0$. Now suppose the induction hypothesis holds up to $n-1$. We have

$$\begin{aligned}p_n &= \int_0^x \mathbb{P}\left\{\sum_{i=1}^{n-1} X_i \le x - z\right\} p(z)dz && p(z) \text{ density of } U[0,1] \\ &= \int_0^x p_{n,x-z}dz \\ &= \int_0^x \frac{(x-z)^{n-1}}{(n-1)!} \\ &= \left[-\frac{(x-z)^n}{n!}\right]_0^x \\ &= \frac{x^n}{n!}.\end{aligned}$$

By induction, we conclude the proof.

In order to compute the expectation, recall the identity

$$\begin{aligned}\mathbb{E}[N] &= \sum_{n=1}^{\infty} \mathbb{P}\{N = n\} \cdot n \\ &= \sum_{n=1}^{\infty} \mathbb{P}\{N \ge n\} \\ &= \sum_{n=0}^{\infty} \mathbb{P}\{N > n\} \\ &= \sum_{n=0}^{\infty} \frac{x^n}{n!} \\ &= \boxed{e^x}.\end{aligned}$$

We can similarly compute the variance as

$$\begin{aligned}
\mathbb{E}[N^2] &= \sum_{n=1}^{\infty} \mathbb{P}\{N = n\} \cdot n^2 \\
&= \sum_{n=1}^{\infty} \mathbb{P}\{N \geq n\}[n^2 - (n-1)^2] \\
&= \sum_{n=0}^{\infty} \mathbb{P}\{N > n\}[2n - 1] \\
&= 2\sum_{n=1}^{\infty} \mathbb{P}\{N > n\} \cdot n - \sum_{n=0}^{\infty} \mathbb{P}\{N > n\} \\
&= 2x \sum_{n=0}^{\infty} \frac{x^n}{n!} - e^x \\
&= e^x(2x - 1).
\end{aligned}$$

It follows that the variance is

$$\mathrm{Var}[N] = \boxed{e^x(2x-1) - e^{2x}}.$$

Problem 4.

let us determine the density of XY, Z^2 separately and recall that the joint density is just the product of densities since XY, Z^2 are independent.

Let F_ξ, p_ξ denote the distribution function and density of a random variable ξ, respectively.

We have

$$\begin{aligned} F_{XY}(t) &= \mathbb{P}\{X \le t\} + \int_t^1 \mathbb{P}\{Y \le t/x\} p_X(x) dx \\ &= t + \int_t^1 \frac{t}{x} dx \\ &= t + [t \log x]_t^1 \\ &= t - t \log t. && t \in [0,1] \end{aligned}$$

It follows that

$$\begin{aligned} p_{XY}(t) &= \frac{d}{dt} F_{XY}(t) \\ &= 1 - \log t - 1 \\ &= -\log t. && t \in [0,1] \end{aligned}$$

Similarly,

$$\begin{aligned} F_{Z^2}(t) &= \sqrt{t} && t \in [0,1] \\ p_{Z^2}(t) &= \frac{1}{2\sqrt{t}}. && t \in (0,1] \end{aligned}$$

The joint density for $t_1, t_2 \in [0,1], t_2 \neq 0$ is thus

$$\boxed{p_{XY,Z^2}(t_1, t_2) = -\frac{\log t_1}{2\sqrt{t_2}}}.$$

We now compute the desired probability

$$\begin{aligned} \mathbb{P}\{XY < Z^2\} &= \int_0^1 \mathbb{P}\{XY \le t\} p_{Z^2}(t) dt \\ &= \int_0^1 (t - t \log t) \frac{1}{2\sqrt{t}} dt \\ &= \frac{1}{2} \int_0^1 \sqrt{t} - \sqrt{t} \log t dt \\ &= \left[-\frac{t^{3/2}(3 \log x - 5)}{9} \right]_0^1 \\ &= \frac{5}{9} \end{aligned}$$

as desired.

Problem 5.

(1)

Write N to be the random variable indicating the number of draws until we stop. Clearly for $n \leq 3$,

$$\mathbb{P}\{N = n\} = 0.$$

Let us compute the distribution function $F(n) := \mathbb{P}\{N \leq n\}$. This the probability that after n draws, we have at least one card from each shade. This is equal to 1 minus the probability that we see at most 3 shades. Let p_1, p_2, p_3 denote the probability we do not see 1, 2, 3 shades for some FIXED shades (it is symmetric so it does not matter the particular choice of shades), respectively. By the inclusion-exclusion principle, the desired probability is thus

$$\begin{aligned} F(n) &= 1 - \binom{4}{1}p_1 + \binom{4}{2}p_2 - \binom{4}{3}p_3 \\ &= 1 - 4\left(\frac{3}{4}\right)^n + 6\left(\frac{2}{4}\right)^n - 4\left(\frac{1}{4}\right)^n. \qquad n \geq 4 \end{aligned}$$

This is because we need to count for the $\binom{4}{1}$ ways we avoid a particular shade, but avoid double counting the $\binom{4}{2}$ ways we avoid a particular pair of shades, and adjust for the overcorrection of the $\binom{4}{3}$ ways we avoid any triple of shades.

Then for $n \geq 4$,

$$\begin{aligned} \mathbb{P}\{N = n\} &= F(n) - F(n-1) \\ &= -4\left(\frac{3}{4}\right)^n + 6\left(\frac{2}{4}\right)^n - 4\left(\frac{1}{4}\right)^n + 4\left(\frac{3}{4}\right)^{n-1} - 6\left(\frac{2}{4}\right)^{n-1} + 4\left(\frac{1}{4}\right)^{n-1} \\ &= \boxed{\left(\frac{3}{4}\right)^{n-1} - 3\left(\frac{2}{4}\right)^{n-1} + 3\left(\frac{1}{4}\right)^{n-1}}. \end{aligned}$$

(2)

It is easy to see the that the sum of probabilities tend to 1 since we computed the individual probabilities as the difference between consecutive values of the distribution function. Thus the partial sum up to n is just $F(n)$ and $F(n) \to 1$ as $n \to \infty$ by inspection.