S&DS 351 Homework 1 Solutions

Staff

January 2023

Problem 1

> An urn contains n balls numbered $1, 2, \ldots, n$. We remove k balls at random (without replacement) and add up the numbers of all remaining balls. Find the mean and variance of the total sum.

Let $X_1, \ldots, X_{n-k}$ be random variables denoting the value of the remaining $n-k$ balls. We are interested in the sum of them, $S := \sum_{i=1}^{n-k} X_i$. Thus,

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{i=1}^{n-k} X_i\right] = \sum_{i=1}^{n-k} \mathbb{E}[X_i]$$

It remains to calculate the expectation of X_i. Clearly, X_i takes values in $\{1, 2, \ldots, n\}$ with uniform probability $1/n$ for each. Therefore,

$$\mathbb{E}[S] = \sum_{i=1}^{n-k} \mathbb{E}[X_i] = \sum_{i=1}^{n-k} \sum_{j=1}^{n} j \cdot \frac{1}{n} = \sum_{i=1}^{n-k} \frac{n(n+1)}{2} \cdot \frac{1}{n} = \frac{(n-k)(n+1)}{2}$$

Now, for the variance. For this problem, we will have to be a bit more cautious about book-keeping. Recall that $\mathrm{Var}(S) = \mathbb{E}\left[S^2\right] - (\mathbb{E}[S])^2$. We will calculate each of these terms individually. First,

$$\mathbb{E}\left[S^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n-k} X_i\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^{n-k}\sum_{j=1}^{n-k} X_i X_j\right] = \sum_{i=1}^{n-k}\sum_{j=1}^{n-k} \mathbb{E}\left[X_i X_j\right]$$

We will consider two cases. In the first, suppose $i = j$. Remember that X_i takes values in $\{1, 2, \ldots, n\}$ with uniform probability $1/n$ for each. Now, recalling the sum identity $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$,

$$\mathbb{E}[X_i^2] = \sum_{i=1}^{n} i^2 \cdot \frac{1}{n} = \frac{(n+1)(2n+1)}{6}$$

When $j \neq i$, we have,

$$\begin{aligned}
\mathbb{E}\left[X_i X_j\right] &= \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j\neq i} ij = \frac{1}{n(n-1)}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} ij - \sum_{i=1}^{n} i^2\right) = \frac{1}{n(n-1)}\left(\left(\sum_{i=1}^{n} i\right)^2 - \sum_{i=1}^{n} i^2\right) \\
&= \frac{1}{n(n-1)}\left(\frac{n^2(n+1)^2}{4} - \frac{n(n+1)(2n+1)}{6}\right) = \frac{n+1}{n-1}\left(\frac{n(n+1)}{4} - \frac{2n+1}{6}\right) \\
&= \frac{(n+1)(3n^2-n-2)}{12(n-1)} = \frac{(n+1)(n-1)(3n+2)}{12(n-1)} = \frac{(n+1)(3n+2)}{12}
\end{aligned}$$

And therefore,

$$\begin{aligned}
\mathbb{E}\left[S^2\right] &= \sum_{i=1}^{n-k}\sum_{j\neq i}\mathbb{E}[X_iX_j] + \sum_{i=1}^{n-k}\mathbb{E}[X_i]^2 \\
&= \frac{(n-k)(n-k-1)(n+1)(3n+2)}{12} + \frac{(n-k)(n+1)(2n+1)}{6} \\
&= \frac{(n-k)(n+1)}{12}\Big((n-k-1)(3n+2) + 2(2n+1)\Big) \\
&= \frac{(n-k)(n+1)}{12}\Big((n-k)(3n+2) + n\Big)
\end{aligned}$$

Thus

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

Problem 2

> Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables for which $\mathbb{E}(X_1^{-1})$ exists. Show that, if $m \leqslant n$, then $\mathbb{E}(S_m/S_n) = m/n$, where $S_m = X_1 + X_2 + \ldots + X_m$.

Note that, by symmetry, $\mathbb{E}[X_i/S_n] = \mathbb{E}[X_1/S_n]$ for all i, and thus,

$$1 = \mathbb{E}[1] = \mathbb{E}\left[\frac{S_n}{S_n}\right] = \sum_{i=1}^{n}\mathbb{E}\left[\frac{X_i}{S_n}\right] = n\cdot\mathbb{E}\left[\frac{X_1}{S_n}\right]$$

Therefore, $\mathbb{E}[X_i/S_n] = 1/n$ for all i, and,

$$\mathbb{E}\left[\frac{S_m}{S_n}\right] = \mathbb{E}\left[\sum_{i=1}^{m}\frac{X_i}{S_n}\right] = \sum_{i=1}^{m}\mathbb{E}\left[\frac{X_i}{S_n}\right] = \sum_{i=1}^{m}\frac{1}{n} = \frac{m}{n}$$

Completing the proof.

Problem 3

> Let $\{X_r : r \geqslant 1\}$ be independent and uniformly distributed on the interval $[0,1]$. Let $0 < x < 1$ and define
>
> $$N = \min\{n \geqslant 1 :\; X_1 + X_2 + \ldots + X_n > x\}.$$
>
> Show that $\mathbb{P}(N > n) = x^n/n!$, and hence find the mean and variance of N.

Let x and N be as described. We shall prove this by induction that $\mathbb{P}(S_n \le x) = x^n/n!$. First, note this is obviously true of $n = 1$. Now, assume this is true of $n-1$. Letting $f(z) = 1$ denote the probability density function of X_n, we have,

$$\mathbb{P}(S_n \le x) = \int_0^x \mathbb{P}(S_{n-1} \le x - s)f(s)ds = \int_0^x \frac{(x-s)^{n-1}}{(n-1)!}ds = \frac{x^n}{n!}$$

So by induction, it is true of all n. Furthermore, $\mathbb{P}(S_n \leq x) = \mathbb{P}(N > n)$, we have verified the first property.

Recall also that for positive integer valued random variables, $\mathbb{E}[X] = \sum_{n=0}^{\infty} \mathbb{P}(X > n)$. Therefore, $\mathbb{E}[X] = \sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$. It simply remains to find variance. Note that, $\text{Var}(N) = \mathbb{E}\left[N^2\right] - (\mathbb{E}[N])^2$.

$$\mathbb{E}[N^2] = \sum_{n=0}^{\infty} \mathbb{P}\left(N^2 > n\right) = \sum_{n=0}^{\infty} \mathbb{P}\left(N > \lfloor\sqrt{n}\rfloor\right) = \sum_{n=0}^{\infty} \frac{x^{\lfloor\sqrt{n}\rfloor}}{(\lfloor\sqrt{n}\rfloor)!}$$

It is not difficult to show that for a given number $m \in \mathbb{Z}$, $m = \lfloor\sqrt{n}\rfloor$ for $2m+1$ numbers n. Indeed, solving for $m^2 \leq n < (m+1)^2$, this becomes $0 \leq n - m^2 \leq (m+1)^2 - m^2 = 2m+1$, which is true for $2m+1$ such n. Thus, accounting for multiplicity,

$$\mathbb{E}[N^2] = \sum_{m=0}^{\infty} (2m+1)\frac{x^m}{m!} = 2\sum_{m=0} \frac{mx^m}{m!} + \sum_{m=0}^{\infty} \frac{x^m}{m!} = 2\sum_{m=1}^{\infty} \frac{x^m}{(m-1)!} + e^x = 2x\sum_{j=0}^{\infty} \frac{x^j}{j!} + e^x = 2xe^x + e^x$$

Thus, $\text{Var}(N) = 2xe^x + e^x - e^{2x} = e^x\left(2x + 1 - e^x\right)$.

Problem 4

Let X, Y and Z be independent and uniformly distributed on the interval $[0, 1]$. Find the joint density function of XY and Z^2, and show that $\mathbb{P}(XY < Z^2) = \frac{5}{9}$.

For now, let $\overline{Y} = 1/Y$; note that $\overline{Y} \in [1, \infty)$. Calculating the density function of $\overline{Y}$, observe $\mathbb{P}(\overline{Y} \leq t) = \mathbb{P}(Y \geq 1/t) = 1 - \mathbb{P}(Y < 1/t) = 1 - \mathbb{P}(Y \leq 1/t) = 1 - (1/t)$, where this holds for $1 \leq t < \infty$. Taking the derivative, we have that $f_{\overline{Y}}(t) = \frac{1}{t^2}$, for t. And thus,

$$\mathbb{P}(XY \leq t) = \mathbb{P}\left(\frac{1}{t}X \leq \overline{Y}\right) = \int_1^{\infty} \mathbb{P}\left(\frac{1}{t}X \leq s\right) f_{\overline{Y}}(s)ds = \int_1^{\infty} \mathbb{P}(X \leq ts)\left(\frac{1}{s^2}\right) ds$$

Note that $\mathbb{P}\left(X \leq ts\right)$ depends on the value of s. If $ts \geq 1$, ie $s \geq 1/t$, this value is 1. Otherwise, the value is equal to ts (as we assume $t \geq 0$). Therefore,

$$\begin{aligned}\mathbb{P}(XY \leq t) &= \int_1^{1/t} \frac{ts}{s^2} ds + \int_{1/t}^{\infty} \frac{1}{s^2} ds = t\log(s)\Big|_1^{1/t} - \frac{1}{s}\Big|_{1/t}^{\infty} \\ &= t\log(1/t) - t\log(1) + \frac{1}{1/t} = -t\log(t) + t = t(1 - \log(t))\end{aligned}$$

Where this holds for $0 \leq t \leq 1$. Otherwise, for $t \leq 0$, the value is 0, and for $t \geq 1$, the value is 1. Now, to obtain a density, we have, for $0 \leq t \leq 1$,

$$f_{XY}(t) = \frac{d}{dt}\mathbb{P}(XY \leq t) = \frac{d}{dt}(t - t\log(t)) = 1 - (1 + \log(t)) = -\log(t)$$

Now we need the density function of Z^2. Observe that,

$$f_{Z^2}(t) = \frac{d}{dt}\mathbb{P}(Z^2 \leq t) = \frac{d}{dt}\mathbb{P}(Z \leq \sqrt{t}) = \frac{d}{dt}\sqrt{t} = \frac{1}{2\sqrt{t}}$$

Therefore, by independence, their joint density function is the product of the individual density functions,

$$f_{XY,Z^2}(s,t) = f_{XY}(s)f_{Z^2}(t) = \frac{\log(1/s)}{2\sqrt{t}}$$

Therefore, to calculate $\mathbb{P}(XY \le Z^2)$, we simply integrate over the region where this holds:

$$\begin{aligned}\mathbb{P}(XY \le Z^2) &= \int_0^1 \int_0^t f_{XY,Z^2}(s,t)dsdt = \int_0^1 \int_0^t \frac{-\log(s)}{2\sqrt{t}}dsdt \\ &= -\frac{1}{2}\int_0^1 \frac{1}{\sqrt{t}}\int_0^t \log(s)dsdt = -\frac{1}{2}\int_0^1 \frac{1}{\sqrt{t}}(s\log(s)-s)\Big|_0^t dt\end{aligned}$$

We adopt the convention that $s\log(s) = 0$ at 0, by continuity (this is common practice in, say, information theory), so that we may write:

$$\mathbb{P}(XY \le Z^2) = -\frac{1}{2}\int_0^1 \frac{t\log(t)-t}{\sqrt{t}}dt = -\frac{1}{2}\int_0^1 \left(\sqrt{t}\log(t) - \sqrt{t}\right)dt$$

Doing the integral,

$$\mathbb{P}(XY \le Z^2) = -\frac{1}{2}\left(\frac{2}{3}t^{3/2}\log(t) - \frac{10}{9}t^{3/2}\right)\Bigg|_0^1 = -\frac{1}{2}\left(\frac{2}{3}1\log(1) - \frac{10}{9}(1)\right) = \frac{5}{9}$$

So we are done!

Problem 5

From a set of 52 poker cards (without 2 jokers), we keep taking cards randomly one by one with replacement, until all the cards taken by us can cover all 4 shades.

(1) Compute the probability that we have picked exactly n cards.

(2) Verify that, after taking summation over $n = 1, 2, \ldots$, the sum of the probabilities above equals to 1.

(1) Let N be the precise number of cards you need to pick up to get a representative from every suit. Also, let X_n be the suit of the nth draw. Observe that it suffices to analyze these probabilities, given we complete our collection with a club:

$$\begin{aligned}\mathbb{P}(N=n) &= 4\cdot\mathbb{P}(N=n, X_n = \clubsuit) \\ &= \mathbb{P}(X_n = \clubsuit \mid \{X_1,\ldots,X_{n-1}\} = \{\diamondsuit, \spadesuit, \heartsuit\})\mathbb{P}(\{X_1,\ldots,X_{n-1}\} = \{\diamondsuit, \spadesuit, \heartsuit\})\end{aligned}$$

Obviously, $\mathbb{P}(X_n = \clubsuit \mid \{X_1,\ldots,X_{n-1}\} = \{\diamondsuit, \spadesuit, \heartsuit\}) = \mathbb{P}(X_n = \clubsuit) = 1/4$, since we are sampling with replacement. It remains to calculate the probability that our first $n-1$ draws collect at least a spade, a heart, a diamond, and no club. Indeed, we can try to count the number of sequences of length $n-1$ that include at least one spade, one heart, one diamond, and no club. Clearly, there are 3^{n-1} total possible strings. 2^{n-1} are just diamonds and hearts, 2^{n-1} are just diamonds and spades, and 2^{n-1} are just spades and hearts. But we double count the 3 strings consisting of all hearts, all spades, and all diamonds. Thus, the total number of nice strings is $3^{n-1} - 3\cdot 2^{n-1} + 3$. And they all have equal probability: $1/4^{n-1}$. Thus,

$$\mathbb{P}(N=n) = 4\cdot\frac{1}{4}\cdot\frac{1}{4^{n-1}}\left(3^{n-1} - 3\cdot 2^{n-1} + 3\right) = \left(\frac{3}{4}\right)^{n-1} - 3\left(\frac{1}{2}\right)^{n-1} + 3\left(\frac{1}{4}\right)^{n-1}$$

(2) If we sum from $n = 4$ to ∞, we find,

$$\begin{aligned}
\sum_{n=4}^{\infty} \mathbb{P}(N = n) &= \sum_{n=4}^{\infty} \left(\frac{3}{4}\right)^{n-1} - 3\sum_{n=4}^{\infty} \left(\frac{1}{2}\right)^{n-1} + 3\sum_{n=4}^{\infty} \left(\frac{1}{4}\right)^{n-1} \\
&= \left(\frac{3}{4}\right)^{3} \sum_{n=0}^{\infty} \left(\frac{3}{4}\right)^{n} - 3\left(\frac{1}{2}\right)^{3} \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^{n} + 3\left(\frac{1}{4}\right)^{3} \sum_{n=0}^{\infty} \left(\frac{1}{4}\right)^{n} \\
&= \left(\frac{27}{64}\right) \frac{1}{1-\frac{3}{4}} - \left(\frac{3}{8}\right) \frac{1}{1-\frac{1}{2}} + \left(\frac{3}{64}\right) \frac{1}{1-\frac{1}{4}} \\
&= \left(\frac{27}{64}\right) 4 - \left(\frac{3}{8}\right) 2 + \left(\frac{3}{64}\right) \frac{4}{3} = \frac{27}{16} - \frac{3}{4} + \frac{1}{16} = 1
\end{aligned}$$