If we randomly put balls into $m(\geq 1)$ boxes until $n(\leq m)$ of them are occupied, what's the expectation of the number of balls needed?

Actually this is the inverse problem of the first problem. Let random variable $A_m(n)$ be the number of of balls needed, then

$$A_m(n) = A_m(n-1) + G_m(n), \forall n \geq 2$$

where $A_m(n)$ and $G_m(n)$ are independent and

$$G_m(n) \mid A_m(n-1) \sim Geo(\frac{m-(n-1)}{m}).$$

Let $a_m(n) = E\left(A_m(n)\right)$ then we have

$$\begin{aligned} a_m(n) &= E\left(A_m(n)\right) = E\left(A_m(n-1) + G_m(n)\right) \\ &= a_m(n-1) + \frac{m}{m-(n-1)}, \forall n \geq 2 \end{aligned} \tag{1}$$

Solving this equation iteratively, we have

$$a_m(n) = \sum_{i=0}^{n-1} \frac{m}{m-i} = m \sum_{i=0}^{n-1} \frac{1}{m-i}, \forall m \geq 1, 1 \leq n \leq m.$$

Notice that $a_m(m) = m\sum_{i=0}^{m-1} \frac{1}{m-i} = m\sum_{i=1}^{m} \frac{1}{i} \approx m(log\, m + \gamma)$ when m is large, where $\gamma \approx 0.58$ is the Euler constant.

Using a similar way, we can find the second moment of $A_m(n)$ which is

$$E\left(A_m^2(n)\right) = \sum_{i=0}^{n-1} \left(\frac{m^2+mi}{(m-i)^2} + \frac{2m}{m-i} a_m(i) \right),$$

where $a_m(0)$ is defined as 0. So the variance of $A_m(n)$ is

$$Var(A_m(n)) = E\left(A_m^2(n)\right) - a_m^2(n) = \sum_{i=0}^{n-1} \left(\frac{m^2+mi}{(m-i)^2} + \frac{2m}{m-i} a_m(i) \right) - a_m^2(n).$$

Many international students in U.S. collect coins for fun. Suppose for each quarter a person collects, it's equally likely to be from one of the 50 states in U.S. We're interested in how fast can a person collect quarters from all 50 states in U.S. Let $m = 50$, the expectation of $A_m(n)$ and the $2-\sigma$ interval, i.e. $(\mu - 2\sigma, \mu + 2\sigma)$ are show in Figure 1. From Figure 1 we can see that for fixed $m = 50$,

(i) both the expectation and variance of $A_m(n)$ increase faster and faster as n increases

(ii) we can collect coins very fast at first, e.g. we can expect to collect 40 distinct quarters in the first 80 quarters we collect

(iii) we can predict $A_m(n)$ for relative small n fairly well, e.g. the 2-σ interval for $n = 40$ is approximately (57,100).

(iv) it's very hard for us to collect the last few coins, e.g. on average we need around 225 quarters in order to collect 50 distinct quarters which is much large than the expected number of quarters needed to collect 45 distinct quarters (which is around 111)

(v) it's also becoming hard to predict $A_m(n)$ when n is very large (close to 50), e.g. the 2-σ interval has a length around 250 when $n = 45$

Figure 1: Expectation and 2-σ Intervals for Balls Needed when m=50