If we randomly put $n(\geq 1)$ balls into $m(\geq 1)$ boxes, what's the expectation of the number of empty boxes?

Let random variable $A_m(n)$ be the number of empty boxes, then conditioning on $A_m(n-1)$ we have

Table 1: Distribution of $A_m(n) \mid A_m(n-1)$ when $n \geq 2$

$A_m(n) \mid A_m(n-1)$	$A_m(n-1)-1$	$A_m(n-1)$
P	$\frac{A_m(n-1)}{m}$	$1-\frac{A_m(n-1)}{m}$

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

$$a_m(n) = E\left(A_m(n)\right) = E\left(E\left(A_m(n) \mid A_m(n-1)\right)\right) = \frac{m-1}{m} a_m(n-1), \forall n \geq 2$$

Solving this equation using method of iteration, we have

$$a_m(n) = (m-1)\left(\frac{m-1}{m}\right)^{n-1} = m\left(\frac{m-1}{m}\right)^{n}, \forall m \geq 1, n \geq 1.$$

From the above formula we know that

(i) the expected number of empty boxes decays geometrically

(ii) as m becomes larger, the geometrical decay becomes slower

(iii) the decay of empty boxes is very slow for moderate or large m, since $\frac{m-1}{m}$ is close to 1

Notice that $\forall c > 0$, $a_m(cm) \to me^{-c}$ as $m \to \infty$, i.e. the expect percent of empty boxes if we put cm balls into m boxes is approximately e^{-c} when m is large.

In statistics, usually we're interested in the first two moments of random variables. We can also find $E(A_m^2(n)$ using a similar method. Let $b_m(n) = E(A_m^2(n))$, then

$$b_m(n) = E(A_m^2(n)) = E\left(E(A_m^2(n) \mid A_m(n))\right) = \frac{m-2}{m} b_m(n-1) + (\frac{m-1}{m})^{n-1}, \forall n \geq 2$$

Solving this equation using method of iteration, we have

$$b_m(n) = (m^2-m)\left(\frac{m-2}{m}\right)^{n} + m\left(\frac{m-1}{m}\right)^{n}.$$

So the variance of the $A_m(n)$ is

$$Var(A_m(n)) = b_m(n) - a_m^2(n) = (m^2-m)\left(\frac{m-2}{m}\right)^{n} + m\left(\frac{m-1}{m}\right)^{n} - m^2\left(\frac{m-1}{m}\right)^{2n}.$$

Intuitively the variance of $A_m(n)$ increase first as n increase and then decrease as n increase. The plot of $Var(A_m(n))$ against n when $m = 100$ is in Figure 1, which consists with our intuition. From Figure 1 we can see that the variance of the number of empty boxes when $m = 100$ is around 10 (yields a standard deviation around 3) in the worst case, so we can predict the number of empty boxes reasonably well. Figure 2 presents the 2-σ interval for the number of empty boxes when $m = 100$.

Using a similar way, we can calculate any finite moment of $A_m(n)$. However, usually we're not interested in moments that higher than 2.

Figure 1: Variance of Number of Empty Boxes when $m = 100$

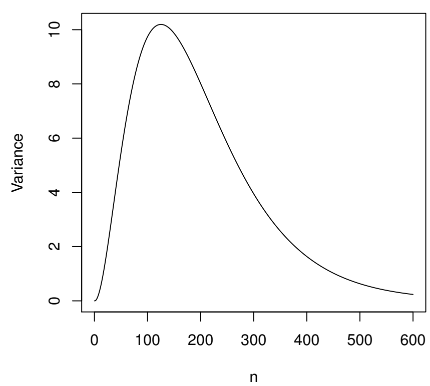

Figure 2: 2-σ Intervals for Number of Empty Boxes when $m = 100$