Problem 3:

m balls into n bins

1. X_i r.v. for bin in i-th ball

Show $\Pr[X_i = X_j] = \frac{1}{n}$ for $i \neq j$

$$\Pr[X_i = X_j] = \sum_{k=1}^{n} \Pr[X_i = X_j = k] = \sum_{k=1}^{n} \Pr[X_i = k \wedge X_j = k] = \sum_{k=1}^{n} \Pr[X_i = k] \cdot \Pr[X_j = k]$$

↑ Law of total probability

↑ independent events

$$= \sum_{k=1}^{n} \frac{1}{n} \cdot \frac{1}{n} = \sum_{k=1}^{n} \frac{1}{n^2} = \frac{1}{n}$$

2. C r.v for #collisions (collision: two balls are in the same bin)

View this as a Bernoulli trial:

- For every distinct (unordered) pair of balls: "success" if in the same bin
  "not success" otherwise
- Success probability $p = \frac{1}{n}$ (see 1.)
- Number of pairs (trials): $N = \binom{m}{2}$

$\Rightarrow E[C] = pN = \binom{m}{2} \frac{1}{n}$ (exp. of binomial distribution)

$\mathrm{Var}[C] = N \cdot p \cdot (1-p) = \binom{m}{2} \frac{1}{n} \cdot \left(1 - \frac{1}{n}\right)$

But: still need to check independence of trials! (Definition of independence)

Claim: For all distinct pairs of ball (a,b) and (c,d): $\Pr[X_a = X_b \wedge X_c = X_d] = \Pr[X_a = X_b] \cdot \Pr[X_c = X_d]$

→ easy to check, similar to 1.

3. Assumption: $m \leq \sqrt{n}$

$$\Pr[C \geq 1] = \Pr[C \geq \alpha \cdot \underbrace{\binom{m}{2} \cdot \frac{1}{n}}_{=E[C]}] \leq \frac{1}{\alpha} = \frac{\binom{m}{2}}{n} = \frac{\frac{m \cdot (m-1)}{2}}{n} \leq \frac{\frac{m^2}{2}}{n} \leq \frac{\frac{n}{2}}{n} = \frac{1}{2}$$

for $\alpha = \frac{n}{\binom{m}{2}}$

↑ Markov

4. Assumption $m \geq 3\sqrt{n}$

$$\binom{m}{2} = \frac{m \cdot (m-1)}{2} \geq \frac{(3\sqrt{n})^2}{2} = \frac{9}{2} n$$

$$\Pr[C \leq 1] = \Pr[-C \geq -1] = \Pr\left[\binom{m}{2} \cdot \frac{1}{n} - C \geq \binom{m}{2} \frac{1}{n} - 1\right]$$

$$\leq \Pr\left[\left|\underbrace{\binom{m}{2} \frac{1}{n}}_{=E[C]} - C\right| \geq \underbrace{\binom{m}{2} \frac{1}{n} - 1}_{\alpha \cdot \mathrm{Var}[C]}\right] \leq \frac{1}{\alpha^2} = \frac{\left(1 - \frac{1}{n}\right)^2}{\left(1 - \frac{n}{\binom{m}{2}}\right)^2} \overset{n \geq 2}{\leq} \frac{\left(1 - \frac{1}{2}\right)^2}{\left(1 - \frac{n}{\frac{9}{2}n}\right)^2} = \frac{\frac{1}{4}}{\left(\frac{7}{9}\right)^2}$$

$$= \frac{81}{4 \cdot 49} \leq \frac{1}{2}$$

$|C - E[C]|$

$= \alpha \cdot \binom{m}{2} \frac{1}{n} \left(1 - \frac{1}{n}\right)$

with $\alpha = \frac{1 - \frac{n}{\binom{m}{2}}}{1 - \frac{1}{n}}$

2. Modification of reservoir sampling

Assumption: $i > k$

We prove by induction that $\Pr[e_i \in S \text{ after step } i+t] = p\left(1 - \frac{p}{k}\right)^t$

Base case: $t = 0$ $\Pr[e_i \in S \text{ after step } i+0] = p = p \cdot \left(1 - \frac{p}{k}\right)^0$

Inductive step $(t \to t+1)$: $\Pr[e_i \in S \text{ after step } i+t+1] = \Pr[e_i \in S \text{ after step } i+t] \cdot \Pr[e_i \text{ survives}]$

$= (*)$

$\Pr[e_i \text{ survives}] = \underbrace{\Pr[e_{i+t+1} \text{ not kept}]}_{1-p} + \underbrace{\Pr[e_{i+t+1} \text{ kept}]}_{p} \cdot \underbrace{\Pr[e_i \text{ not replaced when } e_{i+t+1} \text{ kept}]}_{\left(1-\frac{1}{k}\right) = \frac{k-1}{k}}$

$= 1 - p + p \cdot \frac{k-1}{k} = 1 - \frac{p}{k}$

$(*) = \underbrace{p \cdot \left(1-\frac{p}{k}\right)^t}_{\text{by IH}} \cdot \left(1-\frac{p}{k}\right) = p \cdot \left(1-\frac{p}{k}\right)^{t+1}$

2. exp. number of steps $= \frac{1}{1 - \text{survival pr.}}$ (geometrically distributed)