Counting Turtle Confidence — Detailed Solution

COMP 480/580 — Assignment 1, Question 2
Reference: Lecture 1 (Mark and Recapture Estimation)

Goal

We perform R independent repetitions of the mark-and-recapture experiment. Each repetition produces a measurement M_i (the number of marked turtles seen in the second sample). We use the sample mean

$$\overline{M} = \frac{1}{R} \sum_{i=1}^{R} M_i$$

as an estimator of $\mathbb{E}[M]$. The estimator for the population size n is

$$\hat{n} = \frac{k_1 k_2}{\overline{M}}.$$

We want a high-probability confidence interval for $\overline{M}$ (and therefore for $\hat{n}$).

Step 1 — Expectation and variance of M

Notation:

- n: true number of turtles (unknown),
- k_1: number initially marked and released,
- k_2: size of the second sample (drawn without replacement),
- M: number of marked turtles observed in the second sample.

From the lecture notes,

$$\mathbb{E}[M] = \frac{k_1 k_2}{n}.$$

Since M follows a hypergeometric distribution, its variance is

$$\mathrm{Var}(M) = k_2 \cdot \frac{k_1}{n} \left(1 - \frac{k_1}{n}\right) \cdot \frac{n - k_2}{n - 1}.$$

If we set $p := k_1/n$, then $\mathbb{E}[M] = k_2 p$ and

$$\mathrm{Var}(M) = k_2 p (1 - p) \cdot \frac{n - k_2}{n - 1}.$$

A useful upper bound is

$$\mathrm{Var}(M) \leq k_2 p = \mathbb{E}[M].$$

Step 2 — Chebyshev bound for the sample mean

Let $M_1, \ldots, M_R$ be independent repetitions. Then

$$\overline{M} = \frac{1}{R}\sum_{i=1}^{R} M_i, \quad \mathrm{Var}(\overline{M}) = \frac{\mathrm{Var}(M)}{R}.$$

By Chebyshev's inequality, for any $a > 0$,

$$\Pr\left(|\overline{M} - \mathbb{E}[M]| \geq a\right) \leq \frac{\mathrm{Var}(\overline{M})}{a^2} = \frac{\mathrm{Var}(M)}{Ra^2}.$$

To guarantee probability at least $1 - \delta$, choose

$$a \geq \sqrt{\frac{\mathrm{Var}(M)}{R\delta}}.$$

Using the conservative bound $\mathrm{Var}(M) \leq \mathbb{E}[M]$, we may write

$$a \leq \sqrt{\frac{\mathbb{E}[M]}{R\delta}}.$$

Step 3 — Interval for $\hat{n}$

With probability at least $1 - \delta$,

$$\overline{M} \in [\mathbb{E}[M] - a,\ \mathbb{E}[M] + a].$$

Because $\hat{n} = \frac{k_1 k_2}{\overline{M}}$ and $1/x$ is monotone decreasing for $x > 0$, we obtain

$$\hat{n} \in \left[\frac{k_1 k_2}{\mathbb{E}[M] + a},\ \frac{k_1 k_2}{\mathbb{E}[M] - a}\right].$$

This is valid provided $\mathbb{E}[M] > a$ (to ensure positivity of the denominator).

Step 4 — Choosing R for a relative error target

Suppose we want $\overline{M}$ to be within a relative error f of $\mathbb{E}[M]$:

$$a \leq f\,\mathbb{E}[M].$$

Plugging in Chebyshev's choice for a gives

$$\sqrt{\frac{\mathrm{Var}(M)}{R\delta}} \leq f\,\mathbb{E}[M] \quad \Longrightarrow \quad R \geq \frac{\mathrm{Var}(M)}{\delta f^2\,\mathbb{E}[M]^2}.$$

Using $\mathrm{Var}(M) \leq \mathbb{E}[M]$, we obtain the simpler sufficient condition

$$R \geq \frac{1}{\delta f^2\,\mathbb{E}[M]}.$$

Numeric example

Take $k_1 = 100$, $k_2 = 100$, $n = 10000$. Then

$$\mathbb{E}[M] = \frac{100 \cdot 100}{10000} = 1.$$

The variance is

$$\mathrm{Var}(M) = 100 \cdot \frac{100}{10000}\left(1 - \frac{100}{10000}\right)\frac{9900}{9999} \approx 0.98.$$

Choose $\delta = 0.05$ and fractional error $f = 0.1$. Then

$$R \geq \frac{0.98}{0.05 \cdot 0.01 \cdot 1^2} \approx 1960.$$

Using the conservative bound, we get

$$R \geq \frac{1}{0.05 \cdot 0.01 \cdot 1} = 2000.$$

When is estimation hard?

1. If $\mathbb{E}[M] = k_1 k_2 / n$ is small, many repetitions are needed since R scales like $1/\mathbb{E}[M]$.
2. The variance is often comparable to $\mathbb{E}[M]$, making concentration weak.
3. If $\mathbb{E}[M]$ is close to zero, the interval for $\hat{n}$ may become vacuous (division by small values).
4. Independence of repetitions is required; poor mixing of turtles invalidates the bound.

Summary

By Chebyshev's inequality:

$$\overline{M} \in [\mathbb{E}[M] - a,\ \mathbb{E}[M] + a], \quad a = \sqrt{\frac{\mathrm{Var}(M)}{R\delta}},$$

so that

$$\hat{n} \in \left[\frac{k_1 k_2}{\mathbb{E}[M] + a},\ \frac{k_1 k_2}{\mathbb{E}[M] - a}\right].$$

A sufficient condition for fractional error f and failure probability δ is

$$R \geq \frac{1}{\delta f^2\, \mathbb{E}[M]}.$$

This highlights that estimation becomes difficult when $\mathbb{E}[M]$ is small.