Question 3: Thompson Sampling

(a) Posterior Update

Let arm i be pulled $N_{i,t}$ times until time t, and let $S_{i,t}$ be the number of successes:

$$S_{i,t} = \sum_{u \leq t : i_u = i} r_{i,u}$$

The prior for μ_i is:

$$\mu_i \sim \text{Beta}(1, 1)$$

The likelihood of observing $S_{i,t}$ successes and $N_{i,t} - S_{i,t}$ failures is:

$$\mu_i^{S_{i,t}} (1 - \mu_i)^{N_{i,t} - S_{i,t}}$$

The posterior is: $\boxed{P_{i,t} = \text{Beta}(1 + S_{i,t}, 1 + N_{i,t} - S_{i,t})}$

(b) Mean and Variance of Posterior

Given $\mu_i \sim \text{Beta}(\alpha, \beta)$:

- Mean:

$$\mathbb{E}[\mu_i] = \frac{\alpha}{\alpha + \beta}$$

- Variance:

$$\text{Var}[\mu_i] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Apply to $\alpha = 1 + S_{i,t}, \beta = 1 + N_{i,t} - S_{i,t}$:

$$\mathbb{E}[\mu_i] = \frac{1 + S_{i,t}}{2 + N_{i,t}}$$

$$\boxed{\text{Var}[\mu_i] = \frac{(1 + S_{i,t})(1 + N_{i,t} - S_{i,t})}{(2 + N_{i,t})^2(3 + N_{i,t})}}$$

(c) Exploration vs Exploitation

Thompson Sampling balances exploration and exploitation by sampling:

$$\hat{\mu}_{i,t} \sim \text{Beta}(1 + S_{i,t}, 1 + N_{i,t} - S_{i,t})$$

- **Exploration**: Arms with few pulls (small $N_{i,t}$) have high-variance posterior distributions, leading to a higher chance of being selected.
- **Exploitation**: Arms with high rewards and many pulls have concentrated posterior mass near the mean, leading to consistent selection.

Thus, Thompson Sampling naturally trades off between exploring uncertain arms and exploiting high-reward arms.