Introduction To Bandits

by Elyn Franson
Mentored by Ronan Perry
DRP Spring 2025

What are Bandits?

- Suppose I am a doctor and there are two treatments I can give my patients.

- Suppose I give treatment 1 to a patient and they recover but, I give treatment 2 to a patient and they don’t recover.

- What treatment should I give to my next patient?

- Suppose I give my third patient treatment 1 and they don’t recover. What should I do now? Giving them treatment 1 is the *greedy* strategy.

- Is there a way to formalize a decision strategy? Are there better strategies?

- These sorts of problems related to treatments, drugs, advertisements, ect are called *Bandits*

Can we do better than greedy?

- How can greedy go wrong?

- Recall our example: What is estimated probability of a patient recovering under treatment 2?

    - Answer: 0 (the mean of 0 is 0)

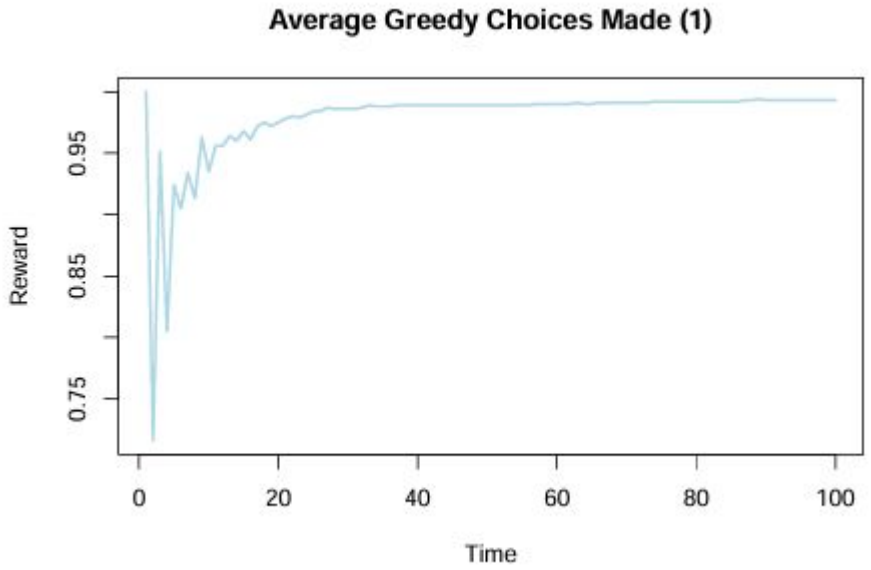
- Will I ever give another patient treatment 2?

    - Answer: No

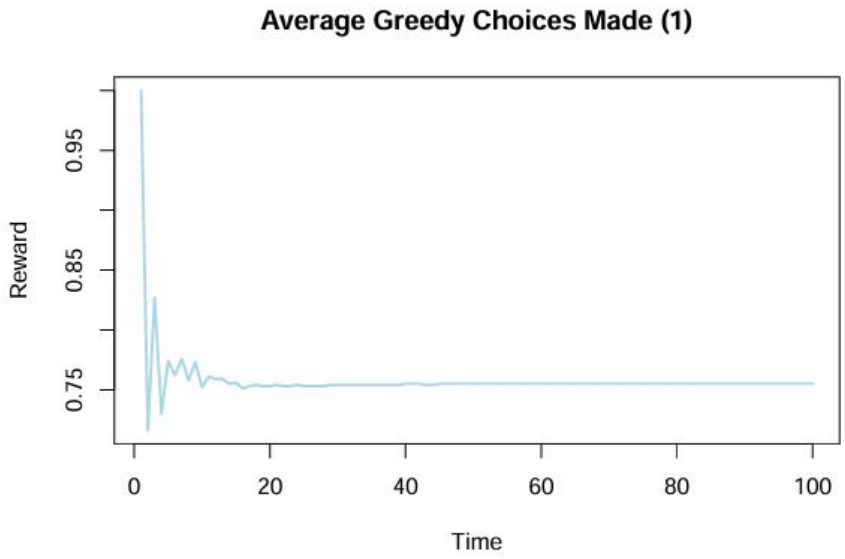
- What are we not considering here?

    - Answer: We could have simply gotten unlucky with treatment 2, e.g., the variance is high.

Greedy - The Simplest Method



p1 = 0.2 and p2 = 0.7



p1 = 0.6 and p2 = 0.7

Greedy Alternatives

Epsilon-Greedy

- Makes a simple choice to sample epsilon% of the other method to explore

Upper Confidence Bound (UCB)

- Chooses treatment with the best upper confidence bound

Thompson Sampling

- Assumes beta distribution and updates according to the beta distribution

Epsilon Greedy

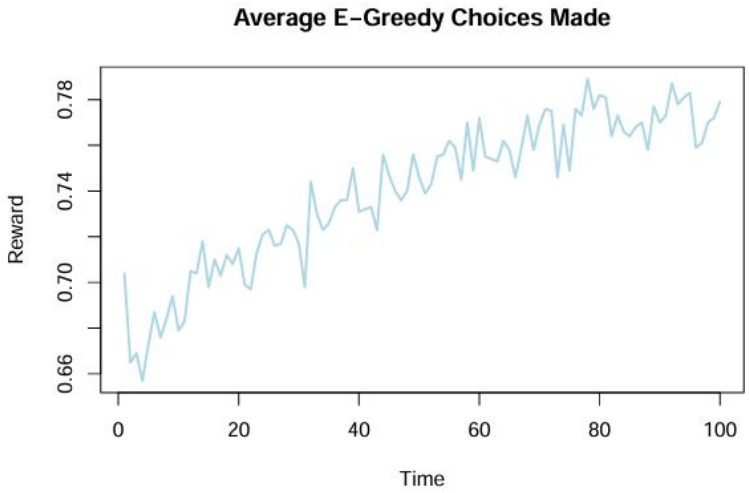
- Explores some of the other data options epsilon% of the time

- Helpful when you don't have a big sample to start and want to explore

- Introduces variability

- In out example improved choices made from 0.75 to 0.78

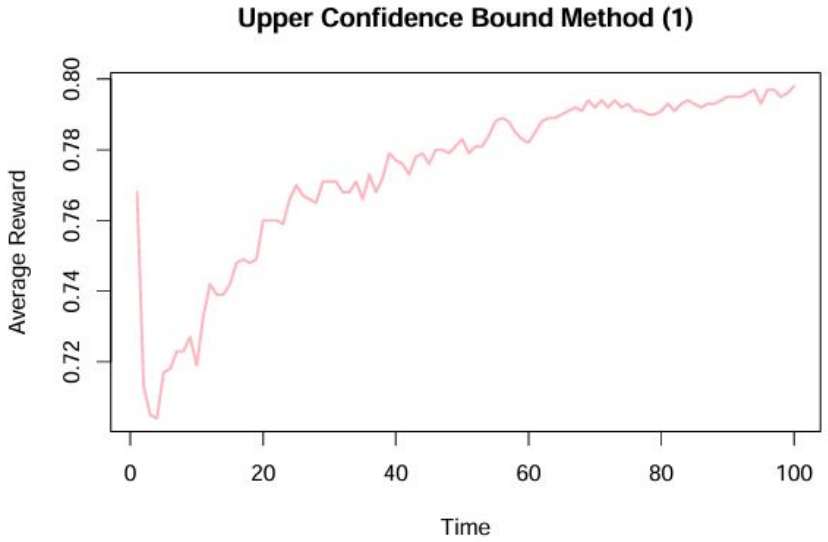
- Will never fully converge



Upper Confidence Bound

- Chooses method with the best upper confidence bound instead of best mean

- Converges to 0.8 slight improvement from e-greedy 0.78



TS Greedy Choices Made

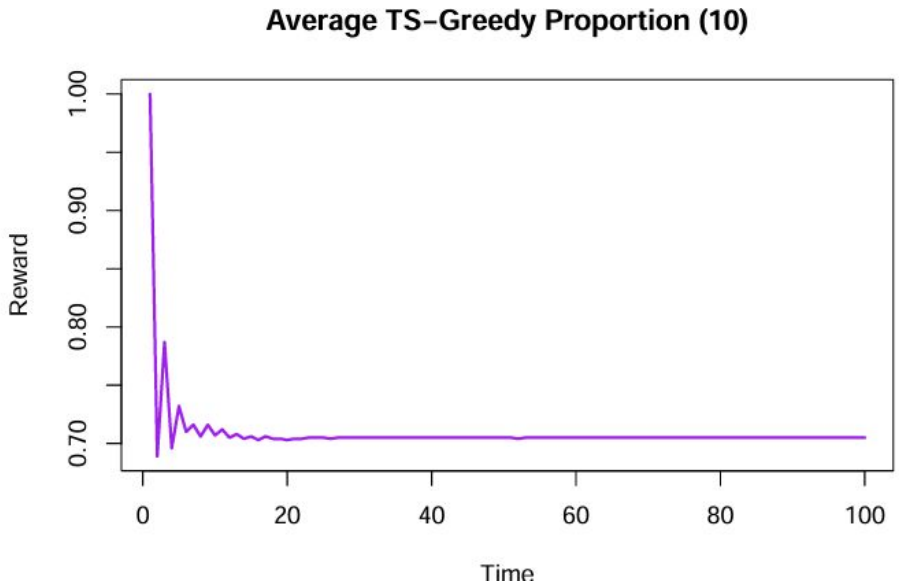
- Similar to greedy except start with a prior belief and update it -> posterior

- Alpha marks number of success

- Beta marks number of failures

- Choose based on E[p] = (alpha / alpha + beta) mean

- Assume drug has uniform prior (0,1) give patient drug and you see a success update



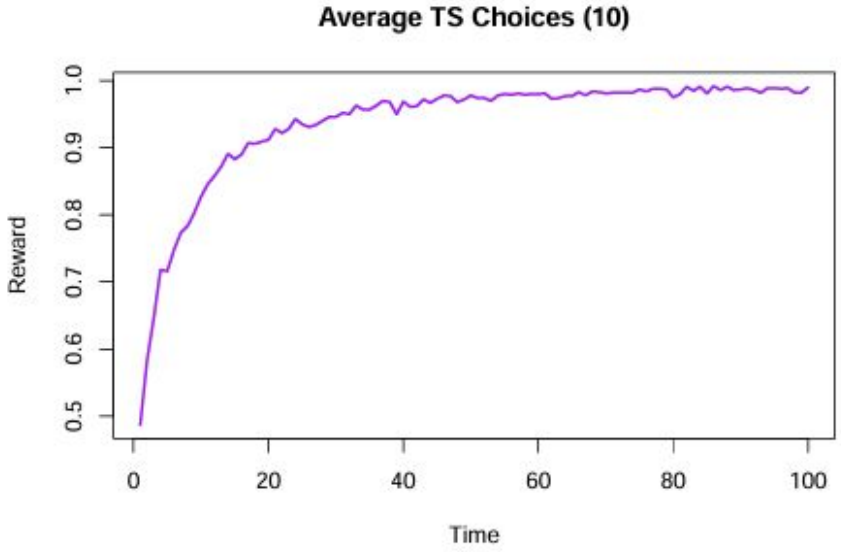
Thompson Sampling

- Start with prior belief

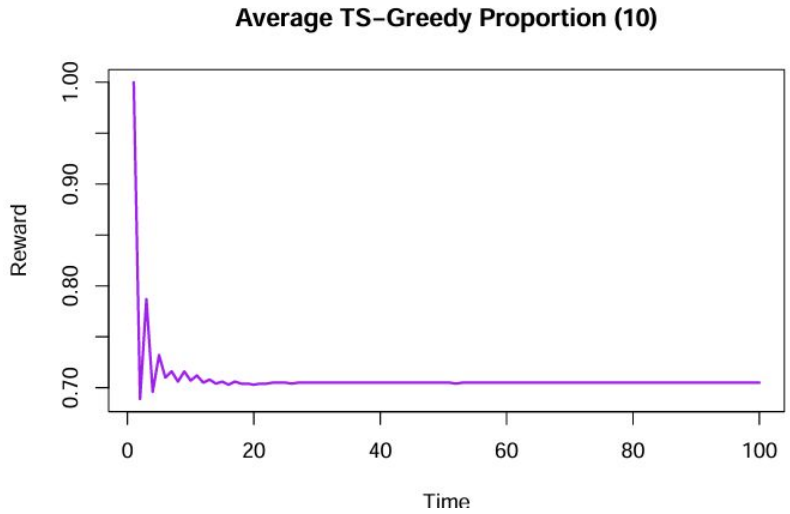
- Draw randomly from each beta distribution choose the one with the highes mean and then choose that one and update alpha and beta

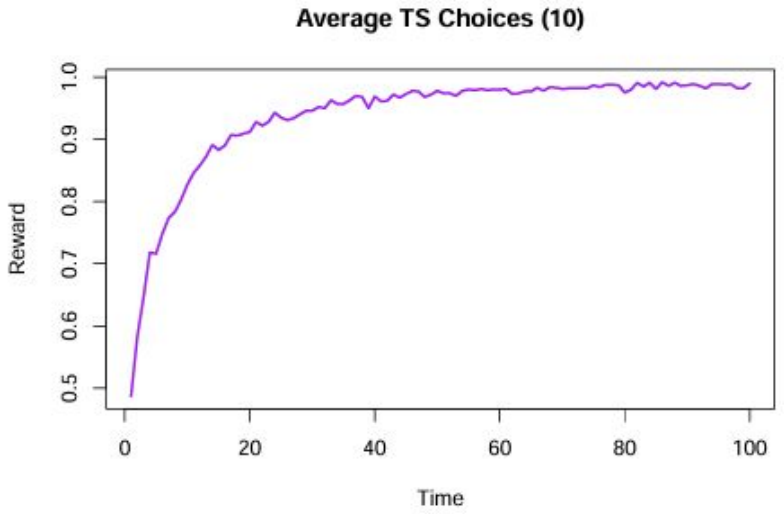
- Posterior is the likelihood of our data given the prior is true times the prior

- Prior heavily effects results



TS vs TS Greedy





Comparing Methods

- Greedy: Takes the best mean and doesn't account for variance
- E-Greedy: Always takes the other action epsilon% of the time
- Upper Confidence Bound: Uses upper confidence bound instead of the mean
- Thompson Sampling: Uses bayesian methods and assumes distribution

Thank You!

Questions?