7-beers-matching

Kunal Marwaha

January 22, 2018

Introduction

A few friends and I bought 7 types of bad beers (think Natty, Miller Lite, Yuengling...) and poured them into 7 uniquely labeled cups. We each tasted all 7 beers and guessed which was which. (Most of us did poorly; I think the high was 4/7.)

How would we do if we guessed randomly? What if there were n beers?

For all parts, justify your answers where possible. Give a brief proof if you can.

Part 1

Let's simplify the problem. Suppose there are 3 beers (Awful, Bud, and Coors). I randomly pour them into three cups, and you guess which is which, randomly. I check your answers after you have guessed all three (so no re-using guesses!).

a) What's the chance that you guess all 3 correctly?

b) What's the chance that you guess exactly 2/3 correctly?

c) What's the chance that you guess exactly 1/3 correctly?

d) What's the chance that you guess exactly 0/3 correctly?

e) What's the expected number of beers that you will guess correctly?

Hint: Think of this as an ordering problem.

Part 2

Let's define a random variable R_n (with $n \in \mathbb{N}$), which represents the distribution of guessing some number of n beers correctly. Precisely, $Pr(R_n = k)$ represents the probability that you randomly guess exactly k out of n beers correctly.

a) Let's first make the problem even smaller. Suppose there are only two beers (Awful and Bud). What is the distribution of R_2? What is its expected value?

b) What is the distribution and expected value of R_1? What does R_1 represent?

c) Reflect on the previous results. What can you gather about $Pr(R_n = k)$, where $k > n$? What about when $k < 0$? Why?

d) What is $Pr(R_n = n)$? Why?

e) What is $Pr(R_n = n - 1)$? Why?

f) What is $E(R_n)$? How does it depend on n? Why? (Hint: Use linearity of expectation.)

Part 3

Let's tease out a recurrence relation to better understand R_n.

a) Suppose we're guessing with 3 beers. We make a guess on the first one, and we're left with two beers and two cups. What are the possible outcomes? Are they all represented with R_2?

> Let's define a new variable, $R_{n,l}$ (with $l \in \mathbb{N}$ and $l \leq n$), which represents the distribution of guessing some number of n beers correctly, with l "bad" beer names that do not match any cups. In particular, $R_{n,0} = R_n$, since all beer names match exactly 1 cup.

b) Define a recurrence relation for $R_n = R_{n,0}$. (Hint: Use Part (3a) to define R_3 in terms of $R_{2,l}$ for various l. Then, generalize.)

c) Suppose, in Part (3a), we missed the guess. How do you represent the two beers and two cups with $R_{n,l}$? What are the possibilities when we guess again? State the recurrence relation for this particular case. Does it match up with the recurrence for $R_{n,0}$?

Part 4

In order to unlock the riddle, we need more information about $R_{n,l}$.

a) What does $R_{n,n}$ represent? What is its distribution and expected value?

b) What does $R_{n,n-1}$ represent? What is its distribution and expected value?

c) The recurrence relation for $R_{n,l}$ is very similar to the one for R_n, but it depends on l. How does it change? Check that Part (3c) matches this recurrence. (Hint: As long as $n > l$, there will be at least one cup that you have a nonzero chance of guessing correctly (why?). Assume that you guess that cup next. What are the outcomes?)

d) Use all of the above information to compute the recurrence for R_4 and each of its subproblems. Do this by hand. What does R_4 represent? What are the base cases in the recurrence?

Part 5

Let's write a small script, and look for asymptotic approximations.

a) Write a recursive program above to compute the distribution of $R_{n,k}$. Check R_n for $n = 1, 2, 3, 4$ to see that the distribution and expected value match up with our earlier calculations.

b) What is the distribution of R_7? Expected value? What's the chance that the high-scorer got exactly 4/7 if they guessed randomly? At least 4/7? Is it statistically significant at the $p = 0.05$ level? (This can be modeled as a 1-sided statistical test to see if the high-scorer really 'knows their bad beers'.)

c) Compute the distribution of R_{20}, R_{100}, and R_{1000}. (Hint: You may want to introduce caching to your script.) How does $Pr(R_n = 0)$ change as n increases? What about $Pr(R_n = 1)$ or for other small numbers? What patterns do you notice? (Hint: Look at reciprocals and ratios.)

d) It turns out that the full distribution of R_n can be generated by the sequence $Pr(R_n = 0)$. How? Represent $Pr(R_n = k)$ in terms of $Pr(R_m = 0)$ for various m. Check that Part (2e) and Part (2f) still hold.

e) What does this mean for $Pr(R_n = k)$ as $n - k$ increases? (There is an asymptotic approximation.) How soon (i.e. what $x = n - k$) are you within 1/1000 of the approximation?

Hope you learned a little math on the way to getting drunk :-)