Caps and robbers - what can you expect?

Laura A. Zager and George C. Verghese

December 9, 2006

The World Series has just come to a thrilling finish, and you've decided to celebrate the end of baseball season by throwing a party. You've invited n of your best baseball buddies over, and as they arrive, they all toss their baseball caps into the closet. The party is a huge success, but as the hour gets late, it's time for everyone to collect their caps and head home. Rather than sort through the messy pile of caps, you decide to streamline the process and hand each guest a cap drawn at random from the closet as they leave. Anticipating grumbles from friends who end up with caps other than their own, you might ask yourself "What is the expected number of guests who receive their own caps?"

Answering this question is a common exercise in undergraduate introductory probability courses, typically referred to as the "matching hats" problem (see [1], for example). Its solution is good practice with *indicator random variables*, which take the value 1 when an event occurs and 0 otherwise. Indicator random variables have a special property that makes them very useful: their expected value is simply the probability that the event occurs. Here, we define I_k to be 1 when cap k is assigned to its owner and 0 otherwise, and let the variable I denote the total number of guests who receive their own cap. Then we can use the linearity of the expected value to write

$$\mathrm{E}[I] = \mathrm{E}[I_1 + I_2 + \cdots + I_n] = n\mathrm{E}[I_k] = nP(I_k = 1).$$

Now, we've reduced the problem to finding the probability that a given cap k has been matched with its owner, and this is simply $1/n$. Thus,

$$\mathrm{E}[I] = n\frac{1}{n} = 1.$$

That the answer is independent of n, the number of party guests, is quite surprising! Beyond the expected number of correct matches, there are many other probabilities that are interesting to explore in this simple scenario, and that lead to many creative solutions (see, for example, Hathout's treatment of this problem in [2]). Here, we present some twists on this problem and continue to look at expected values; our solutions have satisfying and intuitive limiting behavior, and involve fun tricks with iterated expectation and binomial and hypergeometric distributions.

Problem 1A: Matching Within Teams Your n best baseball buddies are drawn from the fan clubs of the s different teams in the league. Each team's fan club is infinitely large, and each fan is only a member of a single club. If each of your guests is equally likely to belong to any of the s clubs, what is the expected number of guests who get a cap of their team (but not necessarily their own)?

Solution We again focus on a specific cap k. Let the random variable C_k be 1 when cap k is matched to a guest from the same team and '0' otherwise, and let C be the total number of caps correctly matched to a guest from the same team. Of course, the variable C is also the number of guests who receive one of their own team's caps. Then

$$\mathrm{E}[C] = \mathrm{E}\left[C_1 + C_2 + \cdots + C_n\right] = n\mathrm{E}\left[C_k\right].$$

Now, define the variable F_k to be the number of your buddies who are fans of the team represented by cap k. If we know F_k, then we know that cap k has probability F_k/n of being matched to a guest of the correct team. Denote the expected value of C_k given that we know F_k by $E[C_k|F_k]$. Wouldn't it be convenient to be able to assume F_k, compute $E[C_k|F_k]$, then use the expectation of F_k to give us a final answer for $E[C_k]$? This is indeed the case, and is a very useful technique called *iterated expectation.* This technique works as follows:

$$\mathrm{E}[C] = n\mathrm{E}\left[C_k\right] = n\mathrm{E}\left[\mathrm{E}\left[C_k | F_k \text{ same-team guests}\right]\right] = n\mathrm{E}\left[\frac{F_k}{n}\right] = \mathrm{E}\left[F_k\right].$$

What is the expected value of F_k? We know that at least one guest is a fan of the team represented by cap k (namely, cap k's owner), and each of the other $n-1$ guests is a fan of the same team independently with probability $1/s$. The number of guests (besides cap k's owner) who are fans of the team represented by cap k is a *binomial random variable*, which can be thought of as the total number of successes in a sequence of $n-1$ independent trials, each of which has probability of success $1/s$. The binomial random variable is one of the first discrete random variables studied in introductory probability courses, and has expected value $(n-1)/s$. With this observation, we can write down the expected value of F_k (and consequently, $\mathrm{E}[C]$):

$$\mathrm{E}[C] = \mathrm{E}[F_k] = 1 + \frac{n-1}{s} = \frac{n+s-1}{s}.$$

Now we consider the limiting behavior of this result. When your buddies are all fans of the same team

($s = 1$), then all of the guests will be matched with their team's cap, so $\mathrm{E}[C] = n$ and everybody goes home happy. If there is a large number of teams in the league ($s \to \infty$), it is unlikely that any one team has more than one fan at your party; in this case, a guest being matched to a cap from his or her team is equivalent to being matched to his or her own cap, and we've returned to the original matching hats problem, in which $\mathrm{E}[C] = 1$. If your party has a huge guest list ($n \to \infty$), we expect that a guest will be assigned a cap from his own team with probability approximately $1/s$, and thus the expected number of correct team pairings will be approximately n/s.

We can take a different approach by starting from a "team-centered" perspective rather than a "cap-centered" one. This second perspective provides a different means of solving many of the problems discussed in this paper, so we'll use it here to come up with an alternate solution for Problem 1A. Define the random variable T_i to be the number of guests who are fans of team i who receive a team-i cap, and let T be the total number of guests who receive a hat from their team. Then

$$\mathrm{E}[T] = \mathrm{E}\left[T_1 + T_2 + \cdots + T_s\right] = s\mathrm{E}\left[T_i\right].$$

Now, we can apply the technique of iterated expectation again, but this time assume that we know the number G_i of guests who root for team i. Then

$$\mathrm{E}[T] = s\mathrm{E}\left[\mathrm{E}\left[T_i | G_i \text{ team-}i \text{ guests}\right]\right].$$

Each of the G_i team-i guests has probability G_i/n of being assigned a team-i cap. If we know G_i, T_i is a binomial random variable on G_i trials with probability of success G_i/n. Substituting the expression for the expectation of a binomial random variable, we find that

$$\mathrm{E}[T] = s\mathrm{E}\left[G_i \frac{G_i}{n}\right] = \frac{s}{n}\mathrm{E}\left[G_i^2\right].$$

Since each of the n guests at your party is a fan of team i with probability $1/s$, the total number G_i of team-i fans is a binomial random variable on n trials with probability of success $1/s$. It's a good exercise to show that $E[G_i^2] = n\frac{1}{s}\left(1 - \frac{1}{s} + n\frac{1}{s}\right)$. Using this result yields the correct solution:

$$\mathrm{E}[T] = \frac{s}{n}\left[n\frac{1}{s}\left(1 - \frac{1}{s} + n\frac{1}{s}\right)\right] = \frac{n + s - 1}{s}.$$

Problem 1B: Unique Matching Within Teams Now suppose that instead of distributing all of the caps at random, you only give a team-i cap to a team-i fan (but not necessarily their own). Under this improved distribution scheme, what is the expected number of guests who receive their own cap?

Solution To solve this problem, we'll take the "team-centered" approach. Define the random variable D_i to be the number of fans of team i who receive their own cap, and let D be the total number of guests who receive their own cap. Then:

$$\mathrm{E}[D] = \mathrm{E}\left[D_1 + D_2 + \cdots + D_s\right] = s\mathrm{E}\left[D_i\right].$$

Again, we use G_i to represent the number of team-i guests at your party. This time, though, we won't assume that we know G_i exactly, but that we do know whether G_i is at least one or not. Using this kind of assumption invokes the *law of total probability*, which allows us to break up the expected value computation into two pieces as follows:

$$\mathrm{E}[D] = s\left(\mathrm{E}\left[D_i | G_i \geq 1 \text{ team-}i \text{ guests}\right] P(G_i \geq 1) + \mathrm{E}\left[D_i | G_i = 0 \text{ team-}i \text{ guests}\right] P(G_i = 0)\right).$$

As we observed in the team-centered solution to Problem 1A, G_i is binomially distributed over n trials with probability of success $1/s$. Thus,

$$P(G_i \geq 1) = 1 - P(G_i = 0) = 1 - \left(1 - \frac{1}{s}\right)^n.$$

Given G_i and the new cap-distribution scheme that we've chosen, we are distributing G_i caps at random to the G_i fans of team k; this is simply the original matching hats problem for G_i party guests, so the first expectation $\mathrm{E}\left[D_i | G_i \geq 1 \text{ team-}i \text{ guests}\right]$ is 1. If $G_i = 0$, then D_i must be zero. Putting these facts together gives us our answer:

$$\mathrm{E}[D] = s\left\{1\left(1 - \left(1 - \frac{1}{s}\right)^n\right) + 0\left(1 - \frac{1}{s}\right)^n\right\} = s\left(1 - \left(1 - \frac{1}{s}\right)^n\right)$$

Let's look at what happens in some special cases. When all of your buddies root for a single team, we are back to the original matching hats problem and its solution, $\mathrm{E}[D] = 1$. If the number of teams in the league is large, we again expect that no team will have more than one fan at the party; with a distribution rule that makes sure that you match guests with a cap from their own team, each guest must be matched with their own cap so that $\mathrm{E}[D] = n$. As the number of guests gets large, we expect that there will be one match within each of the s teams represented, and so $\mathrm{E}[D] = s$.

Next, we take a few steps back to a simpler problem, but we add a twist.

Problem 2: Stolen Caps At the end of the party, you open the closet to redistribute the caps and find that $n - m$ of them have been stolen, presumably at random! Faced with a herd of guests heading towards the door, you decide to choose m of your buddies to receive caps and send the remaining $n-m$ home capless. The m caps are distributed at random to your m lucky buddies. As you try to assess the damage, you might ask yourself, "What is the expected number of guests who receive their own cap?"

Solution As in our original problem, we define the random variable I_k to be 1 when cap k is correctly matched to its owner and let I be the total number of caps correctly matched. Then

$$\mathrm{E}[I] = \mathrm{E}\left[I_1 + I_2 + \cdots + I_m\right] = m\mathrm{E}\left[I_k\right] = mP(I_k = 1).$$

Recall from our original problem that the probability of any one cap being correctly matched to its owner is simply $1/n$, which gives us the answer straight away:

$$\mathrm{E}[I] = m\left(\frac{1}{n}\right) = \frac{m}{n}.$$

When no caps are stolen ($m = n$), then of course we are back to our original problem.

Problem 2A: Stolen Caps, With Teams Matching your guests with their own caps is a tough task, so you'd rather think about the number of your guests who receive a cap from their own team. If each of your guests is equally likely to be a fan of any one of the s different teams, what is the expected number of guests who receive their team's cap when $n - m$ of the caps are stolen?

Solution Parallel to the solution to Problem 1A, define C_k to be the random variable that takes the value 1 when the kth cap is matched to a guest of the correct team. Then

$$\mathrm{E}[C] = \mathrm{E}\left[C_1 + C_2 + \cdots + C_m\right] = m\mathrm{E}\left[C_k = 1\right].$$

Again, we assume that we are given the number of guests F_k who are fans of the team represented by cap k. Then cap k has probability F_k/n of being matched to a guest of the same team, and thus,

$$\mathrm{E}[C] = m\mathrm{E}\left[\mathrm{E}\left[C_k | F_k \text{ same-team guests}\right]\right] = m\mathrm{E}\left[\frac{F_k}{n}\right].$$

We already determined $\mathrm{E}[F_k]$ in the solution to Problem 1A, so we can immediately write down $\mathrm{E}[C]$:

$$\mathrm{E}[C] = m\left(1 + \frac{n-1}{s}\right) = \frac{m}{n}\left(\frac{n+s-1}{s}\right).$$

Interestingly, the solution to this problem is the product of the solutions to Problems 1A and 2! When all of the guests are fans of a single team, each of the m guests who receives a cap receives one from the correct team, so $\mathrm{E}[C] = m$. As the number of teams in the league gets large, matching with one's own team becomes equivalent to matching with one's own cap, so we are back to the question posed in Problem 2 in which $\mathrm{E}[D] = m/n$. As the number of party guests increases, the probability that any of the m hats is distributed to a fan of the correct team is $1/s$, so C becomes binomially distributed with expected value $\mathrm{E}[C] = m/s$.

Problem 2B: Stolen Caps, With Teams and Uniqueness Trying to make the best out of a bad situation, you decide to make sure that each of the m remaining caps goes to a fan of the matching team, but otherwise distribute them randomly. What is the expected number of guests who receive their own cap?

Solution Denote the number of team-i guests paired with their cap by D_i. Our solution to this problem follows the form of Problem 1B, but will require an extra step: we assume that we know R_i, the number of team-i caps available to be distributed. Then

$$\begin{aligned}
\mathrm{E}[D] &= \mathrm{E}\left[D_1 + D_2 + \cdots D_s\right] = s\mathrm{E}\left[D_i\right] \\
&= s\left\{\mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[D_i | G_i \text{ team-}i \text{ guests, } R_i \text{ team-}i \text{ caps available, } G_i \geq 1 \text{ team-}i \text{ guests}\right]\right]\right] P(G_i \geq 1)\right. \\
&\quad \left. + \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[D_i | G_i \text{ team-}i \text{ guests, } R_i \text{ team-}i \text{ caps available, } G_i = 0 \text{ team-}i \text{ guests}\right]\right]\right] P(G_i = 0)\right\}.
\end{aligned}$$

Recalling the simple solution to Problem 2, we see that the expected value of D_i (when we are given G_i and R_i) is simply R_i/G_i for $G_i \geq 1$. For $G_i = 0$, $D_i = 0$. Substituting this observation into the calculation, and recalling that G_i is binomially-distributed, we find:

$$\mathrm{E}[D] = s\left\{\mathrm{E}\left[\mathrm{E}\left[\frac{R_i}{G_i}|G_i \text{ team-}i\text{ guests, } R_i \text{ team-}i\text{ caps available}\right]\right]\left(1-\left(1-\frac{1}{s}\right)^n\right)\right\}.$$

What is the expected value of R_i if we know G_i? R_i is simply the number of team-i caps selected (from the total of G_i) when m caps are selected to remain from the original n. One can imagine this selection process occurring by dropping m balls into n different boxes, where G_i of the boxes are labeled 'Team i'. The distribution that describes the number of balls that end up in the 'Team-i' boxes is the *hypergeometric distribution*, which has the following probability mass function and expected value (see, for example, [3]):

$$P(R_i = x|G_i) = \frac{\binom{G_i}{x}\binom{n-G_i}{m-x}}{\binom{n}{m}} \text{ for } x \in [\max(0, m+G_i-n),\ \min(G_i, m)],$$

$$\mathrm{E}[R_i|G_i] = \frac{mG_i}{n}.$$

Using this expression for $\mathrm{E}[R_i|G_i]$ yields

$$\begin{aligned}\mathrm{E}[D] &= s\left\{\mathrm{E}\left[\frac{1}{G_i}\frac{mG_i}{n}\right]\right\}\left(1-\left(1-\frac{1}{s}\right)^n\right)\\ &= s\frac{m}{n}E[1]\left(1-\left(1-\frac{1}{s}\right)^n\right) = s\frac{m}{n}\left(1-\left(1-\frac{1}{s}\right)^n\right).\end{aligned}$$

As in the solution to Problem 2A, the solution here is the product of the solutions to Problems 1B and 2. When all of the guests are fans of the same team, this problem reduces to Problem 1 and $\mathrm{E}[D] = m/n$. If the number of teams in the league is large, the distribution scheme matches unique caps to unique guests as discussed previously, yielding $\mathrm{E}[D] = m$. As the number of guests gets large, the expected number of correct matches goes to zero.

Problem 2C: Stolen Caps, With Known Matches and Uniqueness Anticipating the challenge of redistributing caps to your party guests when there's a thief on the loose, you add some extra instructions to the invitation for your next party. This time, you invite n fans of one team, and ask r of them to label their caps with their names. Right on cue, the thief strikes again and steals $n - m$ of the caps at random from the closet. This time, though, you were ready for precisely this contingency! To distribute the available caps, you first identify the labeled caps and return them to their owners, then redistribute the unlabeled ones to the $n-r$ guests who did not label their caps. What is the expected number of caps returned to their owner?

Solution Again, let I_k be the random variable that takes the value 1 when available cap k is matched with its owner and 0 otherwise. Then

$$\mathrm{E}[I] = \mathrm{E}[I_1 + I_2 + \cdots + I_m] = m\mathrm{E}[I_k] = mP(I_k = 1).$$

Here, we use the law of total probability, applied to the event that cap k is one of the labeled caps, which occurs with probability r/n. If cap k is labeled, it is correctly matched with probability 1. If cap k is unlabeled, it is equally likely to be matched with any of the $n-r$ guests whose caps were unlabeled. Thus,

$$\begin{aligned}
\mathrm{E}[I] &= mP(I_k = 1) \\
&= m\left[P(I_k = 1|\text{cap-}k \text{ labeled})P(\text{cap-}k \text{ labeled}) + P(I_k = 1|\text{cap-}k \text{ unlabeled})P(\text{cap-}k \text{ unlabeled})\right] \\
&= m\left[(1)\left(\frac{r}{n}\right) + \left(\frac{1}{n-r}\right)\left(1 - \frac{r}{n}\right)\right] \\
&= \frac{m}{n}\left(r+1\right).
\end{aligned}$$

When none of the guests label their caps ($r = 0$), we return to the scenario of Problem 2 ($\mathrm{E}[D] = m/n$). When all but one of the guests label their caps ($r = n-1$), all of the guests can be uniquely matched to their caps, and thus all available caps are matched ($\mathrm{E}[D] = m$).

Once you get the (base)ball rolling, it's easy to come up with many interesting variations on the simple matching hats problem! One might want to consider the case in which the distribution of guests among the s teams is not symmetric (each with probability $1/s$), but is asymmetric, with a guest rooting for team i with probability p_i. It's not difficult to show that applying this generalization to Problem 2B, for example, yields the following expected number of correct matches:

$$\mathrm{E}[D] = \left(\frac{m}{n}\right)\left(s - \sum_{i=1}^{s}(1-p_i)^n\right).$$

With this result, it only takes a bit of calculus to show that $\mathrm{E}[D]$ is maximized when the teams are evenly distributed.

We stumbled onto these problems while computing some lower bounds for the set of graph matching algorithms in [4], and expect that others might find their own applications. It's very satisfying to construct these problems in an analytically tractable way, and to observe the lovely properties and limiting behaviors of the solutions.

Acknowledgements We were fortunate to have a reviewer point out to us the "cap-centered" perspective, which provides very economical solutions to Problems 1A, 1C, 2A and 2C. This material is based upon work supported in part by a National Science Foundation Graduate Research Fellowship. Additional support for this work was provided by the DoD AFOSR URI for 'Architectures for Secure and Robust Distributed Infrastructures,' F49620-01-1-0365 (led by Stanford University).

References

[1] D. Bertsekas and J. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2002.

[2] H. Hathout, The old hats problem revisited, *College Math. J.*, 35 (2004) 97-102.

[3] E. Weisstein, Hypergeometric Distribution, *MathWorld–A Wolfram Web Resource*, http://mathworld.wolfram.com/HypergeometricDistribution.html (last accessed 06 June 2006)

[4] L. Zager, *Graph similarity and matching*, Master's thesis, MIT, Cambridge, MA, 2004.