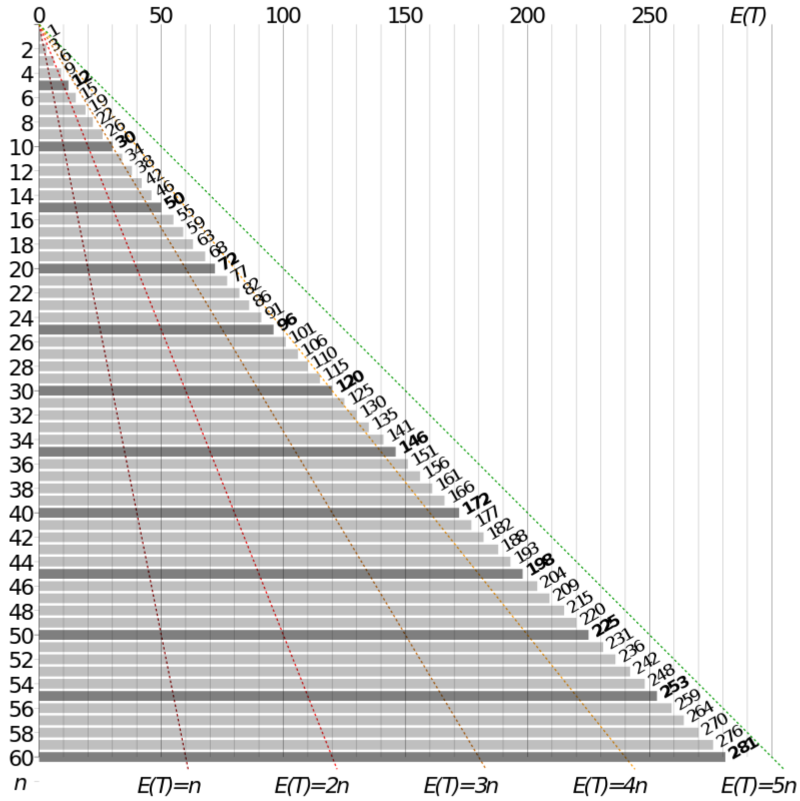
Coupon collector's problem

In [probability theory](), the **coupon collector's problem** describes the "collect all coupons and win" contests. It asks the following question: Suppose that there is an urn of n different coupons, from which coupons are being collected, equally likely, with replacement. What is the probability that more than t sample trials are needed to collect all n coupons? An alternative statement is: Given n coupons, how many coupons do you expect you need to draw with replacement before having drawn each coupon at least once? The mathematical analysis of the problem reveals that the expected number of trials needed grows as $\Theta(n\log(n))$.[1] For example, when n = 50 it takes about 225[2] trials on average to collect all 50 coupons.



Graph of number of coupons, n vs the expected number of tries (i.e., time) needed to collect them all, $E(T)$

Contents

Solution

Calculating the expectation

Let T be the time to collect all n coupons, and let t_i be the time to collect the i-th coupon after $i - 1$ coupons have been collected. Think of T and t_i as random variables. Observe that the probability of collecting a **new** coupon is $p_i = (n - (i - 1))/n$. Therefore, t_i has geometric distribution with expectation $1/p_i$. By the linearity of expectations we have:

$$\begin{aligned}\operatorname{E}(T) &= \operatorname{E}(t_1)+\operatorname{E}(t_2)+\cdots+\operatorname{E}(t_n) = \frac{1}{p_1}+\frac{1}{p_2}+\cdots+\frac{1}{p_n} \\ &= \frac{n}{n}+\frac{n}{n-1}+\cdots+\frac{n}{1} \\ &= n\cdot\left(\frac{1}{1}+\frac{1}{2}+\cdots+\frac{1}{n}\right) \\ &= n\cdot H_n.\end{aligned}$$

Here H_n is the n-th harmonic number. Using the asymptotics of the harmonic numbers, we obtain:

$$\operatorname{E}(T) = n\cdot H_n = n\log n + \gamma n + \frac{1}{2} + O(1/n),$$

where $\gamma \approx 0.5772156649$ is the Euler–Mascheroni constant.

Now one can use the Markov inequality to bound the desired probability:

$$\operatorname{P}(T \geq cnH_n) \leq \frac{1}{c}.$$

Calculating the variance

Using the independence of random variables t_i, we obtain:

$$\begin{aligned}\operatorname{Var}(T) &= \operatorname{Var}(t_1)+\operatorname{Var}(t_2)+\cdots+\operatorname{Var}(t_n) \\ &= \frac{1-p_1}{p_1^2}+\frac{1-p_2}{p_2^2}+\cdots+\frac{1-p_n}{p_n^2} \\ &< \left(\frac{n^2}{n^2}+\frac{n^2}{(n-1)^2}+\cdots+\frac{n^2}{1^2}\right) \\ &= n^2\cdot\left(\frac{1}{1^2}+\frac{1}{2^2}+\cdots+\frac{1}{n^2}\right) \\ &< \frac{\pi^2}{6}n^2\end{aligned}$$

since $\frac{\pi^2}{6} = \frac{1}{1^2}+\frac{1}{2^2}+\cdots+\frac{1}{n^2}+\cdots$ (see Basel problem).

Now one can use the Chebyshev inequality to bound the desired probability:

$$\operatorname{P}\left(|T - nH_n| \geq cn\right) \leq \frac{\pi^2}{6c^2}.$$

Tail estimates

A different upper bound can be derived from the following observation. Let Z_i^r denote the event that the i-th coupon was not picked in the first r trials. Then:

$$P\left[Z_i^r\right] = \left(1-\frac{1}{n}\right)^r \leq e^{-r/n}$$

Thus, for $r = \beta n\log n$, we have $P\left[Z_i^r\right] \leq e^{(-\beta n\log n)/n} = n^{-\beta}$.

$$P\left[T > \beta n \log n\right] = P\left[\bigcup_i Z_i^{\beta n \log n}\right] \le n \cdot P[Z_1^{\beta n \log n}] \le n^{-\beta+1}$$

Extensions and generalizations

- Paul Erdős and Alfréd Rényi proved the limit theorem for the distribution of *T*. This result is a further extension of previous bounds.

$$\mathrm{P}(T < n\log n + cn) \to e^{-e^{-c}}, \text{ as } n \to \infty.$$

- Donald J. Newman and Lawrence Shepp found a generalization of the coupon collector's problem when *m* copies of each coupon need to be collected. Let T_m be the first time *m* copies of each coupon are collected. They showed that the expectation in this case satisfies:

$$\mathrm{E}(T_m) = n\log n + (m-1)n\log\log n + O(n), \text{ as } n \to \infty.$$

Here *m* is fixed. When *m* = 1 we get the earlier formula for the expectation.

- Common generalization, also due to Erdős and Rényi:

$$\mathrm{P}\left(T_m < n\log n + (m-1)n\log\log n + cn\right) \to e^{-e^{-c}/(m-1)!}, \text{ as } n \to \infty.$$

- In the general case of a nonuniform probability distribution, according to Philippe Flajolet,[3]

$$E(T) = \int_0^\infty \left(1 - \prod_{i=1}^n (1 - e^{-p_i t})\right) dt.$$

See also

- Watterson estimator
- Birthday problem

Notes

1. Here and throughout this article, "log" refers to the natural logarithm rather than a logarithm to some other base. The use of Θ here invokes big O notation.
2. E(50) = 50(1 + 1/2 + 1/3 + ... + 1/50) = 224.9603, the expected number of trials to collect all 50 coupons. The approximation $n\log n + \gamma n + 1/2$ for this expected number gives in this case $50\log 50 + 50\gamma + 1/2 \approx 195.6011 + 28.8608 + 0.5 \approx 224.9619$.
3. Flajolet, Philippe; Gardy, Danièle; Thimonier, Loÿs (1992), "Birthday paradox, coupon collectors, caching algorithms and self-organizing search" (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.217.5965&rep=rep1&type=pdf), *Discrete Applied Mathematics*, **39** (3): 207–229, doi:10.1016/0166-218x(92)90177-c (https://doi.org/10.1016%2F0166-218x%2892%2990177-c)

References

- Blom, Gunnar; Holst, Lars; Sandell, Dennis (1994), "7.5 Coupon collecting I, 7.6 Coupon collecting II, and 15.4 Coupon collecting III", *Problems and Snapshots from the World of Probability* (https://books.google.com/books?id=KCsSWFMq2u0C&pg=PA85), New York: Springer-Verlag, pp. 85–87, 191, ISBN 0-387-94161-4, MR 1265713 (https://www.ams.org/mathscinet-getitem?mr=1265713).
- Dawkins, Brian (1991), "Siobhan's problem: the coupon collector revisited", *The American Statistician*, **45** (1): 76–82, doi:10.2307/2685247 (https://doi.org/10.2307%2F2685247), JSTOR 2685247 (https://www.jstor.org/stable/2685247).

- Erdős, Paul; Rényi, Alfréd (1961), "On a classical problem of probability theory" (http://www.renyi.hu/~p_erdos/1961-09.pdf) (PDF), *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, **6**: 215–220, MR 0150807 (https://www.ams.org/mathscinet-getitem?mr=0150807).
- Newman, Donald J.; Shepp, Lawrence (1960), "The double dixie cup problem", *American Mathematical Monthly*, **67**: 58–61, doi:10.2307/2308930 (https://doi.org/10.2307%2F2308930), MR 0120672 (https://www.ams.org/mathscinet-getitem?mr=0120672)
- Flajolet, Philippe; Gardy, Danièle; Thimonier, Loÿs (1992), "Birthday paradox, coupon collectors, caching algorithms and self-organizing search" (http://algo.inria.fr/flajolet/Publications/alloc2.ps.gz), *Discrete Applied Mathematics*, **39** (3): 207–229, doi:10.1016/0166-218X(92)90177-C (https://doi.org/10.1016%2F0166-218X%2892%2990177-C), MR 1189469 (https://www.ams.org/mathscinet-getitem?mr=1189469).
- Isaac, Richard (1995), "8.4 The coupon collector's problem solved", *The Pleasures of Probability* (https://books.google.com/books?id=a_2vsIx4FQMC&pg=PA80), Undergraduate Texts in Mathematics, New York: Springer-Verlag, pp. 80–82, ISBN 0-387-94415-X, MR 1329545 (https://www.ams.org/mathscinet-getitem?mr=1329545).
- Motwani, Rajeev; Raghavan, Prabhakar (1995), "3.6. The Coupon Collector's Problem", *Randomized algorithms* (https://books.google.com/books?id=QKVY4mDivBEC&pg=PA57), Cambridge: Cambridge University Press, pp. 57–63, MR 1344451 (https://www.ams.org/mathscinet-getitem?mr=1344451).

External links

- "Coupon Collector Problem (http://demonstrations.wolfram.com/CouponCollectorProblem/)" by Ed Pegg, Jr., the Wolfram Demonstrations Project. Mathematica package.
- *How Many Singles, Doubles, Triples, Etc., Should The Coupon Collector Expect? (http://www.math.rutgers.edu/~zeilberg/mamarim/mamarimhtml/coupon.html)*, a short note by Doron Zeilberger.