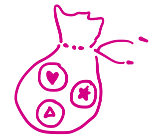
Coupon-Collector Problem

- There are n distinct balls in a bag 
- You pick one at a time with replacement → put it back in the box after you pick it
- You get a coupon  each time you pick a new ball → pick a ball you have not already picked before

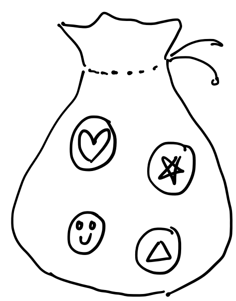
What is the expected number of ball-picks you have to do until you have all n coupons?

X := "the number of" ball-picks until you have all n coupons
hint to use Linearity of Expection

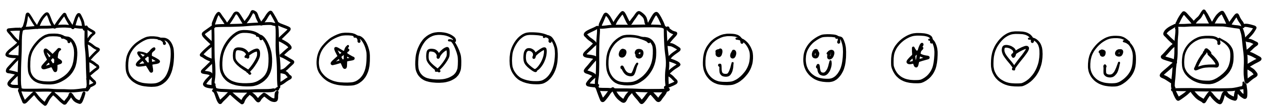
$\mathbb{E}(X) = ?$

Let's do a toy example first:

Pretend $n = 4$



Here's one way we could keep picking with replacement



It took us 13 whole tries to get the 4 distinct coupons

$X = 13$ in this example case

hmm... this seems pretty tricky. After all, once you pick a new ball and get a coupon for it — you can never get that coupon again.

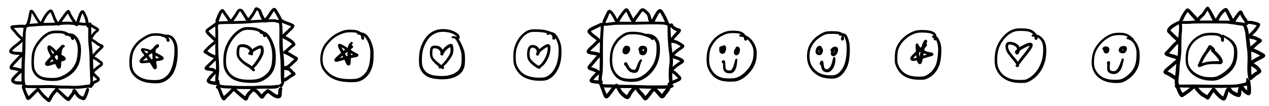
As you keep picking new balls, there are less new balls to pick from. You're like "using up" the new balls

So as you keep picking, it gets harder and harder to get the next coupon. The probability of getting a coupon at one time is not the same as the probability of getting a coupon at another time.

hmm... if only there were some way to split X up so we could use Linearity of Expectation

$X = \sum$?

Let's go back to our toy example...



1 2 4 6

$X = 13$

$1 + 2 + 4 + 6 = 13$

OMG! WOW! Adding up the number of ball-picks until the next new ball/coupon is a nice way to add up to X

$k \in [n]$

$X_k :=$ the number of ball-picks until you get the k^{th} new one — after you've already gotten $k-1$ coupons

$$X = \sum_{k=1}^{n} X_k$$

Cool! Now we've split X up into something we can apply Linearity of Expectation to

$$\underbrace{\mathbb{E}(X)}_{\text{what you want}} = \mathbb{E}\Big(\underbrace{\sum_{k=1}^{n}}_{\text{Substitute a convinient quanitity...}} X_k\Big) = \underbrace{\sum_{k=1}^{n} \mathbb{E}(X_k)}_{\text{...so that we can apply LOE}}$$

By Linearity of Expectation

Awesome! Now to find $\mathbb{E}(X)$, we just need to find $\mathbb{E}(X_k)$. The goal was to smartly pick an X_k which has an $\mathbb{E}(X_k)$ that is easier to calculate than the original $\mathbb{E}(X)$

$\mathbb{E}(X_k) = ?$ hmm... even if it easier than $\mathbb{E}(X)$... $\mathbb{E}(X_k)$ still seems tricky

... we still need to do a bit more work and figure out more about X_k

Maybe we should try and find the probability that X_k is a certain value

After all $\mathbb{E}(X_k) = \sum_r r \cdot \mathbb{P}(X_k = r)$

$\mathbb{P}(X_K = r) = ?$

What's the probability that it takes exactly r ball-picks until we get the K^{th} newest ball / get the K^{th} coupon

hint to use Geometric

i.e. What's the probability that we pick old balls $r-1$ times and then pick a new ball (the K^{th} new ball) on the r^{th} try

K-1 balls are old (i.e. failures)

we want r-1 failures

the rest of the balls are new (i.e. successes)

$$\mathbb{P}(X_K = r) = \left(\frac{K-1}{n}\right)^{r-1} \left(\frac{n-(K-1)}{n}\right)^{1}$$

we want the last one to be a success (for a total of r ball-picks)

*We know K-1 balls are old by how we defined X_K

$$P(X_K = r) = \Big(\underbrace{\frac{K-1}{n}}_{1-p}\Big)^{r-1} \Big(\underbrace{\frac{n-(K-1)}{n}}_{p}\Big)^{1}$$

Woah! This looks ~Geometric(p)

$(1-p)^{r-1}(p)^1$

Nice! Now that we know that it's Geometrically distributed we already know the formula for its Expected Value — we don't even have to do any extra work.

We Know that if $V \sim Geom(p)$ then $\mathbb{E}(V) = \frac{1}{p}$ So for X_K this is...

$$X_K \sim Geom\left(\frac{n-(K-1)}{n}\right)$$

$$\mathbb{E}(X_K) = \frac{1}{\frac{n-(K-1)}{n}} = \frac{n}{n-K+1}$$

Yay! Let's put this all together to finally get $\mathbb{E}(X)$

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} \mathbb{E}(X_k)$$ Linearity of Expectation

$$= \sum_{k=1}^{n} \frac{n}{n-k+1}$$ Plug in what we found for $\mathbb{E}(X_k)$

$$= n \sum_{k=1}^{n} \frac{1}{n-k+1}$$ Factor out n

$$= n \sum_{a=1}^{n} \frac{1}{a}$$ Re-index

This is a fun trick you can do with complicated sums.

Notice how:

$$\sum_{k=1}^{n} \frac{1}{n-k+1} = \frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \dots + \frac{1}{1}$$

We can re-order the terms:

$$\frac{1}{1} + \frac{1}{2} + \dots + \frac{1}{n-1} + \frac{1}{n} = \sum_{a=1}^{n} \frac{1}{a}$$

$$\approx n \log_e n$$

You don't need to know that part, but it's pretty neat.

<u>Conclusion:</u> We expect to take $n \sum_{a=1}^{n} \frac{1}{a}$ ($\approx n \log_e n$) ball-picks until we collect all n distinct coupons

Coupon Collector tldr

$X :=$ the number of ball-picks until you have all n coupons

$k \in [n]$

$X_k :=$ the number of ball-picks until you get the k^{th} new one — after you've already gotten $k-1$ coupons

$$X = \sum_{k=1}^{n} X_k$$

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} \mathbb{E}(X_k)$$ By Linearity of Expectation

$$= \sum_{k=1}^{n} \frac{n}{n-k+1}$$ By result below in $\boxed{\text{box}}$

$$= n \sum_{a=1}^{n} \frac{1}{a}$$ Re-indexing

$$\boxed{\begin{gathered} \mathbb{P}(X_k = r) = \left(\frac{k-1}{n}\right)^{r-1} \left(\frac{n-(k-1)}{n}\right)^{1} \\ X_k \sim \text{Geom}\left(\frac{n-(k-1)}{n}\right) \\ \mathbb{E}(X_k) = \frac{n}{n-k+1} \end{gathered}}$$