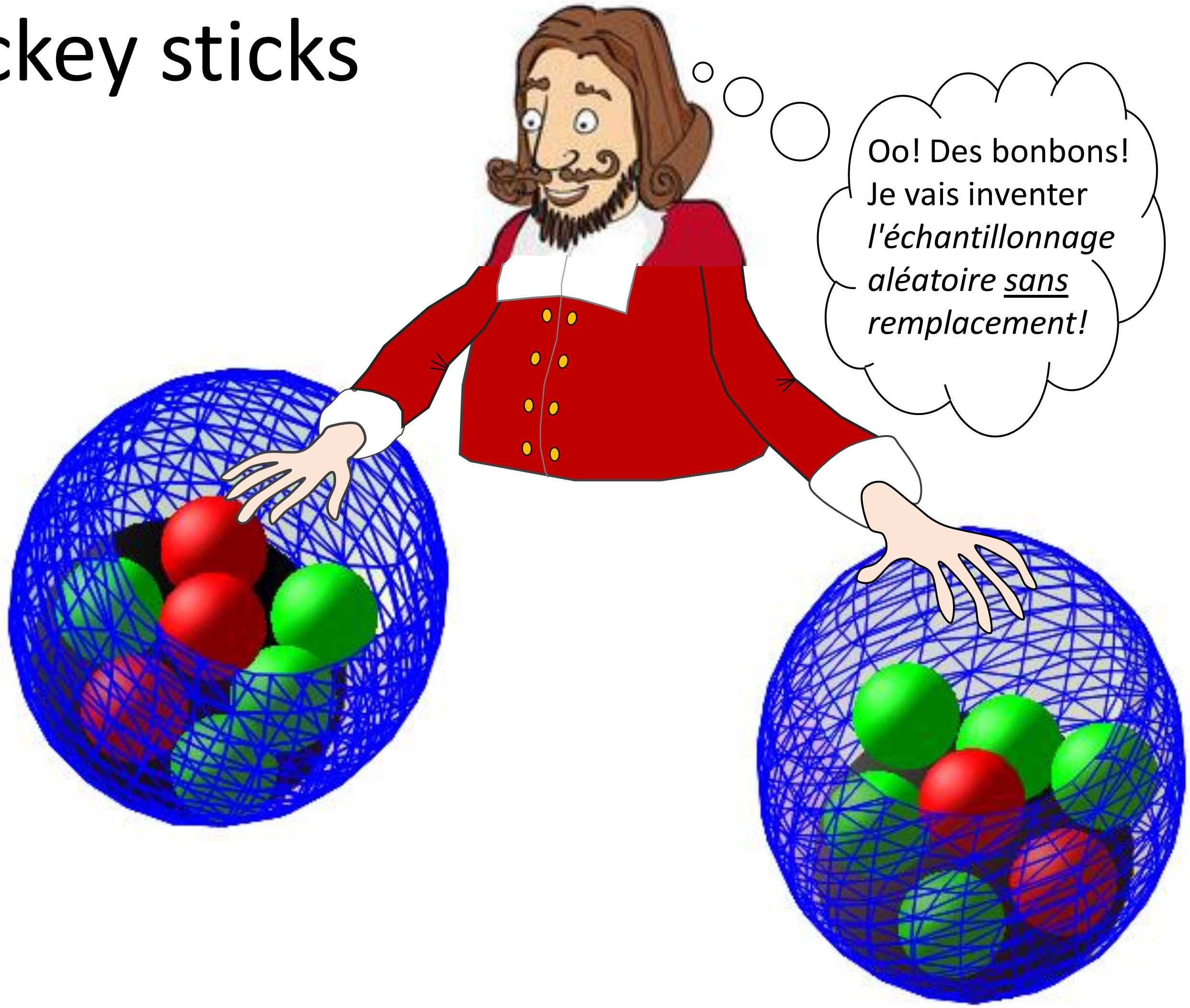
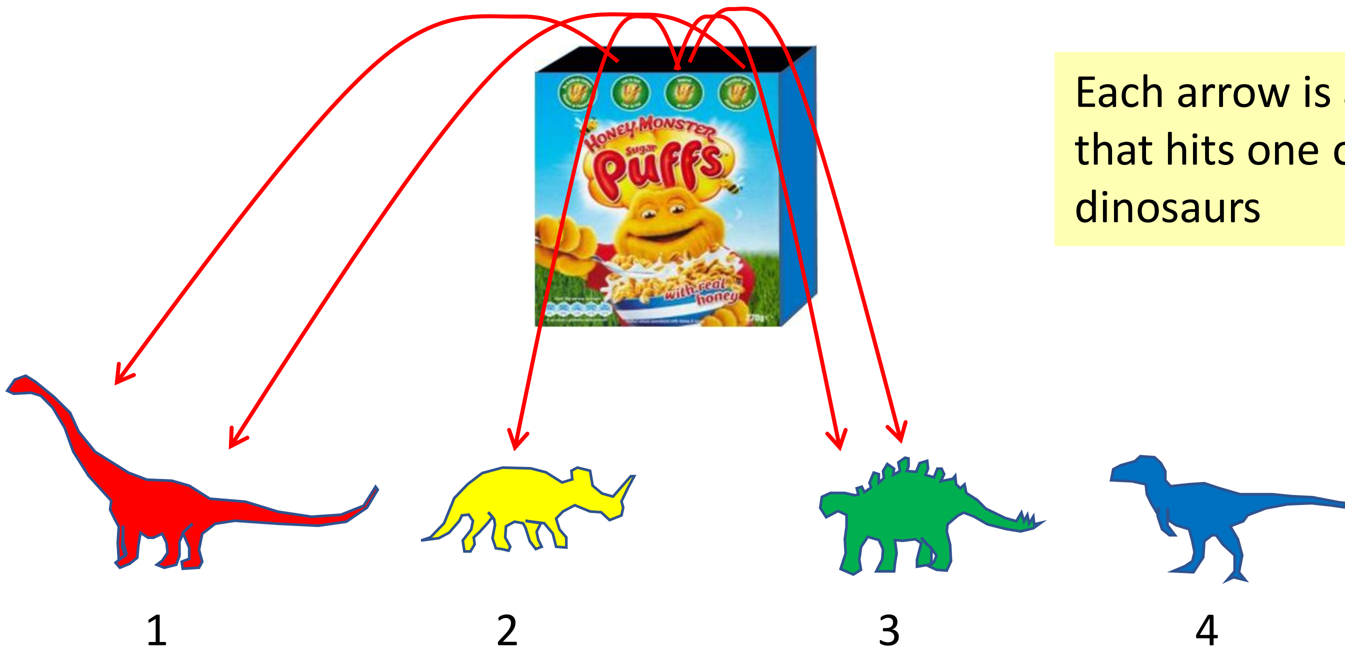
Random sampling with hockey sticks



Recap: Balls in bins I

Putting n identical balls into m bins: coupon collector's but with a single magic sugar puffs packet that generates a stream of dinosaurs



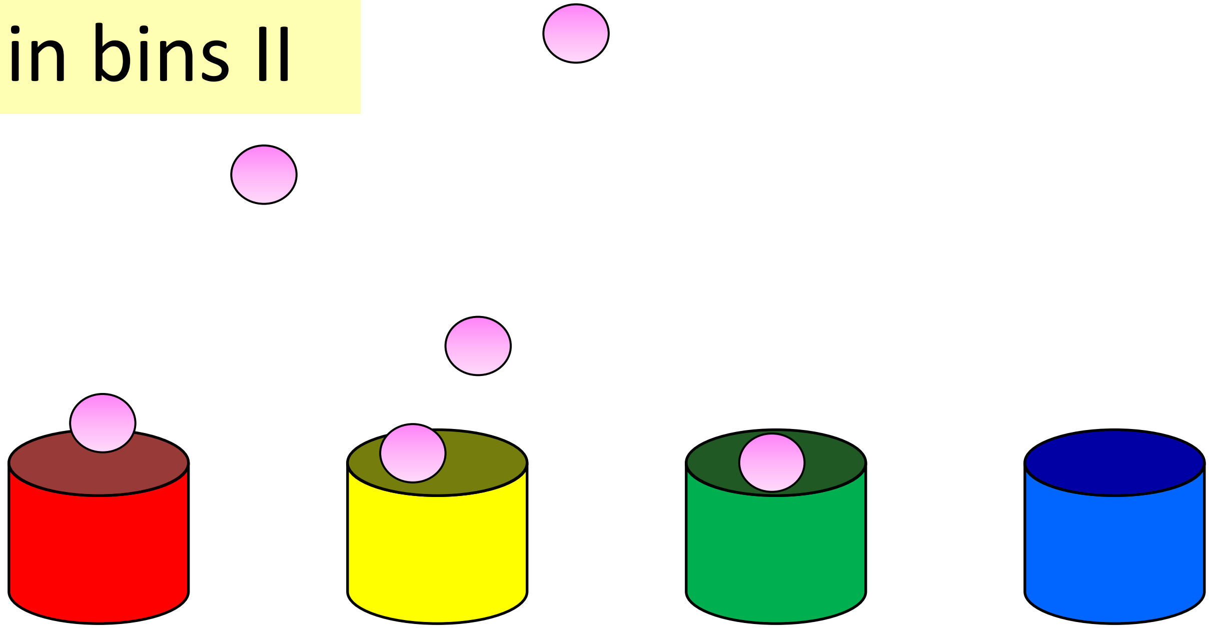
Each arrow is a ball that hits one of the dinosaurs

What is the probability of getting every one of m dinosaurs in a stream of n from the packet?

How many ways of putting n identical balls into m bins?

How many of these place at least one ball into each bin?

Recap: Balls in bins II



So now we have m bins into which we place n balls with repetition allowed and we want at least one ball in each bin.

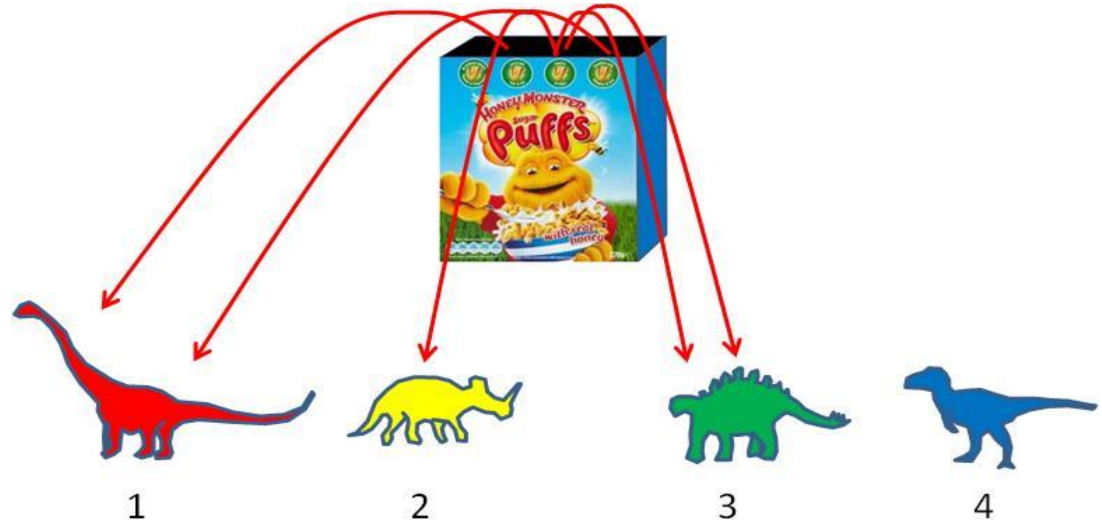
Basic count: number of ways to place n balls into m bins: $\binom{m-1+n}{m-1}$

E.g. 6 balls into 4 bins: take $4-1+6$ places. Choose $4-1$ bin markers and the remaining 6 places are occupied by the balls.

 = 1 ball, 0 balls, 3 balls, 2 balls

Recap: Balls in bins III

What is the probability of getting every one of n dinosaurs in a stream of m from the packet?



How many ways of putting n identical balls into m bins?

$$\binom{m-1+n}{m-1}$$

How many of these place at least one ball into each bin?

Take m of the n balls and place one in each bin.

Now place the remaining $n-m$ balls in all possible ways:

$$\binom{m-1+n-m}{m-1} = \binom{n-1}{m-1}$$

So probability of a surjection with 5 balls from the magic packet is

$$\frac{\binom{4}{3}}{\binom{8}{3}} = \frac{4}{56} \approx 0.07$$

Trying it out experimentally

$$\binom{m-1+n}{m-1}$$

Let’s take m = 4 bins and n = 5 balls. I’ll generate 5 random numbers in the range 1, …, 4.

$$\binom{n-1}{m-1}$$

Do this 1000 times. Check how many trials have every bin nonempty.

$$\frac{\binom{4}{3}}{\binom{8}{3}} = \frac{4}{56} \approx 0.07$$

I get 236. Try again: 250. Again: 256.
Far higher than 7% predicted.

What am I doing wrong?

First rookie mistake!

The words “place 5 **identical** balls into 4 bins” means it happens **simultaneously**.
The balls are not identical if they arrive one by one.

E.g. toss a coin 5 times. This is putting 5 balls into a bin called H and a bin called T.

The outcomes are:

H	**T**
5	0
4	1
3	2
2	3
1	4
0	5

What is the probability of 5 heads? From the table it is 1/6.

Coin tossing: observed or not?

E.g. toss a coin 5 times. This is putting 5 balls into a bin called H and a bin called T.

But what if we **watch** as the coin is tossed. Then the outcomes we may observe are:

H	H	H	H	H
H	H	H	H	T
H	H	H	T	H
		⋮		
H	T	T	T	T
T	T	T	T	T

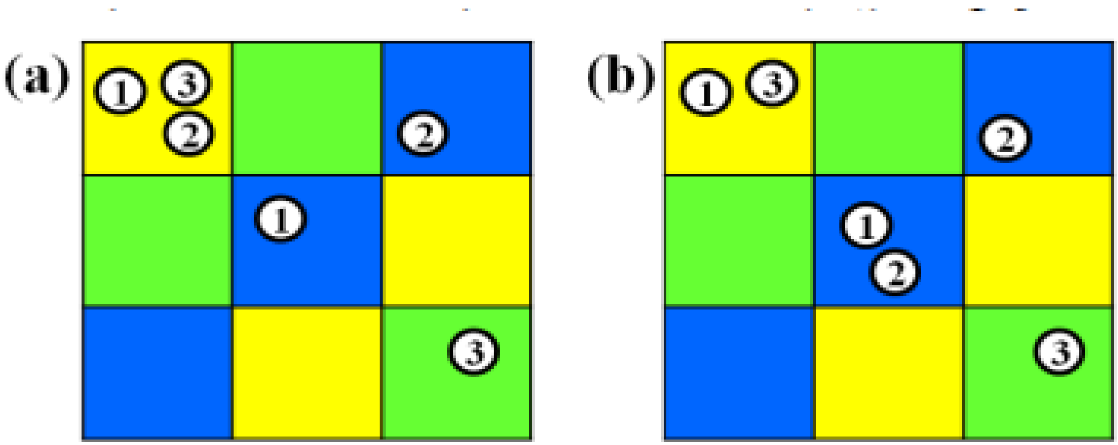
What is the probability of 5 heads? From the table it is 1/32. The table is a **function** from a sequence of 5 tosses to two outcomes.

So if we don't watch the coin tossing we are much more likely to get 5 heads!?

Putting balls in bins is a 'process' isn't it?

But saying "place 5 identical balls into 4 bins" seems to suggest a physical act of 'placement'.

In fact our application really did mean **simultaneous** placement:



n players commit privately to a placement of 2 tokens each on the cells of an $\sqrt{m} \times \sqrt{m}$ grid.

Here, $n = 3$, $m = 9$.

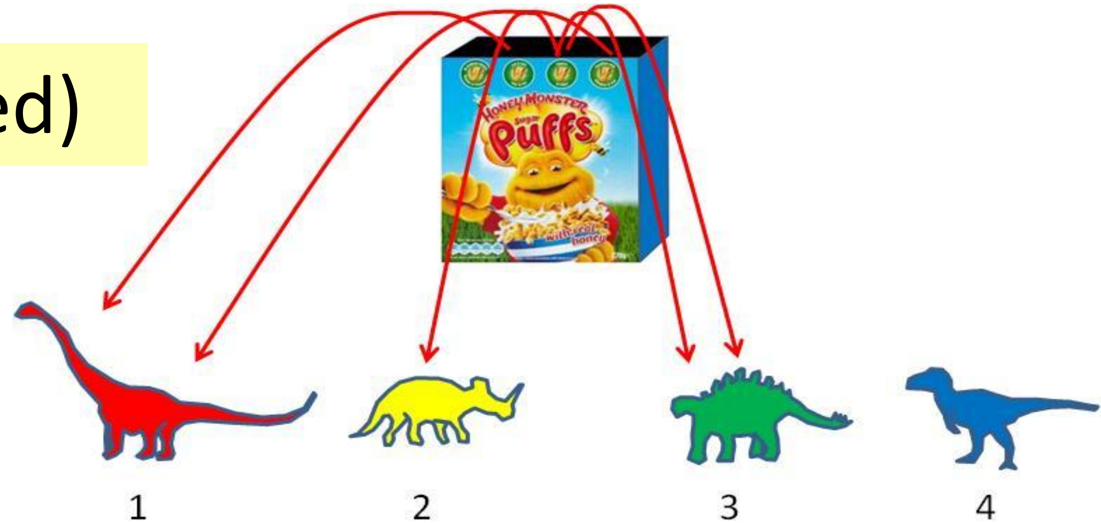
They reveal their choices and their stake money is shared equally according to the tokens which single-occupy cells.

The grid is cosmetic. Really I'm just putting 6 tokens into 9 bins.

But the whole example is a bit cosmetic really, isn't it? How often do you ask questions about pre-existing configurations?

And when can you be sure that it isn't a **function** that your probabilities are measuring.

The magic packet (corrected)



What is the probability of getting every one of n dinosaurs in a ~~stream~~ **simultaneous burst** of m from the packet?

How many ways of putting n identical balls into m bins? $\binom{m-1+n}{m-1}$

How many of these place at least one ball into each bin?

Take m of the n balls and place one in each bin.

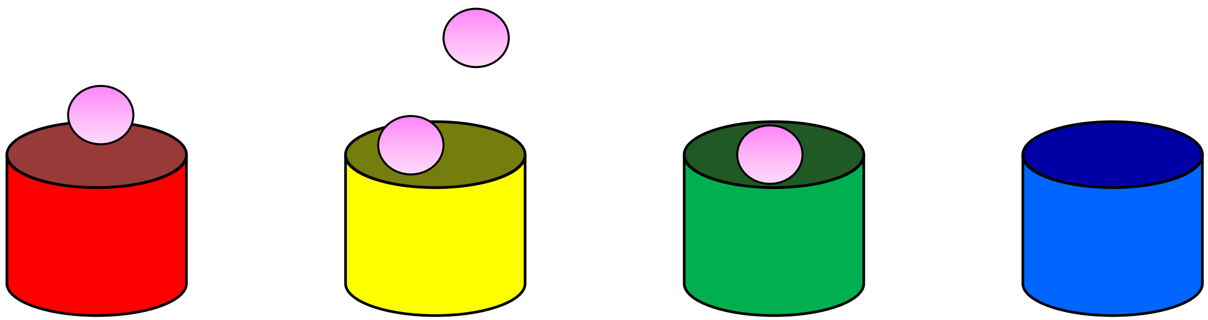
Now place the remaining $n-m$ balls in all possible ways: $\binom{m-1+n-m}{m-1} = \binom{n-1}{m-1}$

So probability of a surjection with 5 balls from the magic packet is

$$\frac{\binom{4}{3}}{\binom{8}{3}} = \frac{4}{56} \approx 0.07$$

How do I simulate this and see, experimentally, a rough 7% figure?

Balls in bins: random sampling I



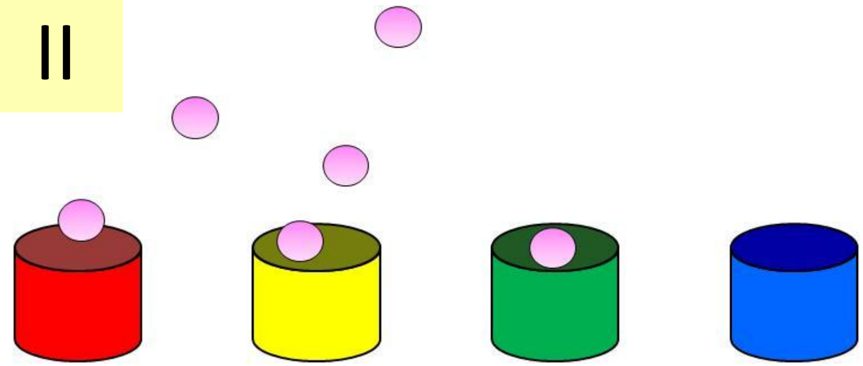
Basic count: number of ways to place n balls into m bins: $\binom{m-1+n}{m-1}$

E.g. 6 balls into 4 bins: take $4-1+6$ places. Choose $4-1$ bin markers and the remaining 6 places are occupied by the balls.

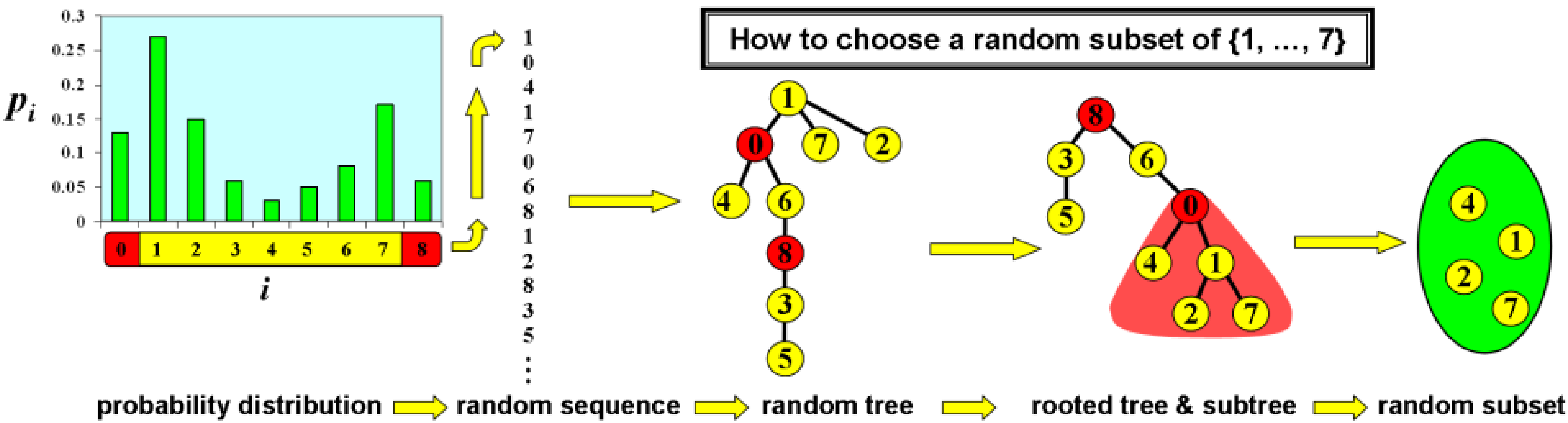
 = 1 ball, 0 balls, 3 balls, 2 balls

Obvious way to choose a random placement of n balls into m bins: choose $m-1$ bin marker placements out of $m-1+n$. But this is choosing a random subset.

Balls in bins: random sampling II

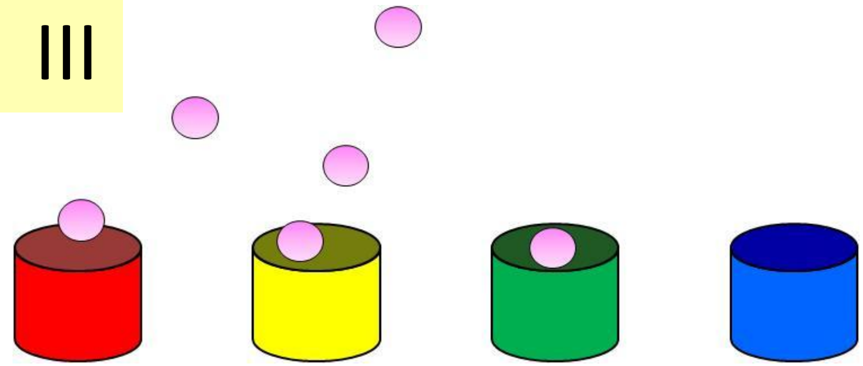


Choosing a random subset is well-studied. E.g. this application by Jim Pitman of the Abel–Hurwitz binomial theorem:



But maybe we can find something simpler for the specific case of randomly putting balls in bins?

Balls in bins: random sampling III



Basic idea based on nested sums (following Butler and Karasik, 2010):

$$\sum_{i_1=0}^{n}\sum_{i_2=0}^{i_1}\cdots\sum_{i_m=0}^{i_{m-1}} 1 = \binom{m+n}{n}$$

$$\updownarrow$$

$$n \geq i_1 \geq i_2 \geq \cdots \geq i_m \geq 0$$

$$\updownarrow$$

$$(n - i_1, i_1 - i_2, \ldots, i_m - 0)$$

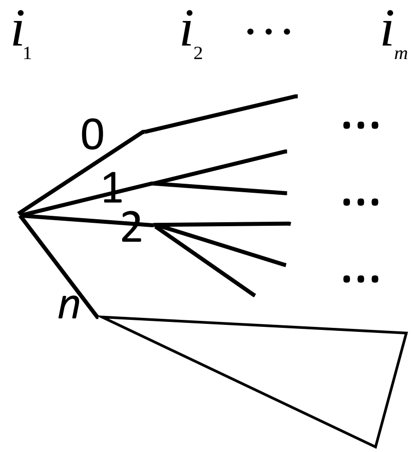
$(m + 1)$-tuple of nonnegative integers summing to n.

putting $m + 1$ balls into n bins

Balls in bins: random sampling IV

This will place n balls into $m + 1$ bins!

$$\sum_{i_1=0}^{n} \sum_{i_2=0}^{i_1} \cdots \sum_{i_m=0}^{i_{m-1}} 1 \longrightarrow$$



$\binom{m+n}{n}$ leaves

$$\updownarrow$$

$$n \ge i_1 \ge i_2 \ge \cdots \ge i_m \ge 0$$

$$\updownarrow$$

$$(n - i_1, i_1 - i_2, \ldots, i_m - 0)$$

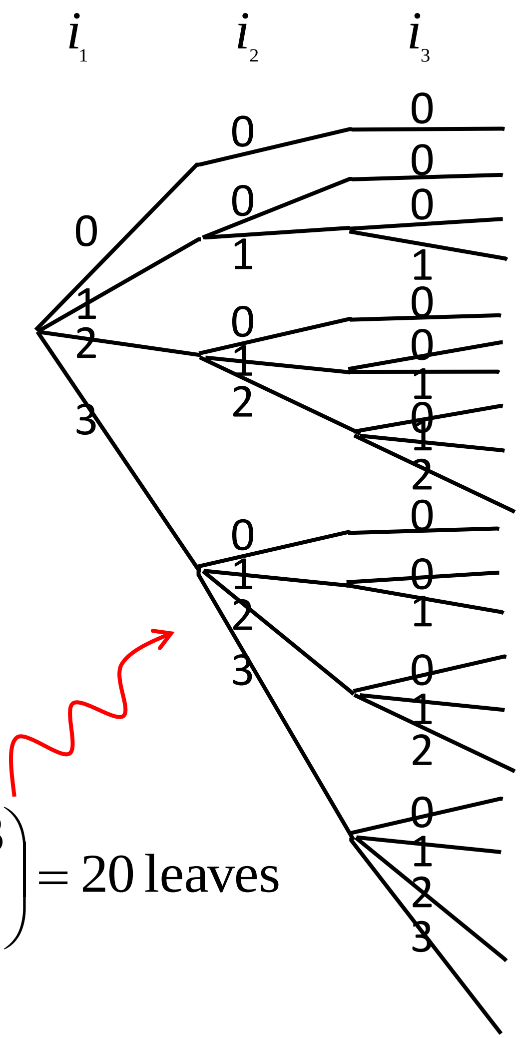
Choose a random leaf by repeatedly choosing a tree branch, starting at the root. This will give values for

$$i_1, i_2, \ \ldots, i_m$$

which will give us our $(m + 1)$-tuple

Balls in bins: random sampling V

E.g. place 3 balls into 4 bins



$$\binom{4-1+3}{4-1} = 20 \text{ leaves}$$

Choose branches: 2, 1, 1

$$3 \ge i_1 \ge i_2 \ge i_3 \ge 0$$

2 1 1

↓

$$(n - i_1, i_1 - i_2, i_2 - i_3, i_3 - 0)$$

(3 – 2, 2 – 1, 1 – 1 , 1 – 0)
(1, 1, 0, 1)

Choose branches: 3, 3, 3

$$3 \ge i_1 \ge i_2 \ge i_3 \ge 0$$

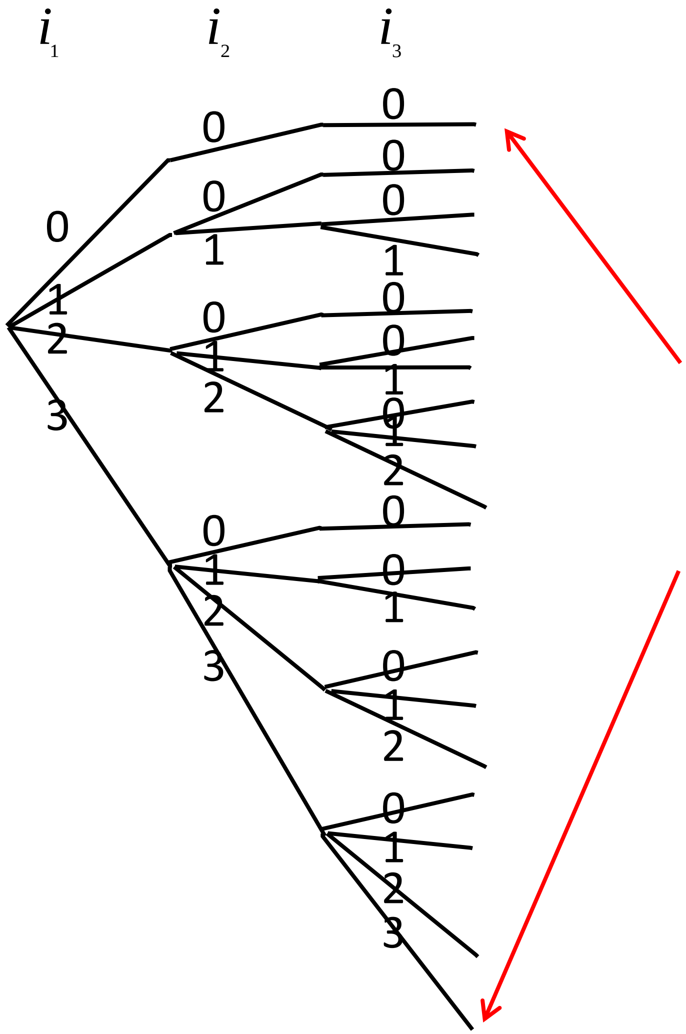
3 3 3

↓

$$(n - i_1, i_1 - i_2, i_2 - i_3, i_3 - 0)$$

(3 – 3, 3 – 3, 3 – 3 , 3 – 0)
(0, 0, 0, 3)

Second rookie mistake!



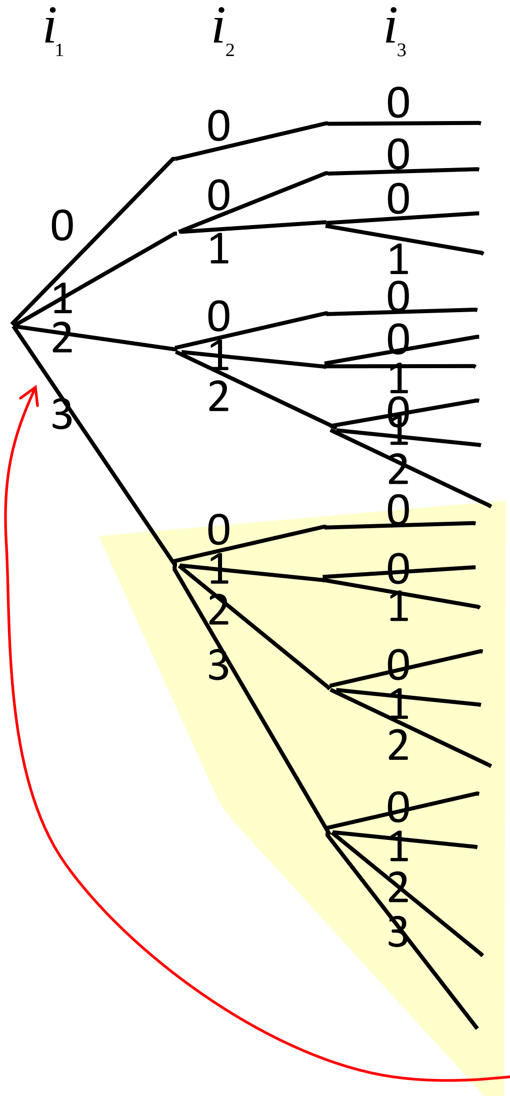
Any leaf (= any 4-tuple) can be chosen.

But completely **not** uniformly at random!

This leaf (ball placement 3, 0, 0, 0) is chosen with probability 1/4.

This leaf (ball placement 0, 0, 0, 3) is chosen with probability (1/4)^3

Choosing tree branches non-uniformly



There are $\binom{6}{3} = 20$ leaves.

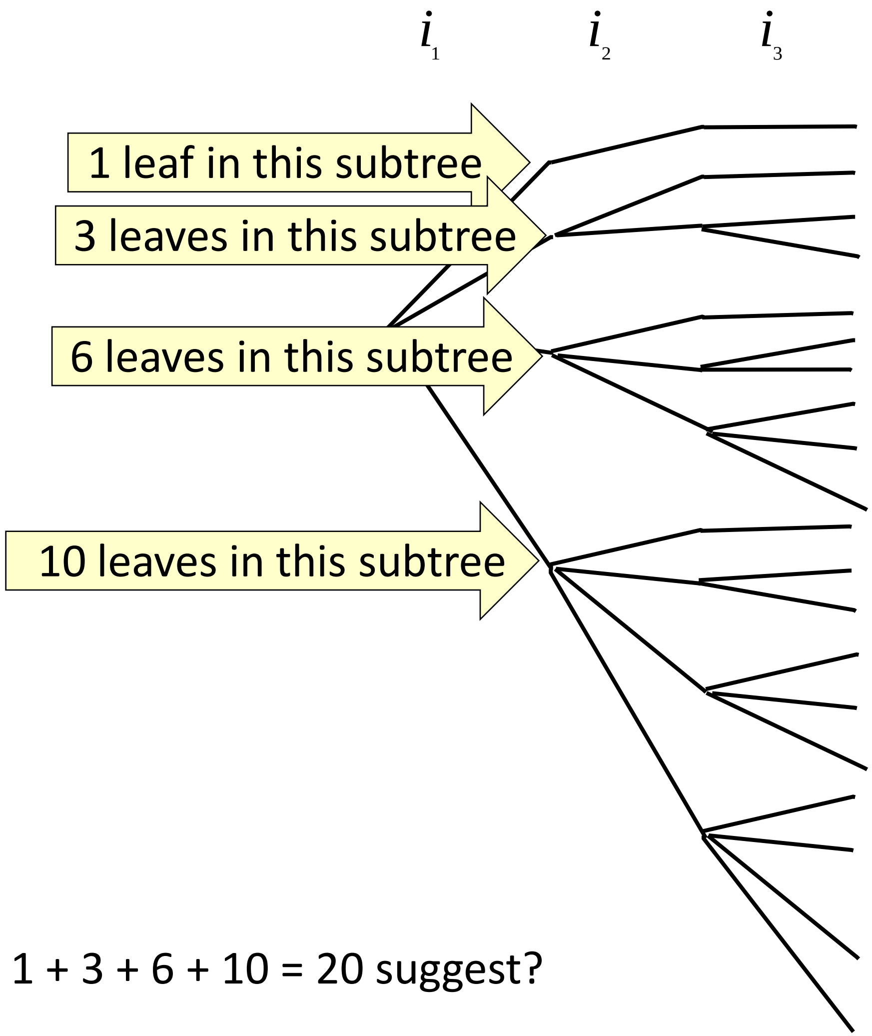
Each leaf is a placement of 3 balls into 4 bins.

We need each leaf to be chosen with probability 1/20.

This subtree accounts for 10 of the 20 leaves.

So this branch must be chosen 50% of the time.

Choosing tree branches non-uniformly



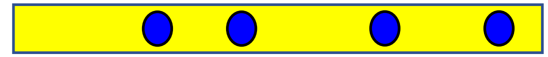
What does 1 + 3 + 6 + 10 = 20 suggest?

The hockey stick identity

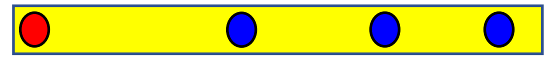
	0	1	2	3	4	5	6	7	8	...
0	1									
1	1	1								
2	1	2	1							
3	1	3	3	1						
4	1	4	6	4	1					
5	1	5	10	10	5	1				
6	1	6	15	20	15	6	1			
7	1	7	21	35	35	21	7	1		
8	1	8	28	56	70	56	28	8	1	
⋮										

$$\binom{n+1}{m+1} = \sum_{r=m}^{n} \binom{r}{m}$$

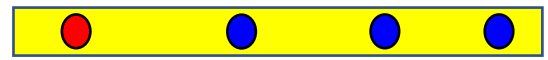
Combinatorial proof:



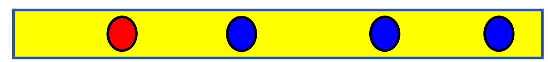
$$\binom{n+1}{m+1}$$



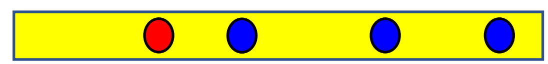
$$= \binom{n}{m}$$



$$+ \binom{n-1}{m}$$



$$+ \binom{n-2}{m}$$



$$+ \binom{n-3}{m} + \dots$$

Chosing tree branches using hockey stick weightings

	0	1	2	3	4	5	6	7	8	...
0	1									
1	1	1								
2	1	2	1							
3	1	3	3	1						
4	1	4	6	4	1					
5	1	5	10	10	5	1				
6	1	6	15	20	15	6	1			
7	1	7	21	35	35	21	7	1		
8	1	8	28	56	70	56	28	8	1	
⋮										

Choose a number from 1 to 20

Suppose we choose 7

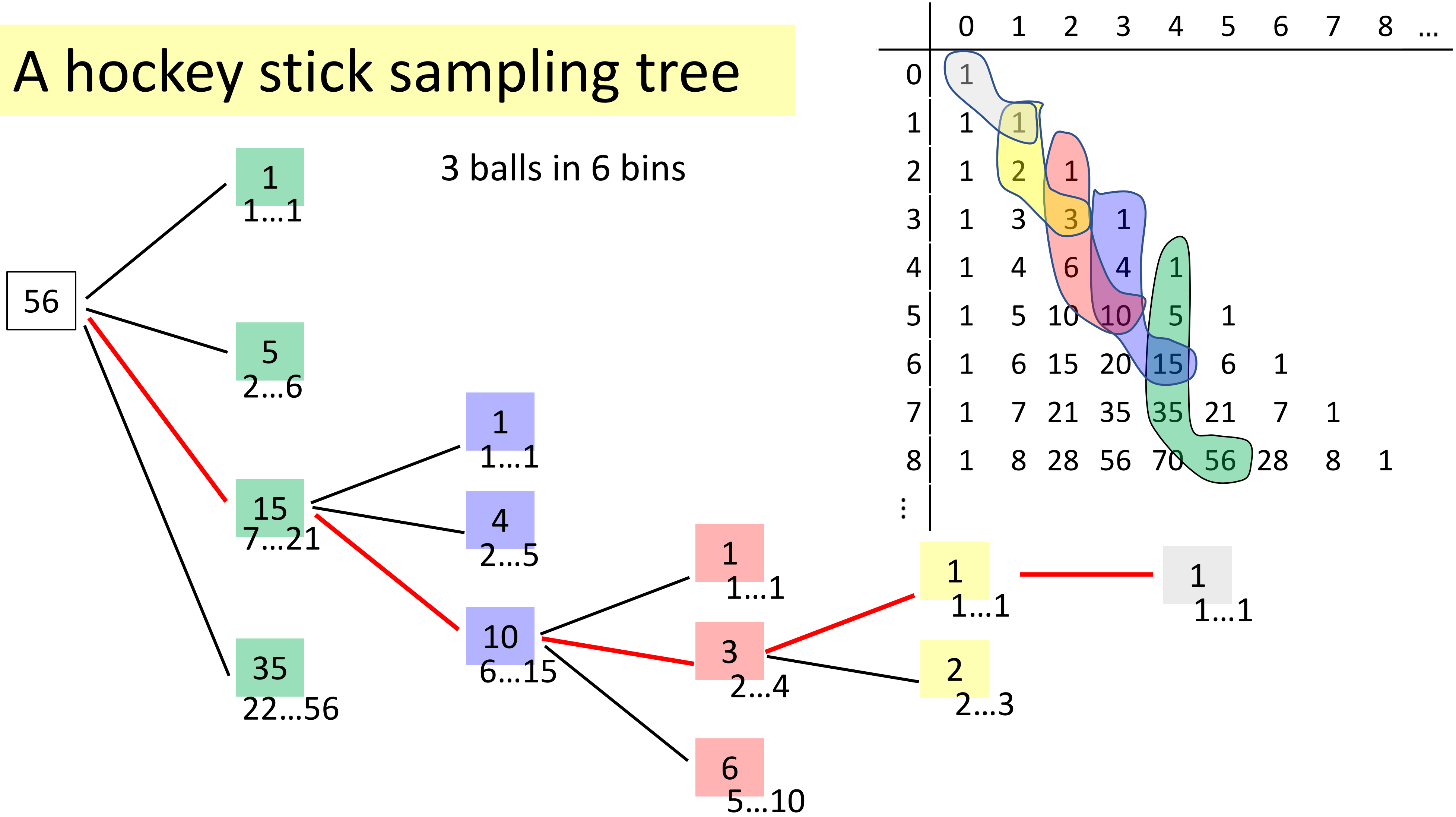
Find where it falls in the hockey stick distribution for 20

It is in the subtree with leaves 5 – 10.

Repeat using hockey stick at 6

A hockey stick sampling tree

3 balls in 6 bins

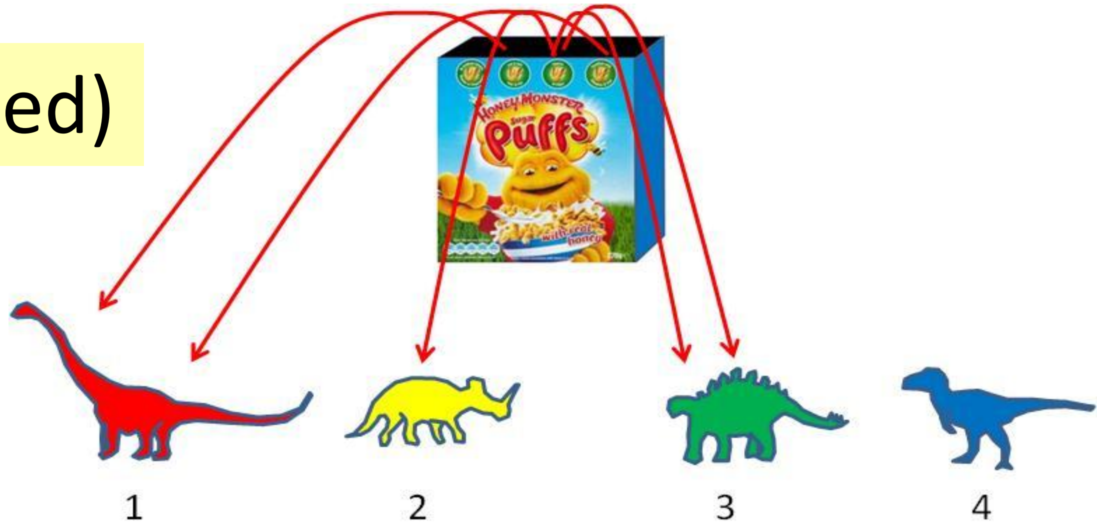


$i_1 = 2$ $i_2 = 2$ $i_3 = 1$ $i_4 = 0$ $i_5 = 0$

3 ≥ 2 ≥ 2 ≥ 1 ≥ 0 ≥ 0 ≥ 0

1 0 1 1 0 0

The magic packet (simultated)



What is the probability of getting every one of n dinosaurs in a ~~stream~~ **simultaneous burst** of m from the packet?

Probability of a surjection with 5 balls from the magic packet is

$$\frac{\binom{4}{3}}{\binom{8}{3}} = \frac{4}{56} \approx 0.07$$

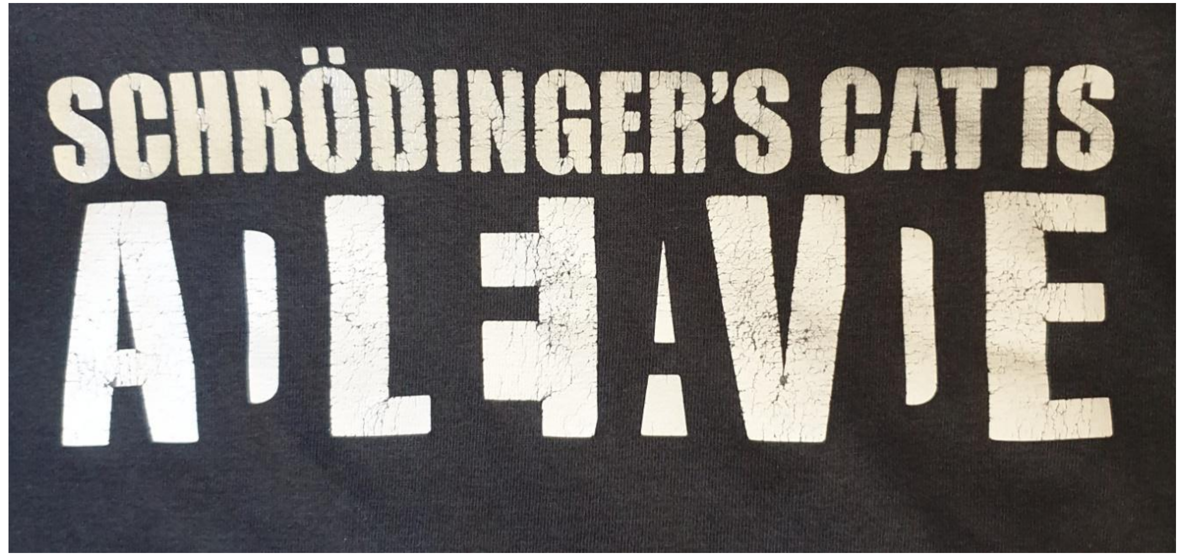
```
total := 0 :
for j from 1 to 100 do
  onto := 0 :
  for i from 1 to 1000 do
    if trial(4, 5, 0) = 0 then onto := onto + 1 end;
  end:
  total := total + onto/1000 :
end:
evalf(total/100);
```

0.07161000000

```
evalf(4/56)
```

0.07142857143

A rookie question!



10 coin tosses.

If all are heads then gun fires.

Is the cat watching the coin??