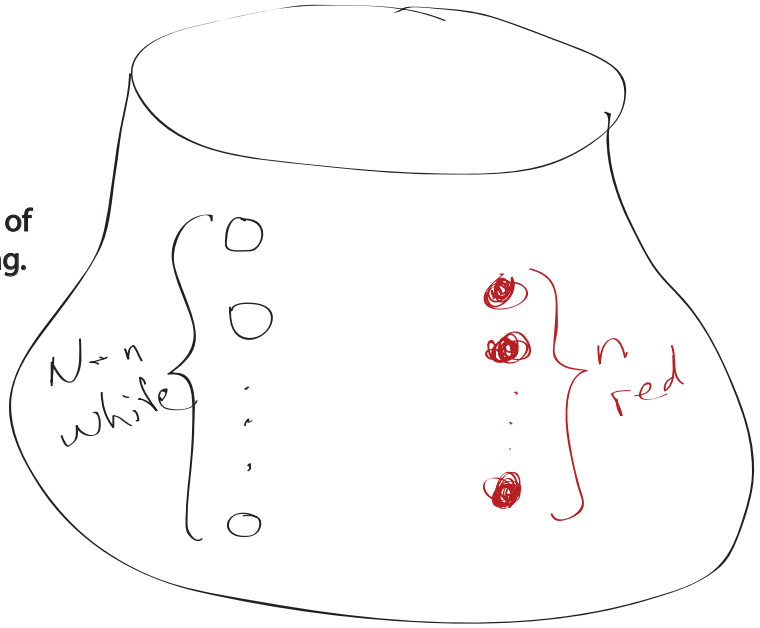
N = total number of balls in the bag
Let n be the number of red balls.
So N-n is the number of white balls.

Let Y be a random variable that describes the number of red balls chosen, if we pull m balls from the bag.

So the parameters are

- **N**
- **n**
- **m**
- **Y**



The bag is considered to be opaque, and so when you reach in, you can see what ball you are taking.
The bag only allows one ball to be removed at a time.
There are two options:

- Option 1: to pull with replacement. So we take a ball out of the bag, check out its colour, and then put the ball back into the bag.
- Option 2: to pull without replacement. So we take a ball out of the bag, check out its colours, and leave it out of the bag.

The mathematics associated with Option 2 (w/out replacement) is more difficult that Option 1.

What is the probability of pull a red ball? n/N

If you pull with replacement, what is the prob of pulling a red ball, after you just pulled a red ball? n/N
And if you pull without repacement? (n-1)/(N-1)

The fact that the probabilities change when you pull without replacement makes the mathematics more difficult.

Assume there are N=20K genes and there is a gene signature for e.g. EGFR activation with n=100 genes.

Suppose you complete a microarray experiment comparing EGFR activated mice to wild-type mice and you identify m=200 genes that are differentially expressed (e.g. from a t-test after correcting for multiple testing).

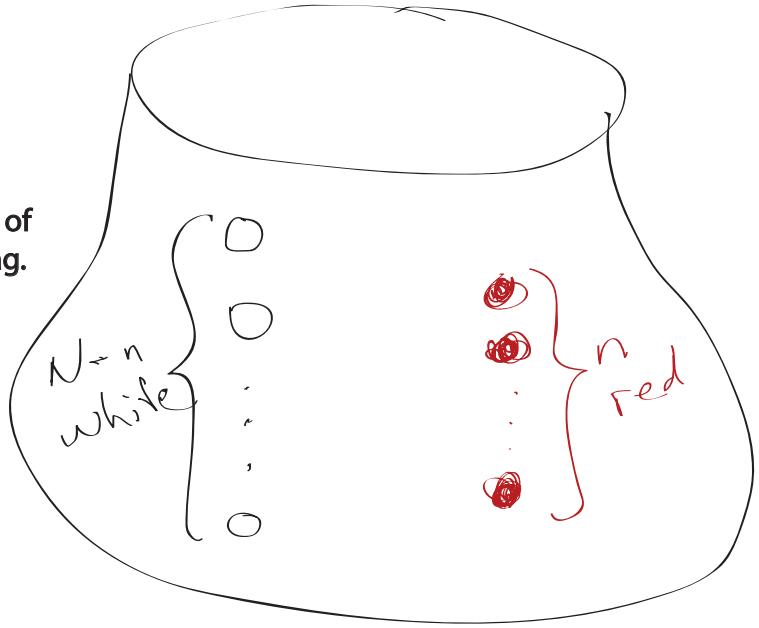
Suppose furthermore that there are Y=10 genes in common between the two signatures. You want to know if this is a surprisingly high degree of overalp.

N = total number of balls in the bag
Let n be the number of red balls.
So N-n is the number of white balls.

Let Y be a random variable that describes the number of red balls chosen, if we pull m balls from the bag.

So the parameters are

N
n
m
Y



Assume there are N=20K genes and there is a gene signature for e.g. EGFR activation with n=500 genes.

Suppose you complete a microarray experiment comparing EGFR activated mice to wild-type mice and you identify m=100 genes that are differentially expressed (e.g. from a t-test after correcting for multiple testing).

Suppose furthermore that there are Y=10 genes in common between the two signatures. You want to know if this is a surprisingly high degree of overalp.

For a very "back of the envelope" rough approximation, note that each time you pull from the bag with replacement, the probability of a red is n/N .... therefore after m=100 pulls you might expect

$$100 * n / N$$

red balls. Using our numbers here this would be approximatley 2.5 = 100*(500/20K) balls. (The real number is 10 so compared to 2.5 in this "rough" manner maybe we do have enrichment ... though to say if that is significant with a p-value of 0.05... might be close!)

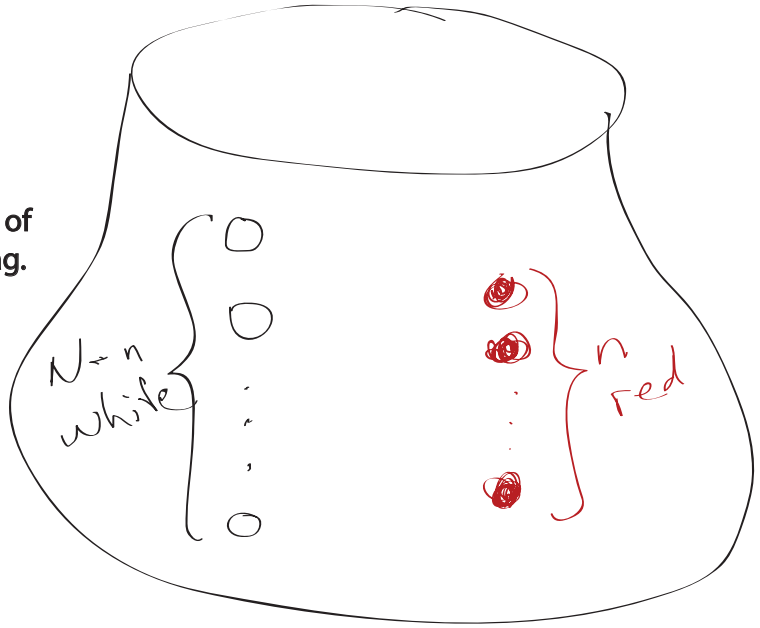
Still using "replacement", a more formal and correct way to do this is to use the binomial distribution.

Here we are again asking what the probability is that we pull Y=10 red balls from a bag with N-n white balls and n red balls.

N = total number of balls in the bag
Let n be the number of red balls.
So N-n is the number of white balls.

Let Y be a random variable that describes the number of red balls chosen, if we pull m balls from the bag.

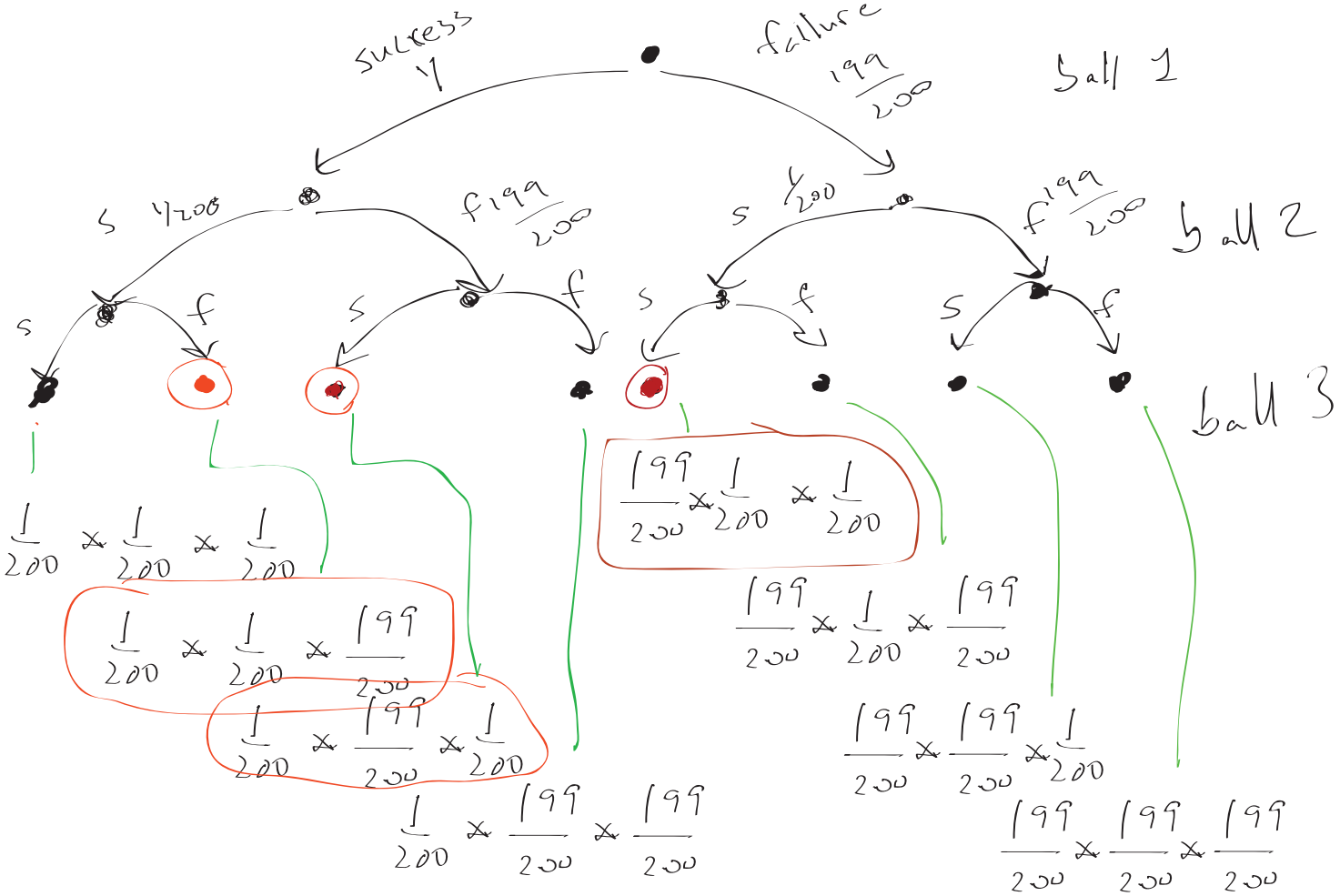
So the parameters are
N
n
m
Y



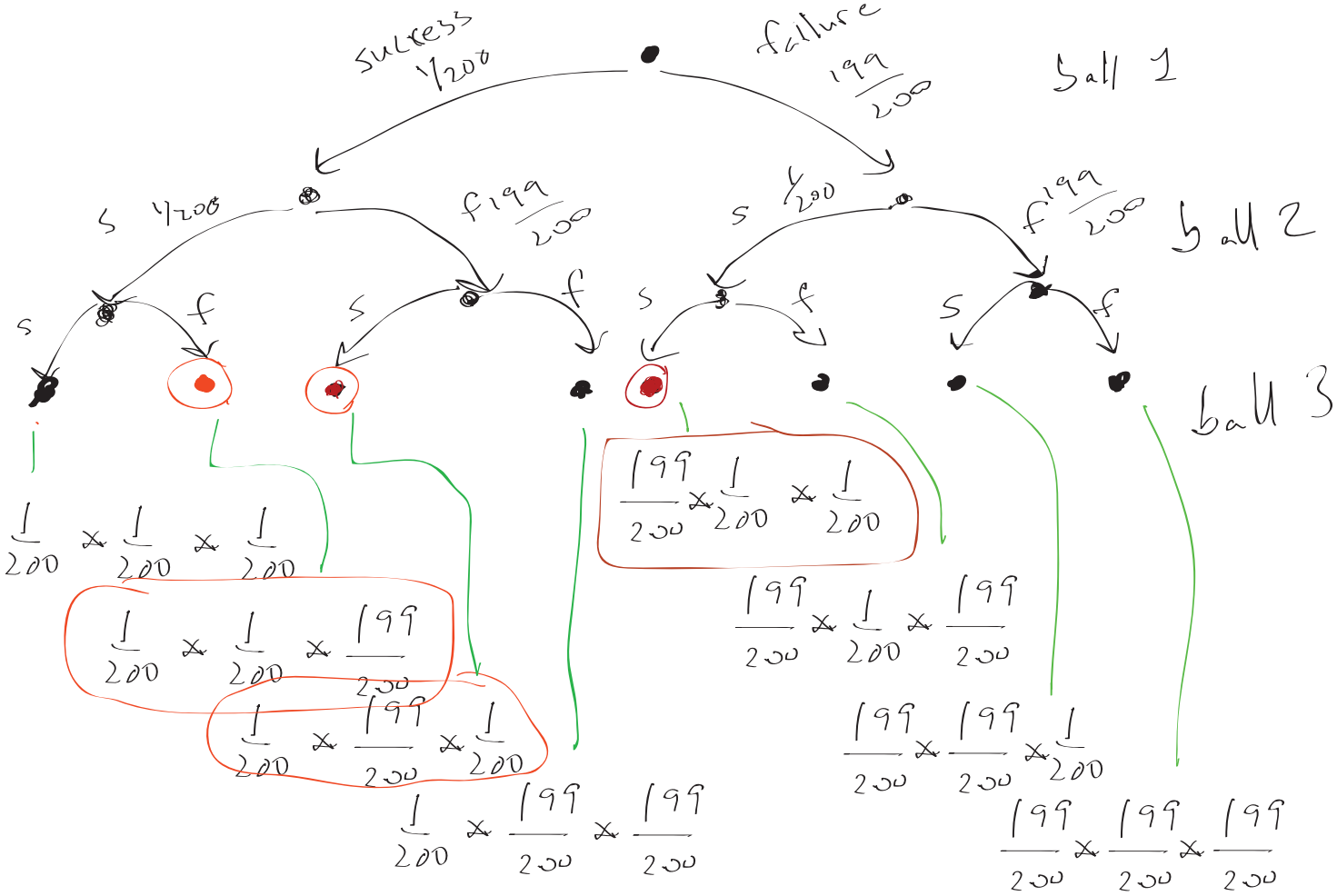
Here we are again asking what the probability is that we pull Y=10 red balls from a bag with N-n white balls and n red balls.

As before the probability of **success** (pulling a red ball) is n/N. The probability of **failure** is (1-n/N). We have Y=10 successes and m-Y = 200-10 = 190 failures.

To make things simpler here, let's suppose that m = 3 and Y =2. (So there is 1 failure). The following tree describes all possibilities. The lineages that have 2 successes and 1 failure are marked in red..



To make things simpler here, let's suppose that m = 3 and Y =2. (So there is 1 failure). The following tree describes all possibilities. The lineages that have 2 successes and 1 failure are marked in red..



There are 3 ways to get 2 successes and 1 failure, the probability that one of these three events occurs is:

$$\left(\frac{1}{200}\times\frac{1}{200}\times\frac{199}{200}\right)+\left(\frac{1}{200}\times\frac{199}{200}\times\frac{1}{200}\right)+\left(\frac{199}{200}\times\frac{1}{200}\times\frac{1}{200}\right)=0.0000746$$

More generally ...

There are 3 ways to get 2 successes and 1 failure, the probability that one of these three events occurs is:

$$\left(\frac{1}{200} \times \frac{1}{200} \times \frac{199}{200}\right) + \left(\frac{1}{200} \times \frac{199}{200} \times \frac{1}{200}\right) + \left(\frac{199}{200} \times \frac{1}{200} \times \frac{1}{200}\right) = 0.0000746$$

$$\left(\begin{array}{c}\#\text{ of ways} \\ \text{to get 2 successes} \\ \text{of 3}\end{array}\right) \cdot \left(\begin{array}{c}\text{Prob. associated with} \\ \text{getting 2 of 3 successes}\end{array}\right)$$

$$= 3 \cdot \left(\frac{1}{200} \times \frac{1}{200} \times \frac{199}{200}\right)$$

$$= 3 \cdot \underbrace{\left(\frac{1}{200} \times \frac{1}{200}\right)}_{\text{success}} \underbrace{\left(\frac{199}{200}\right)}_{\text{failures}}$$

$$= 3 \cdot (\text{prob. of success})^{\#\text{success}} \cdot (\text{prob of failure})^{\#\text{failures}}$$

$$= 3 \cdot \left(\text{prob. of success}\right)^{\#\text{successes}} \cdot \left(\text{prob of failure}\right)^{\#\text{failures}}$$

Most generally ...

Binomial Probability Formula.

$$\text{Prob}[Y \text{ successes of } m]$$

$$= \binom{m}{y} \cdot p^{Y} \cdot (1-p)^{m-Y}$$

$$\binom{m}{y} = \frac{m!}{y!\,(m-y)!}$$

where p = prob of success.

In our previous example

$$= \binom{200}{10} \cdot \left(\frac{n}{N}\right)^{10} \cdot \left(1-\frac{n}{N}\right)^{190}$$

$$= \binom{200}{10} \cdot \left(\frac{1}{200}\right)^{10} \cdot \left(\frac{199}{200}\right)^{190}$$

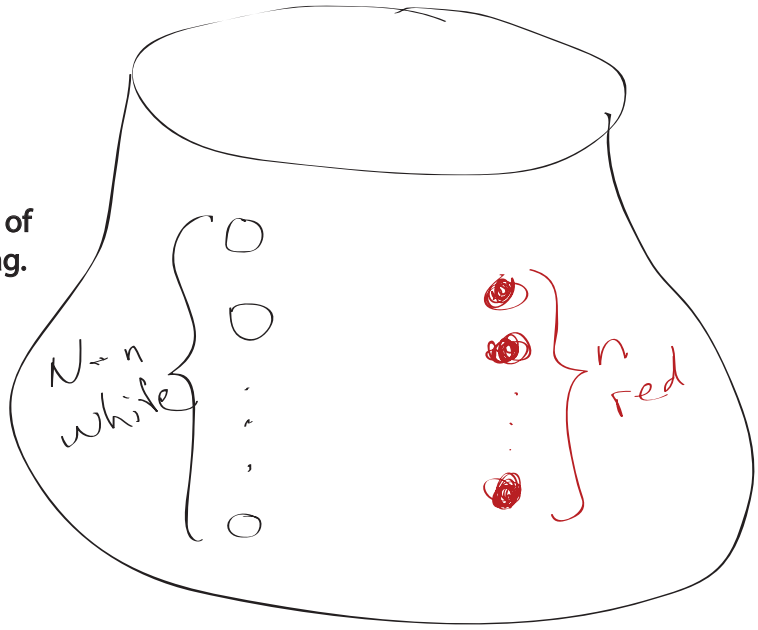
(`binom()` in R computes this)

But a more accurate estimation is possible if we do not allow replacement (once a ball is picked it is removed from the bag).

N = total number of balls in the bag
Let n be the number of red balls.
So N-n is the number of white balls.

Let Y be a random variable that describes the number of red balls chosen, if we pull m balls from the bag.

So the parameters are
N
n
m
Y



Assume there are N=20K genes and there is a gene signature for e.g. EGFR activation with n=100 genes.

Suppose you complete a microarray experiment comparing EGFR activated mice to wild-type mice and you identify m=200 genes that are differentially expressed (e.g. from a t-test after correcting for multiple testing).

Suppose furthermore that there are Y=10 genes in common between the two signatures. You want to know if this is a surprisingly high degree of overalp.

Hypergeometric Prob. distribution fnc

$$Pr_Y(y) = \frac{\binom{n}{y}\cdot\binom{N-n}{m-y}}{\binom{N}{m}}$$

Assume there are N=20K genes and there is a gene signature for e.g. EGFR activation with n=100 genes.

Suppose you complete a microarray experiment comparing EGFR activated mice to wild-type mice and you identify m=200 genes that are differentially expressed (e.g. from a t-test after correcting for multiple testing).

Suppose furthermore that there are Y=10 genes in common between the two signatures. You want to know if this is a surprisingly high degree of overalp.

Hypergeometric Prob. distribution func

$$Pr_Y(y) = \frac{\binom{n}{y} \cdot \binom{N-n}{m-y}}{\binom{N}{m}}$$

$$= \frac{\binom{100}{10} \cdot \binom{20K-100}{200-10}}{\binom{20K}{200}}$$