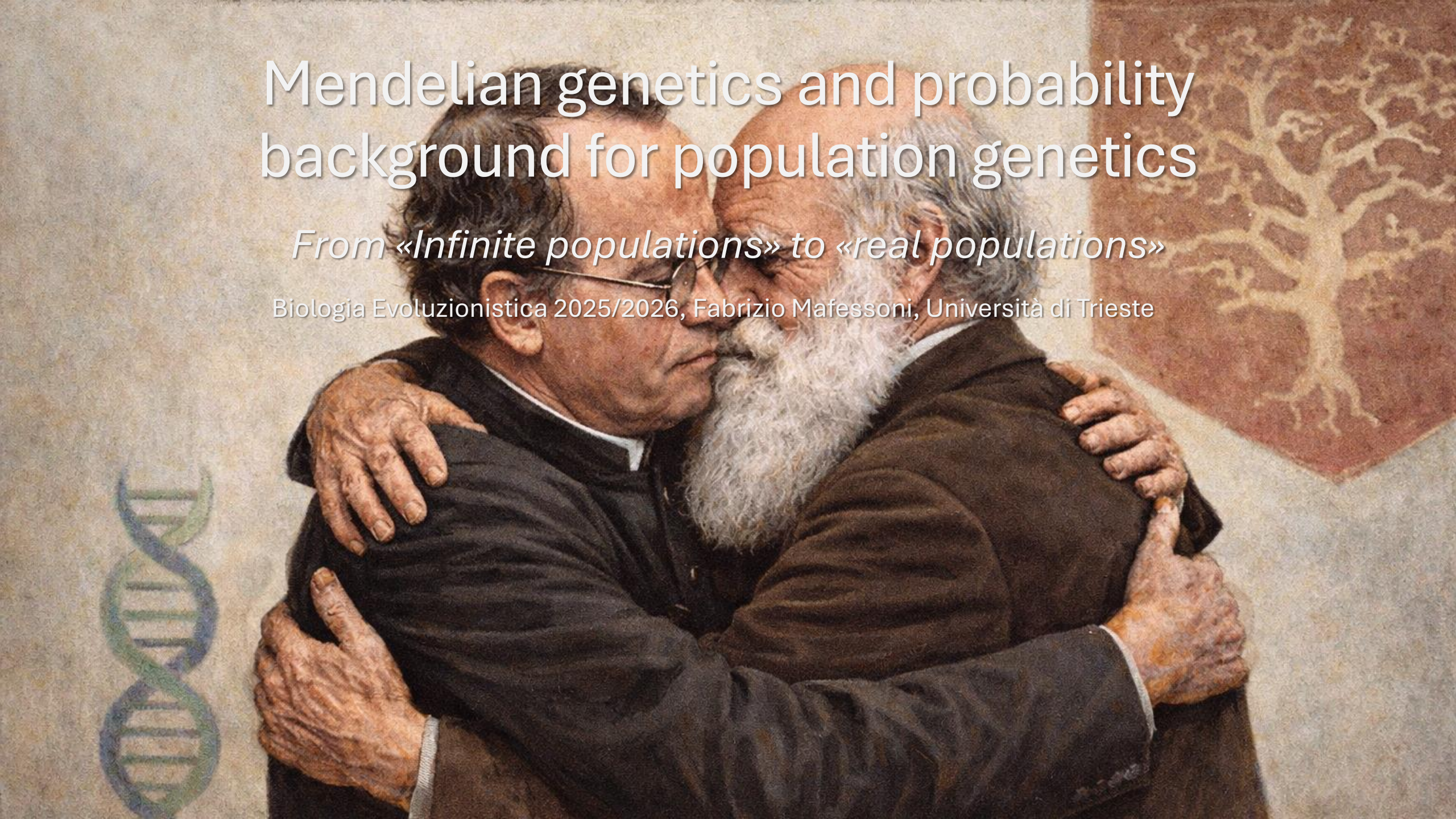
Mendelian genetics and probability background for population genetics

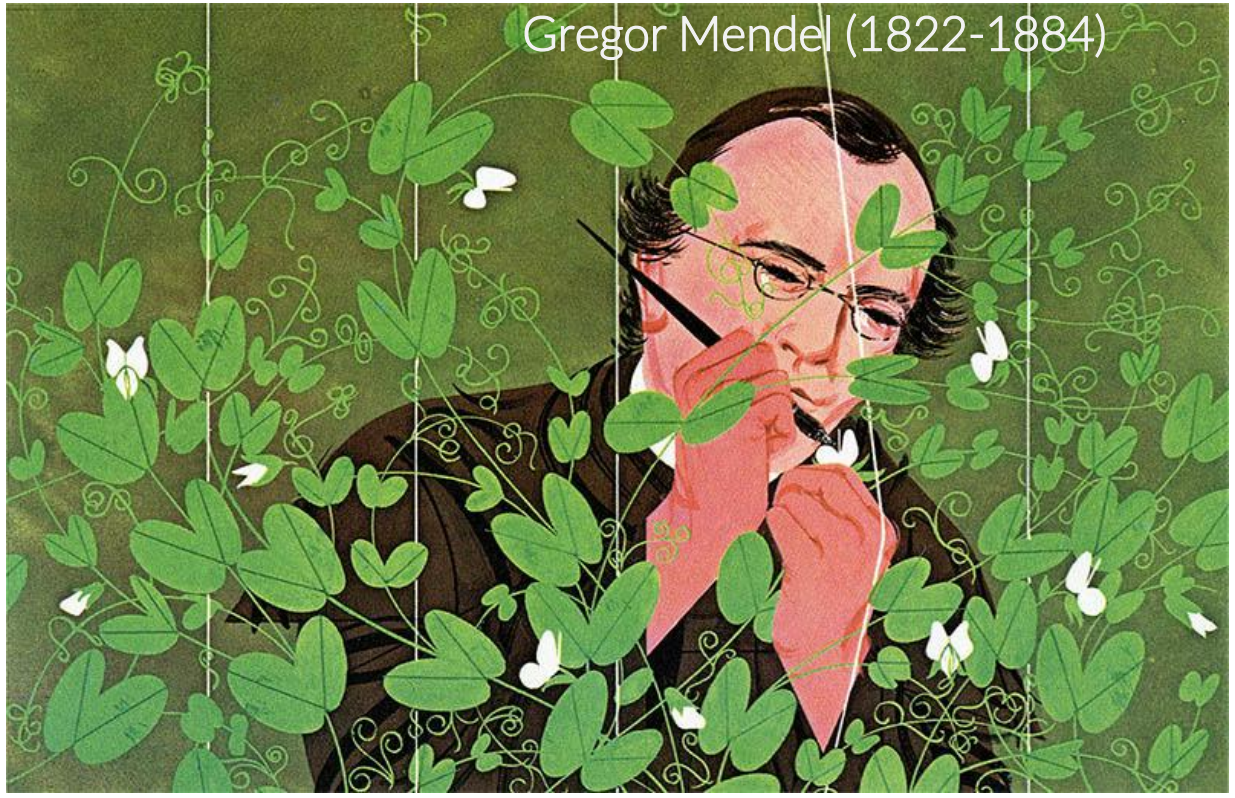
From «Infinite populations» to «real populations»

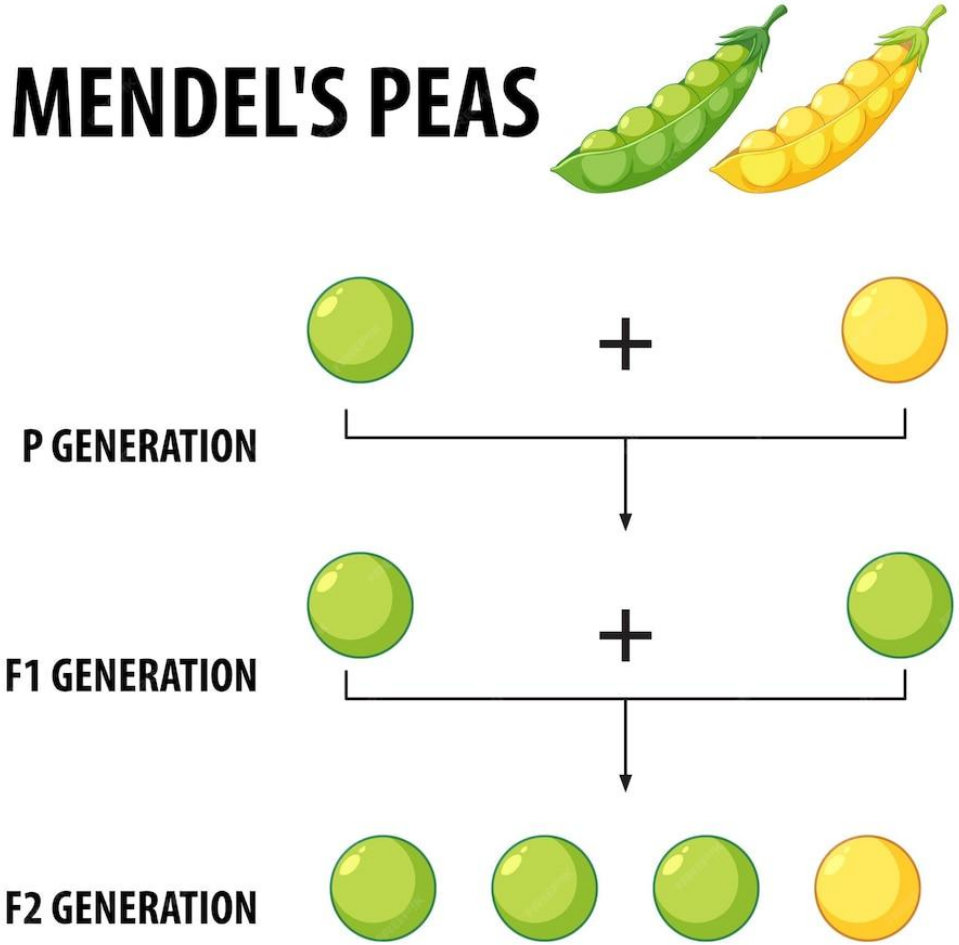
Biologia Evoluzionistica 2025/2026, Fabrizio Mafessoni, Università di Trieste



Mendel's laws of inheritance

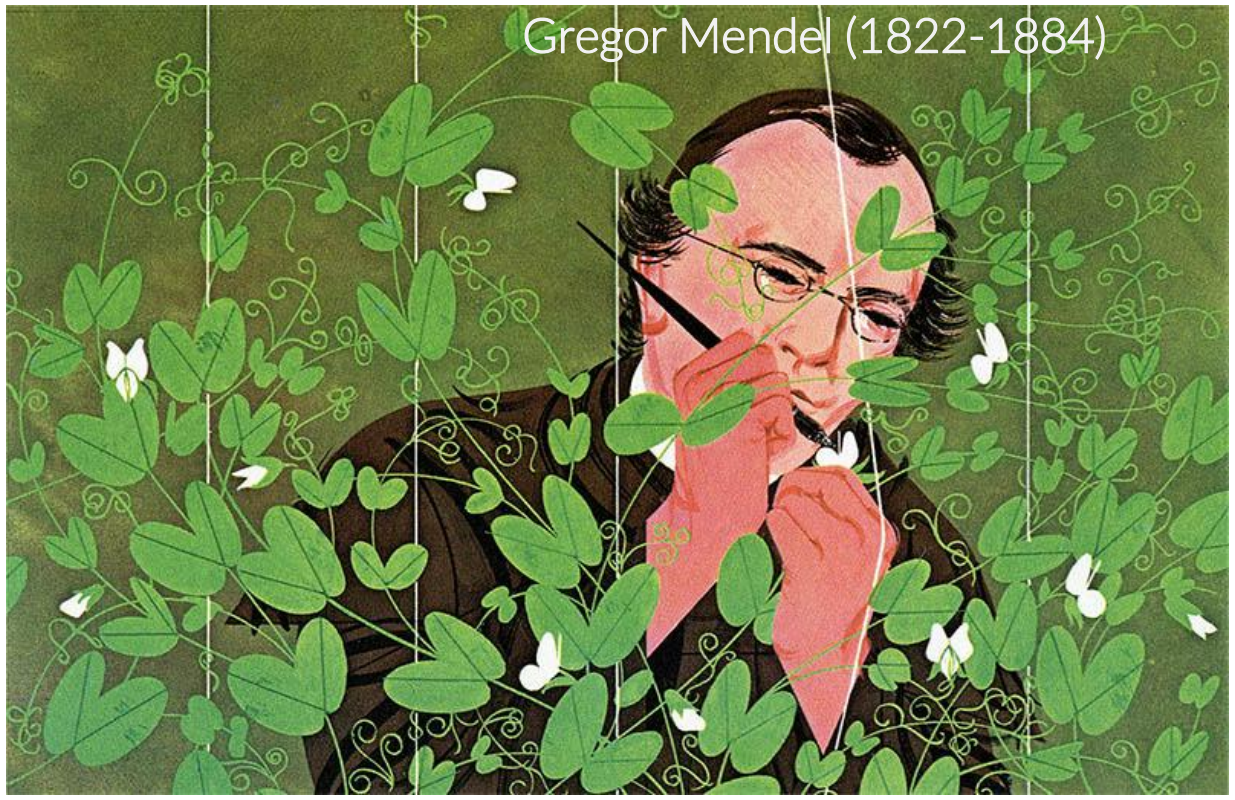
1) Segregation
2) Independent assortment
3) Dominance

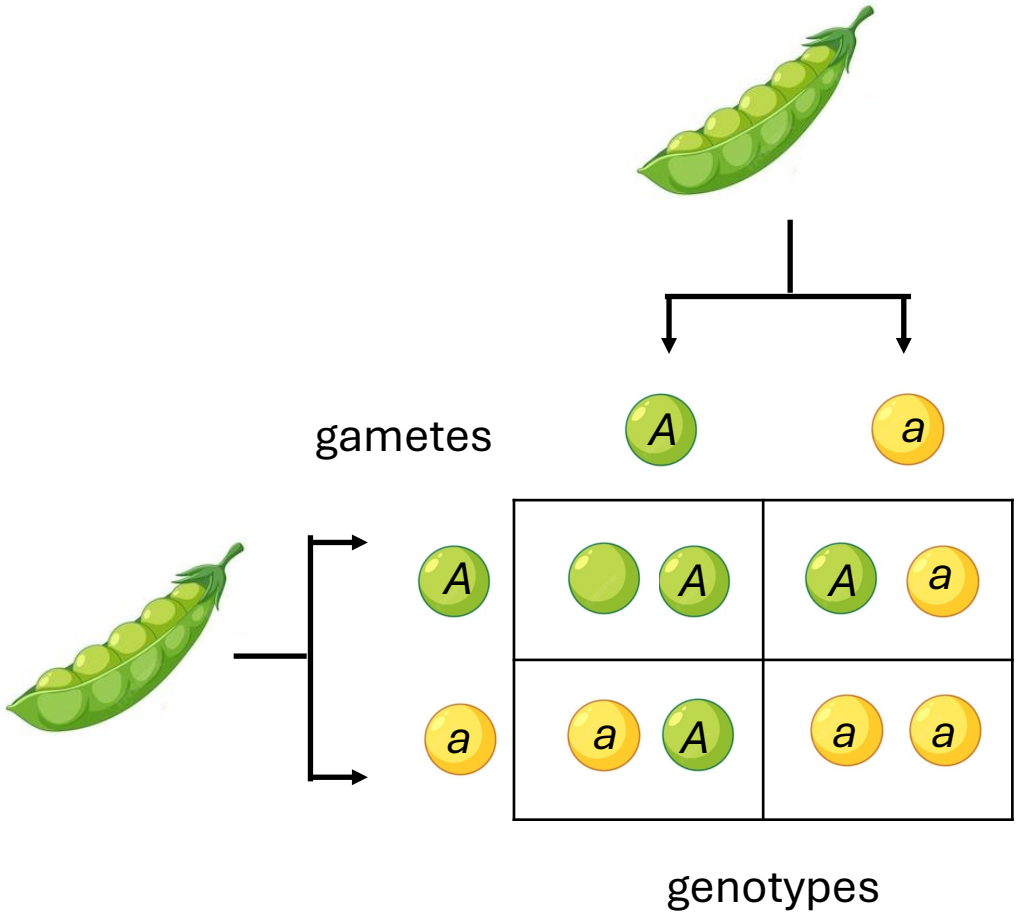




Mendel's laws of inheritance

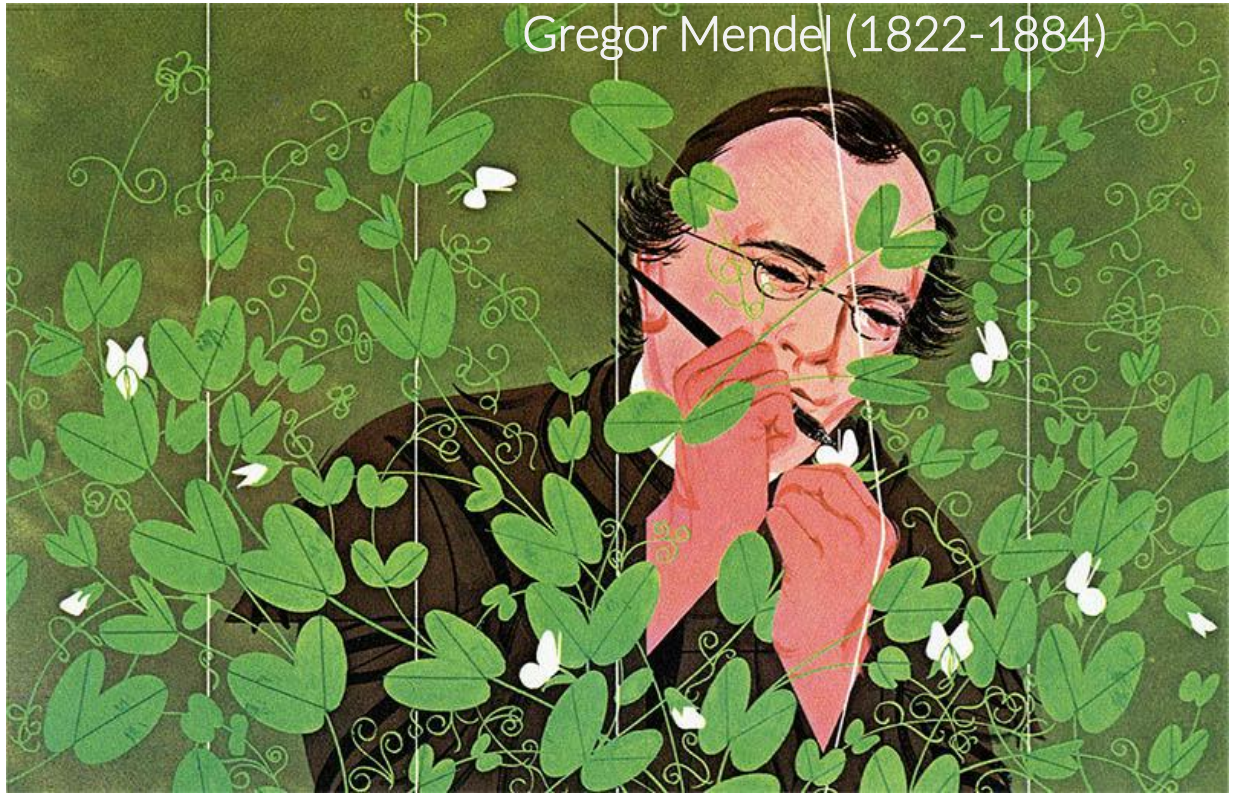
1) Segregation
2) Independent assortment
3) Dominance



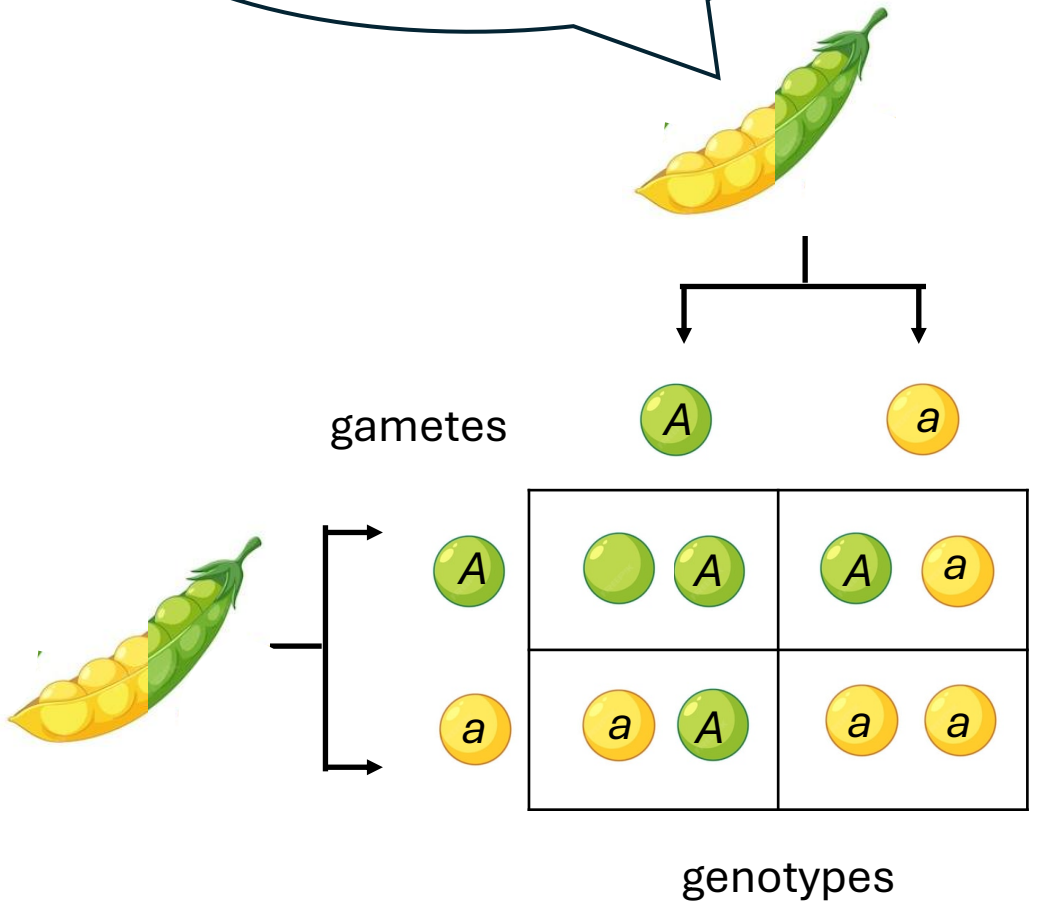


Mendel's laws of inheritance

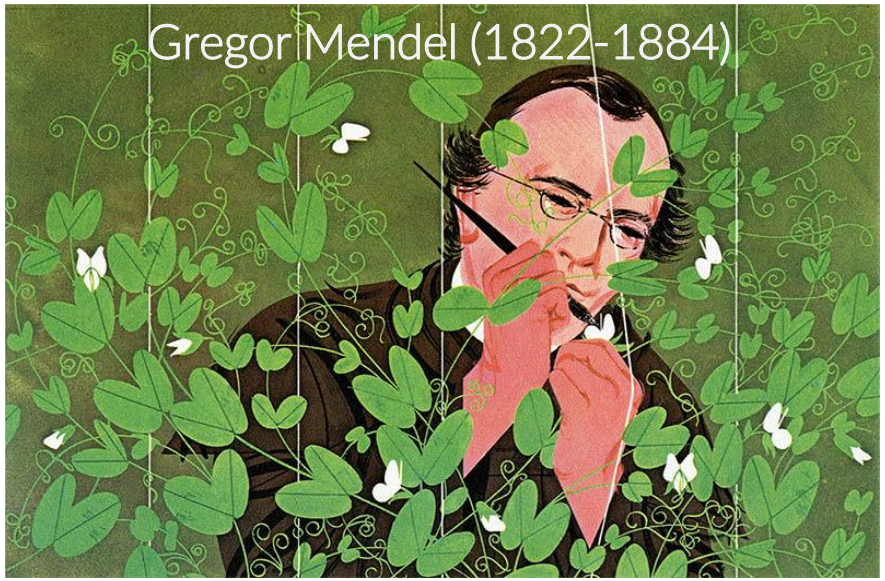
1) Segregation
2) Independent assortment
3) Dominance

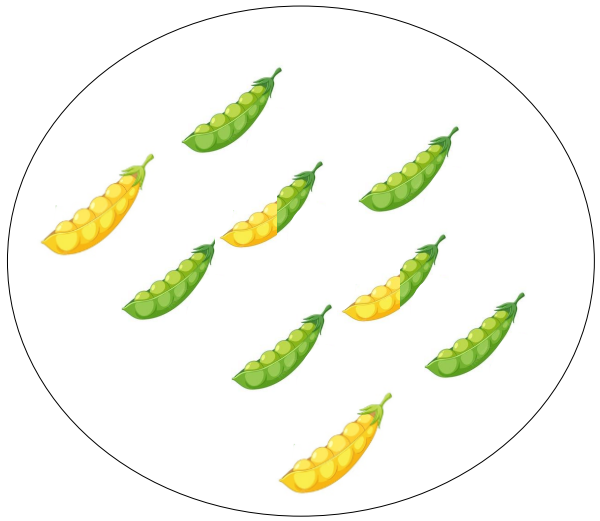


Green is actually dominant in peas but let's visualize them better as **codominant**

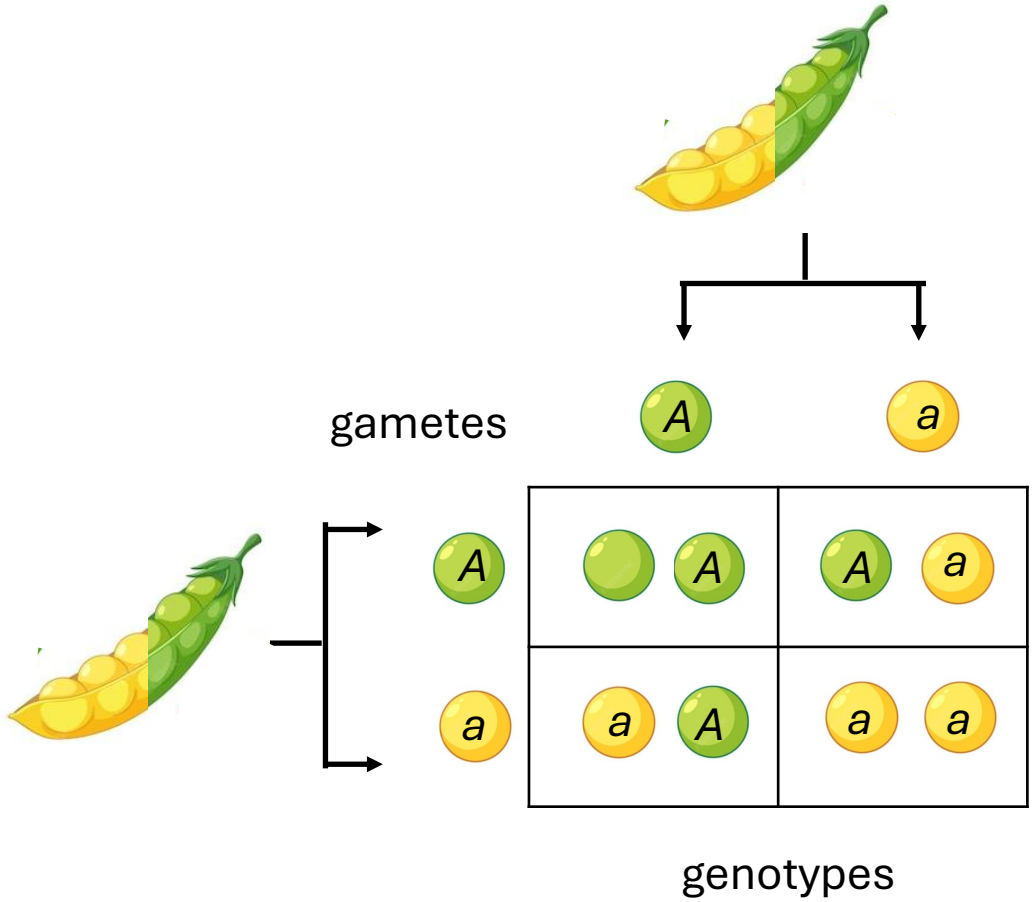


genotypes

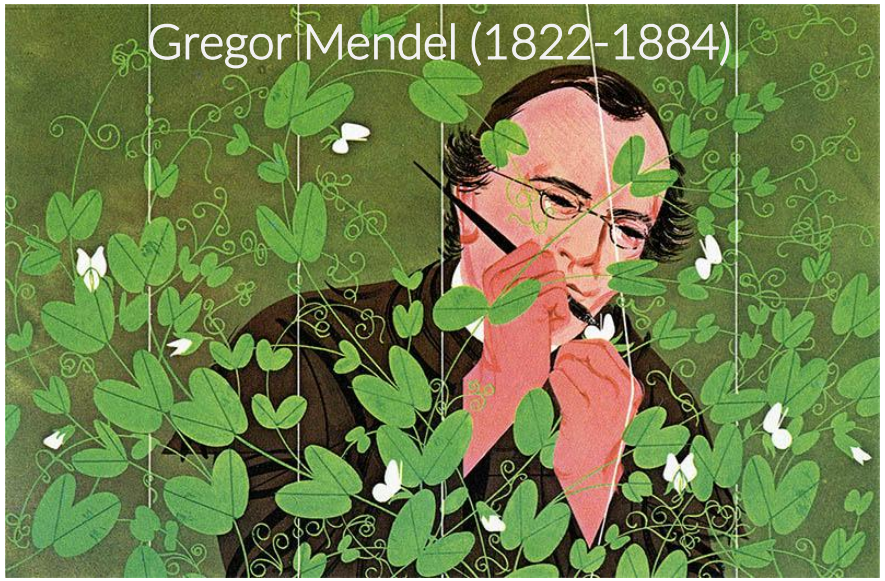


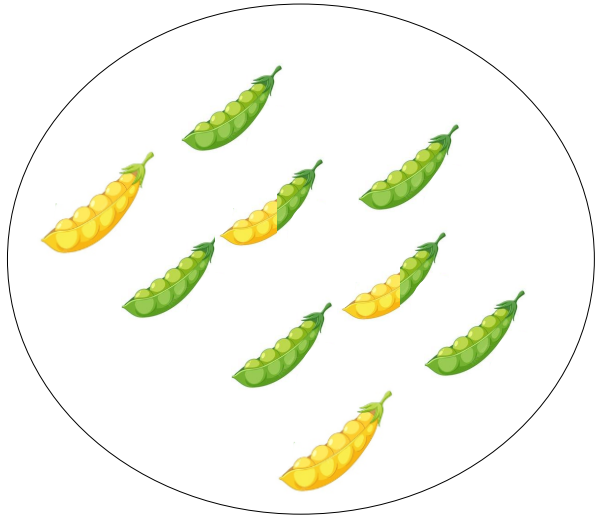


Genotypes



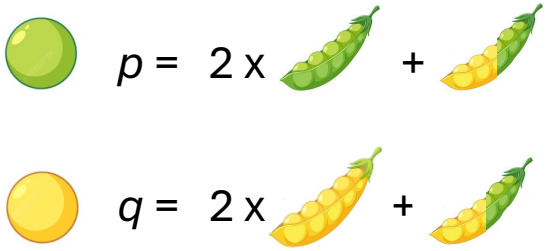
genotypes

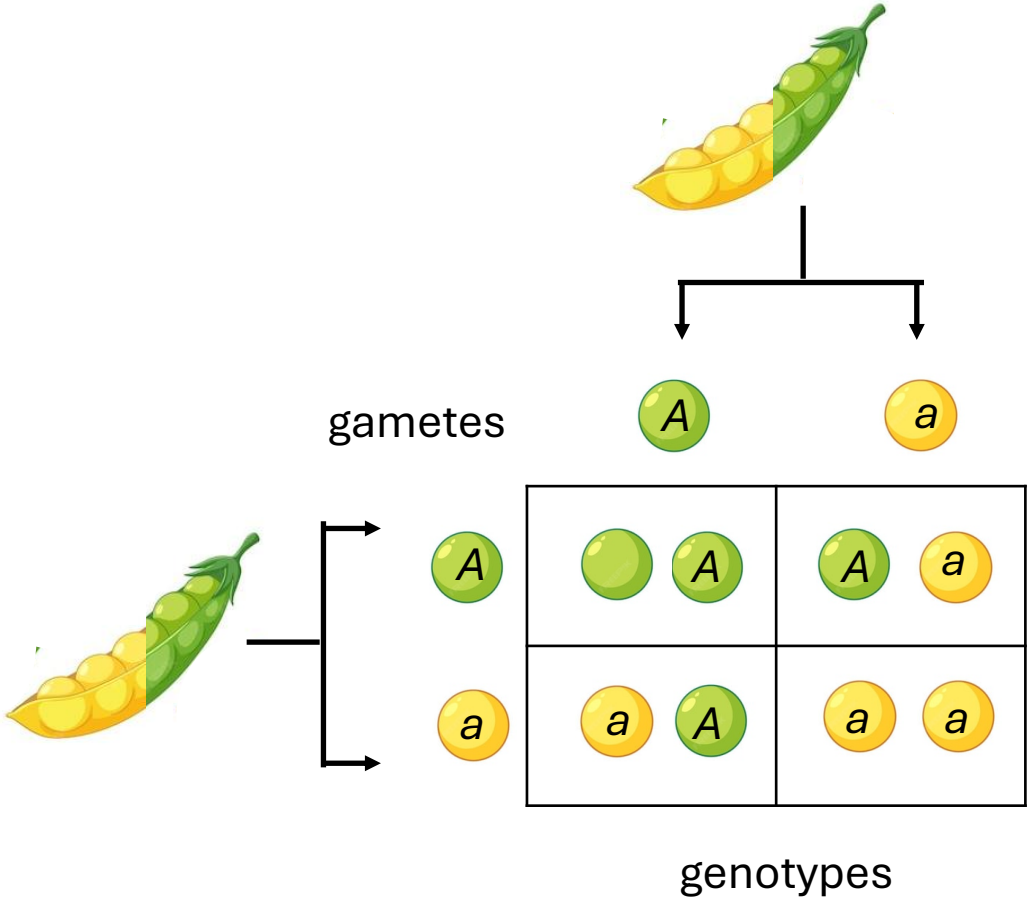




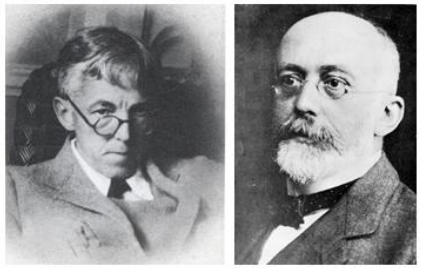
Genotypes

Frequency of gametes/alleles

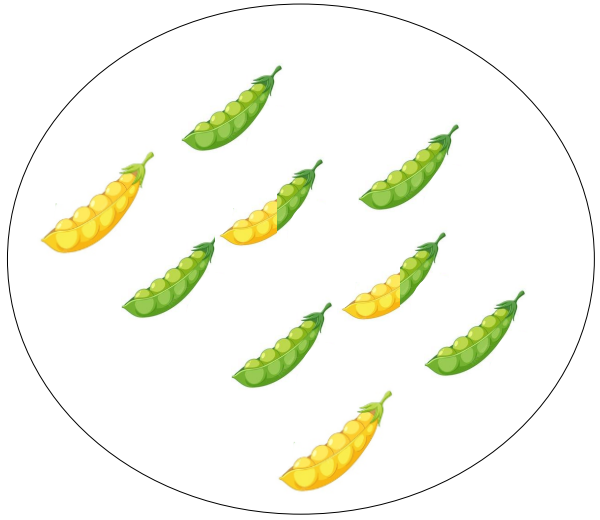




Hardy-Weinberg Equilibrium

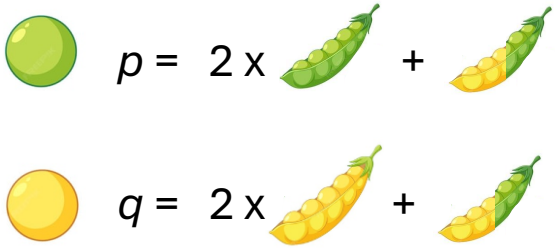


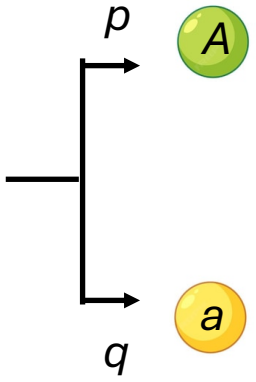
For a randomly mating populations, where the transmission of gametes of an allele with frequency *p* is not affected by external forces (mutation, selection, migration, etc.), the expected proportion of genotypes in next generations is distributed as p^2, $2pq$, q^2

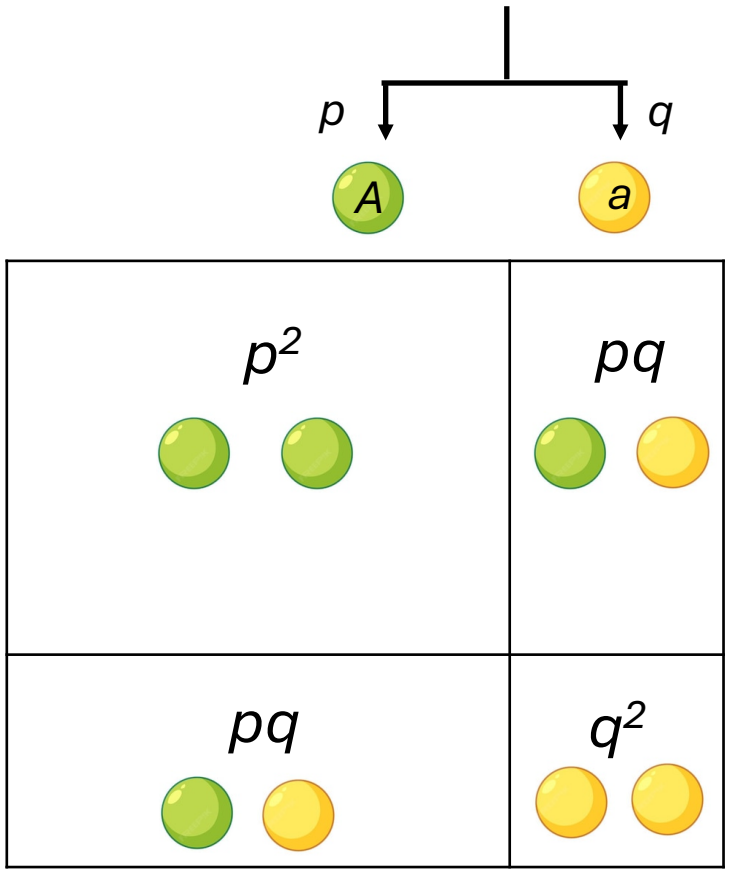


Genotypes

Frequency of gametes/alleles

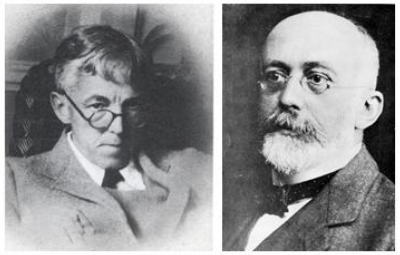






Frequency of (future) genotypes

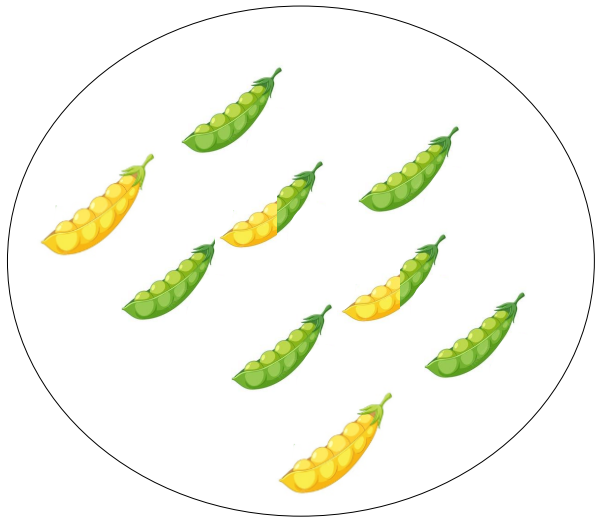
Hardy-Weinberg Equilibrium



=

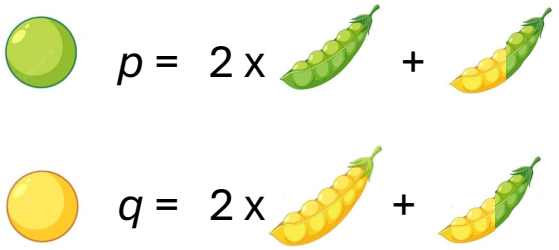
in a population

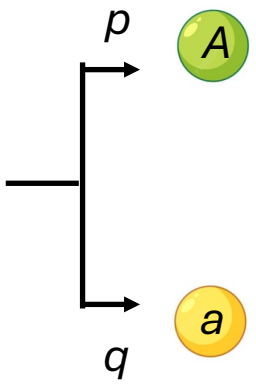
For a randomly mating populations, where the transmission of gametes of an allele with frequency *p* is not affected by external forces (mutation, selection, migration, etc.), the expected proportion of genotypes in next generations is distributed as p^2, $2pq$, q^2

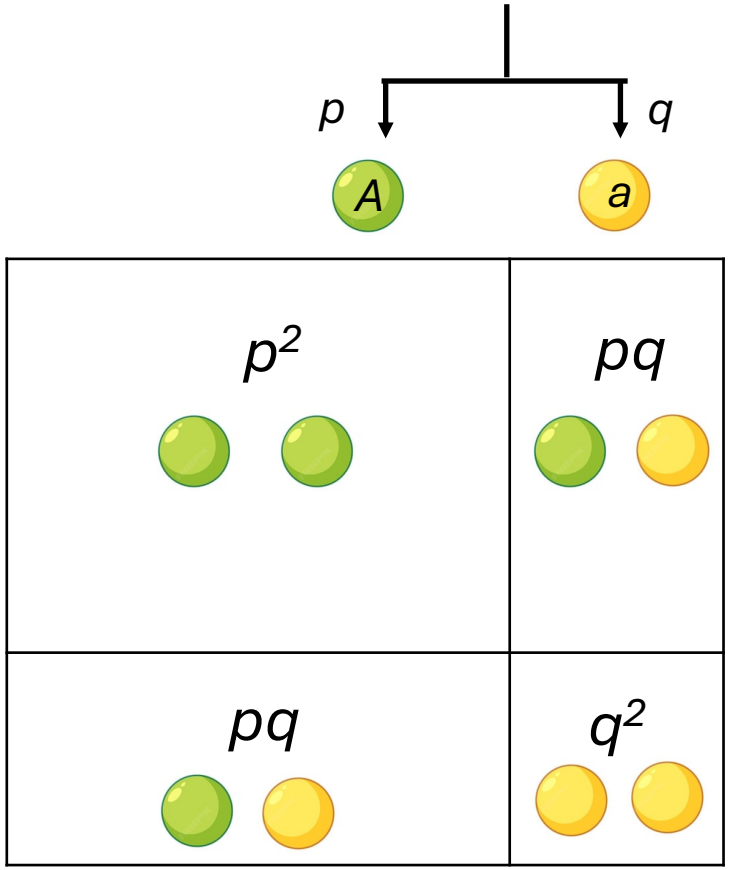


Genotypes

Frequency of gametes/alleles



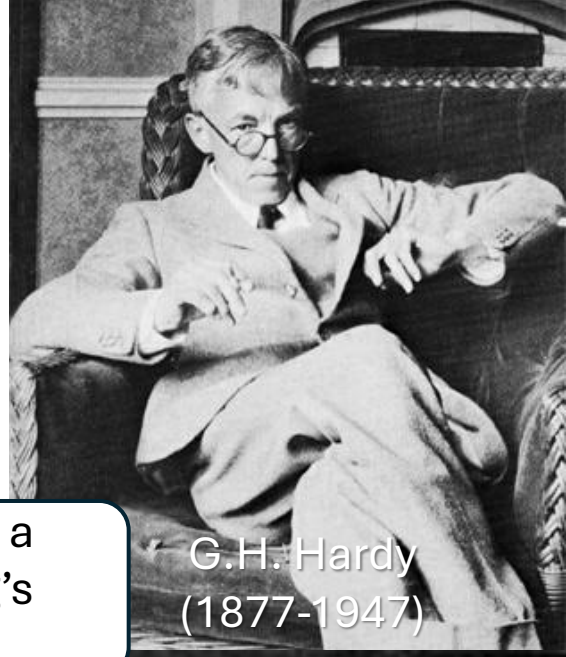




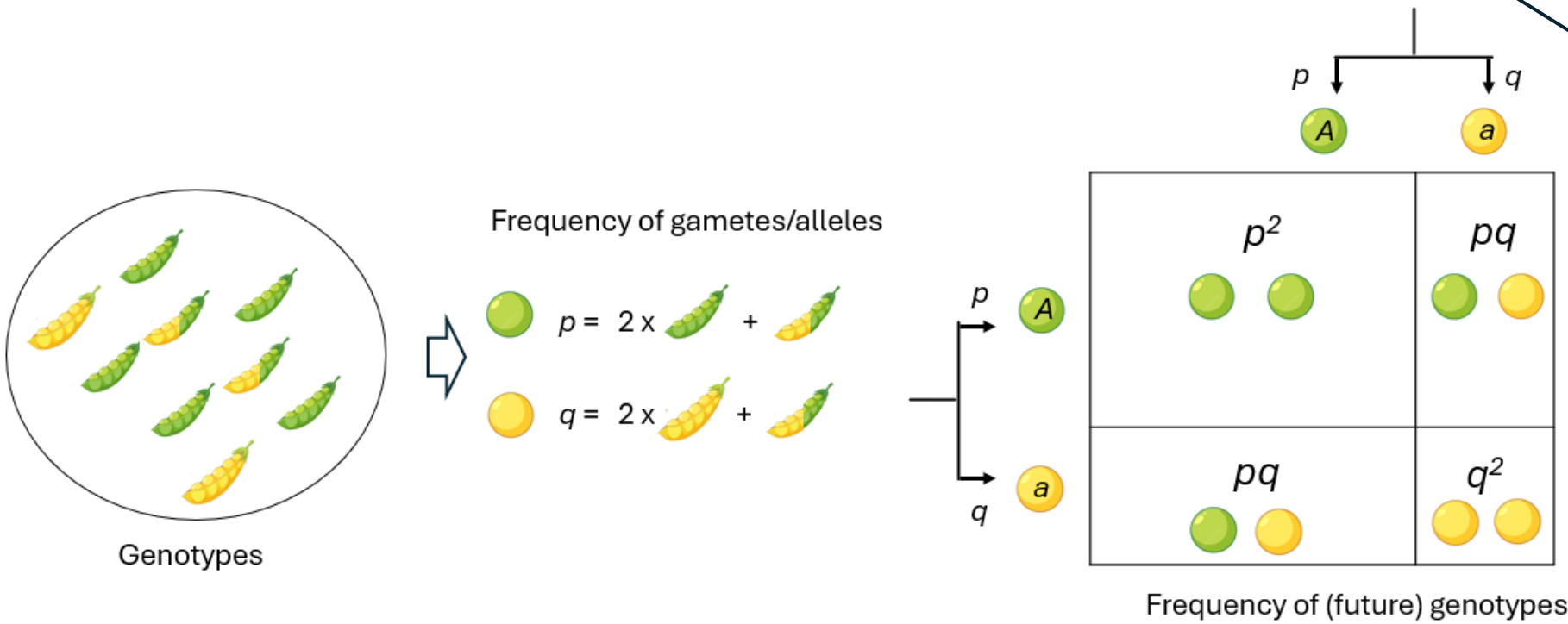
Frequency of (future) genotypes

Hardy-Weinberg Equilibrium

For a randomly mating populations, where the transmission of gametes of an allele with frequency *p* is not affected by external forces (mutation, selection, migration, etc.), the expected proportion of genotypes in next generations is distributed as p^2, $2pq$, q^2



Can you give us an example of a population at Hardy-Weinberg's equilibrium?

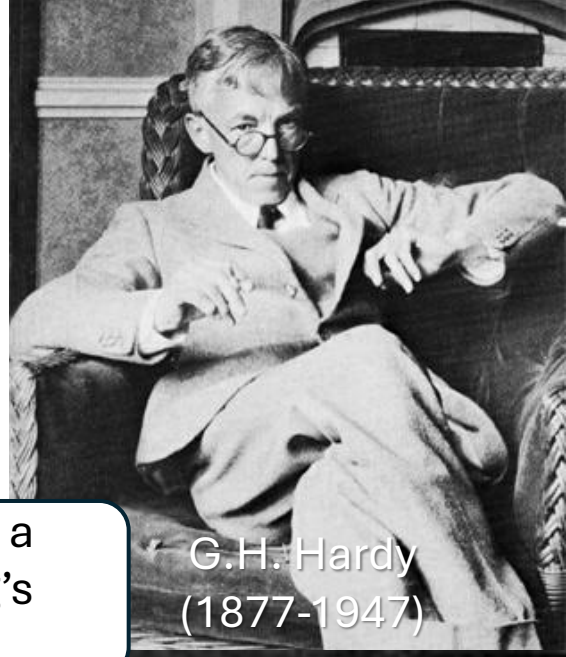


Can you give me an example of a population under Hardy-Weinberg equilibrium?

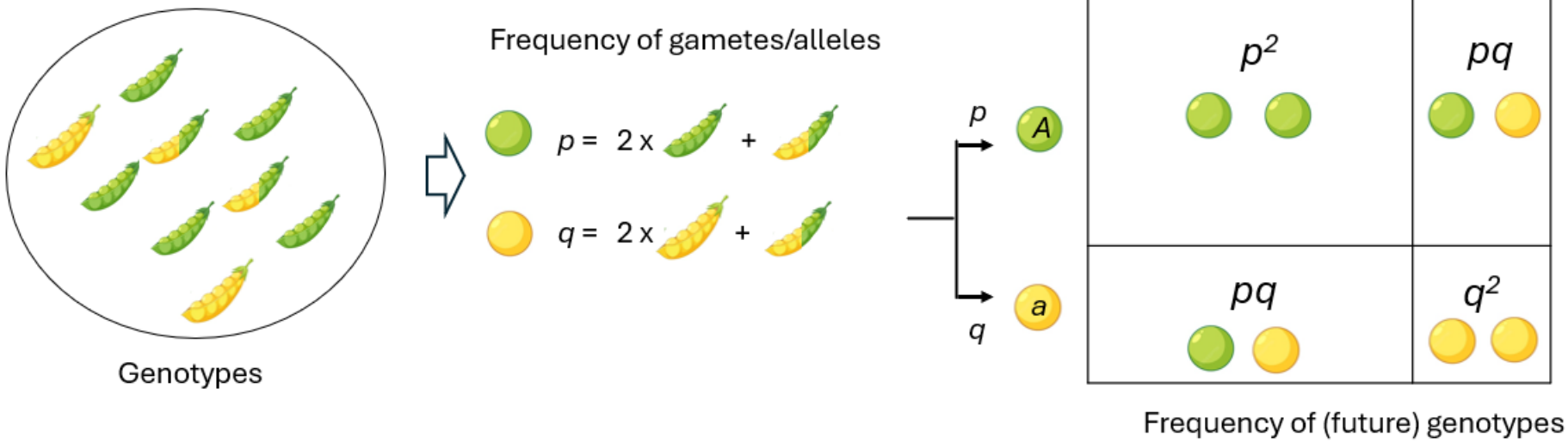
Can you give me an example of a population under Hardy-Weinberg equilibrium (more or less)?

Hardy-Weinberg Equilibrium

For a randomly mating populations, where the transmission of gametes of an allele with frequency *p* is not affected by external forces (**mutation**, selection, migration, etc.), the expected proportion of genotypes in next generations is distributed as p^2, $2pq$, q^2

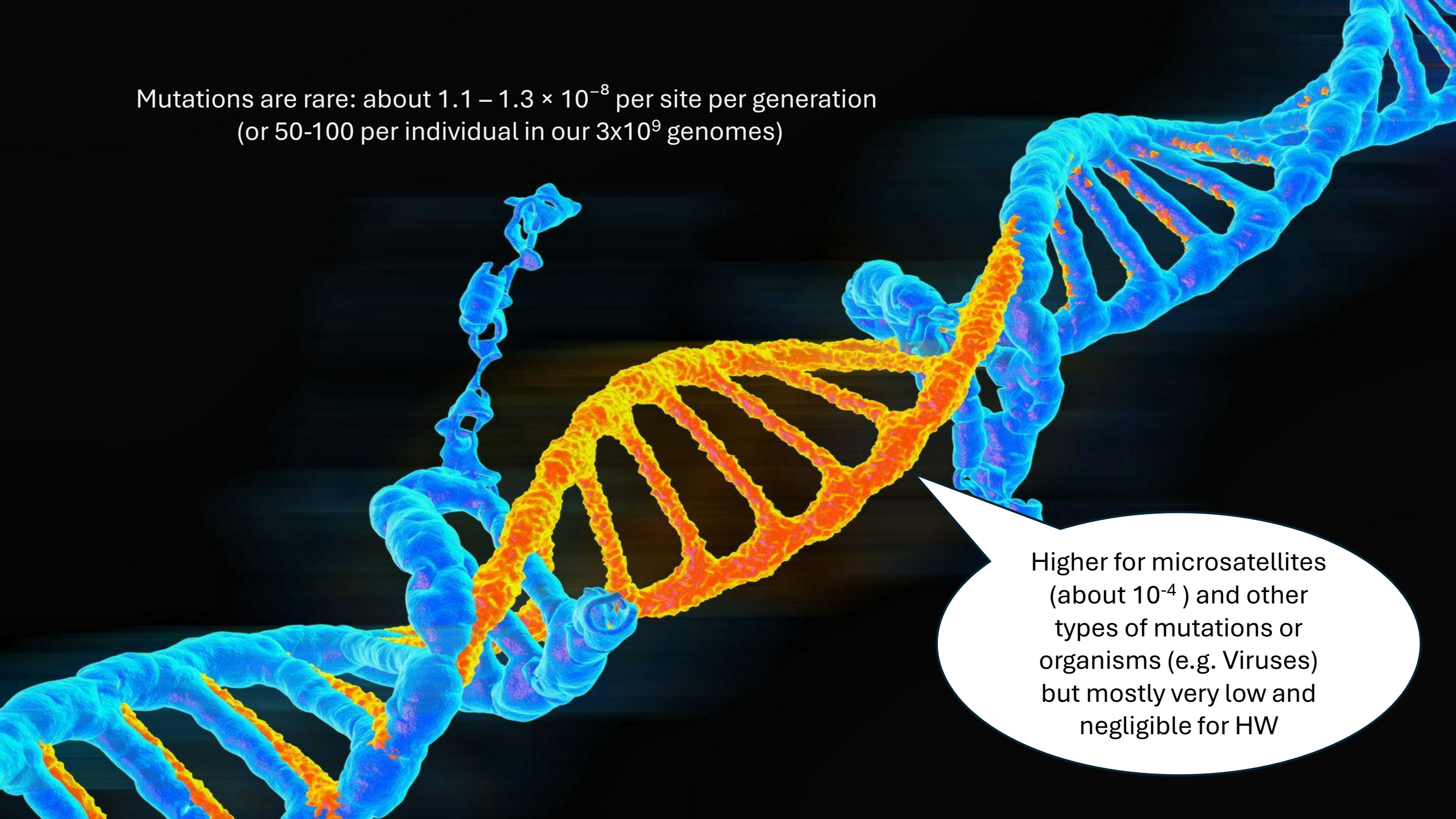


Can you give us an example of a population at Hardy-Weinberg's equilibrium?



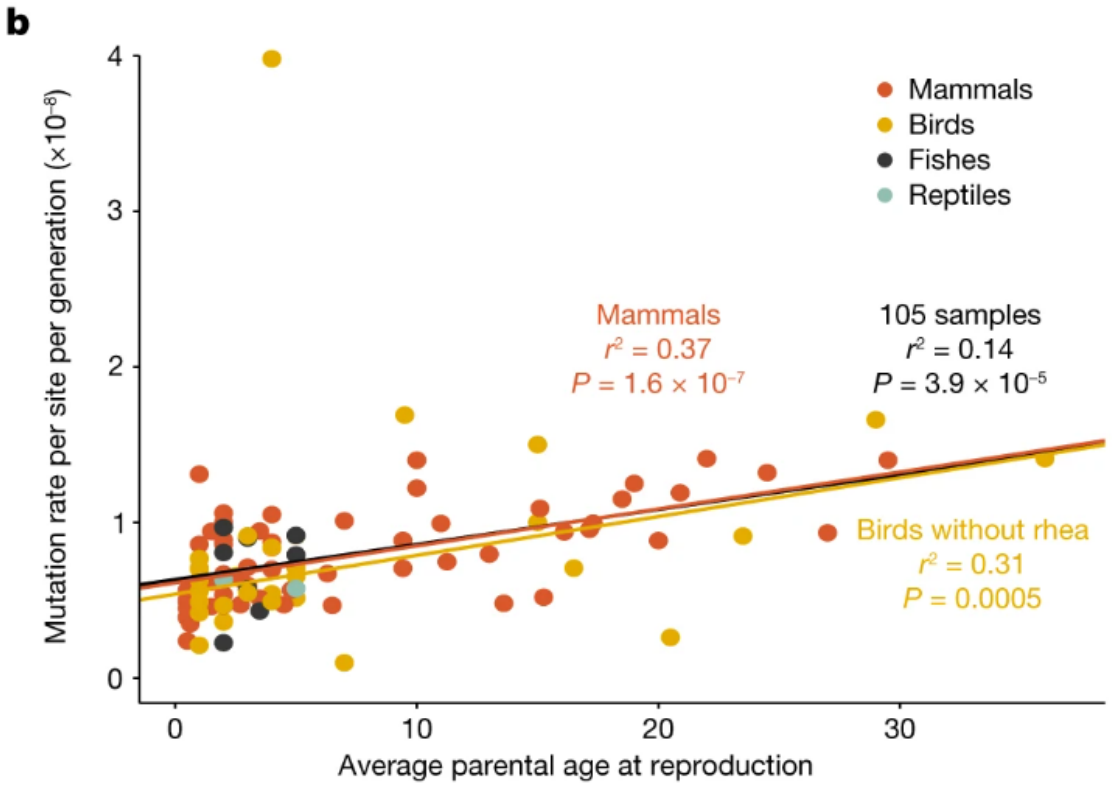
Mutations are rare: about $1.1 – 1.3 \times 10^{-8}$ per site per generation
(or 50-100 per individual in our $3x10^9$ genomes)

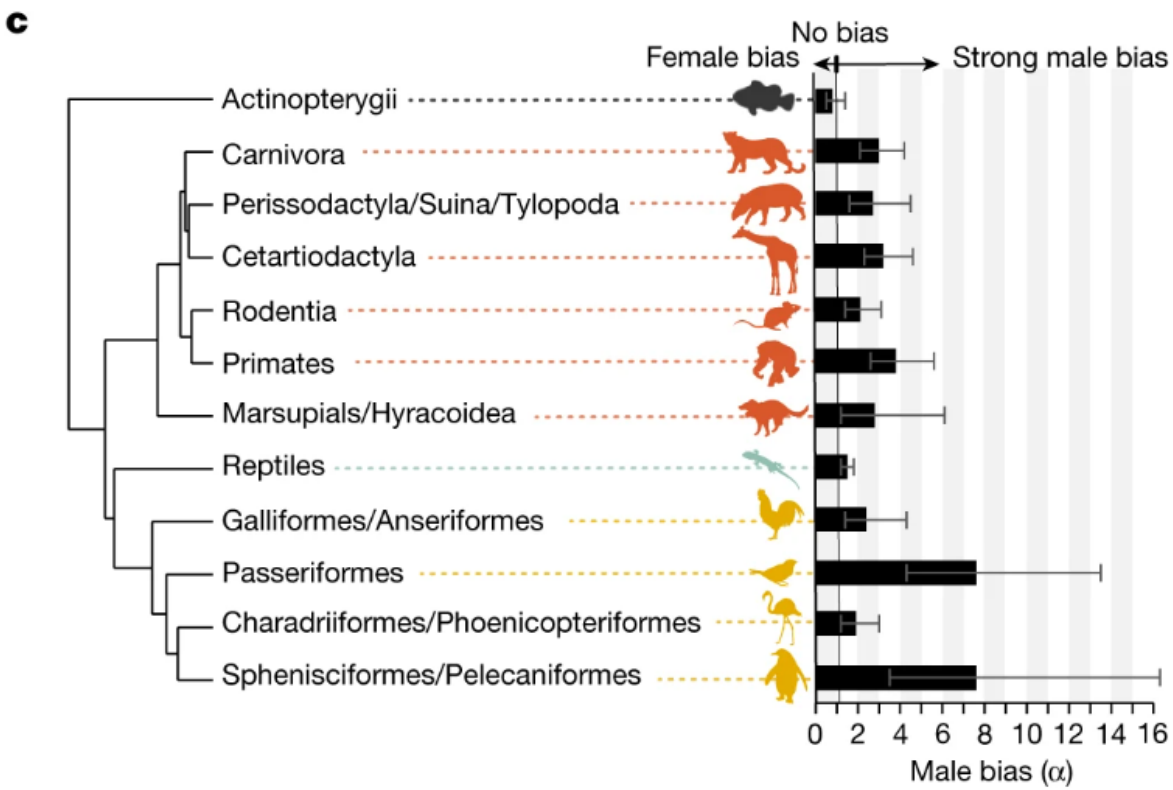
Mutations are rare: about $1.1 – 1.3 \times 10^{-8}$ per site per generation
(or 50-100 per individual in our $3x10^9$ genomes)



Higher for microsatellites (about 10^{-4}) and other types of mutations or organisms (e.g. Viruses) but mostly very low and negligible for HW

And mutations rates are as low in most vertebrates (and even lower in bacteria)

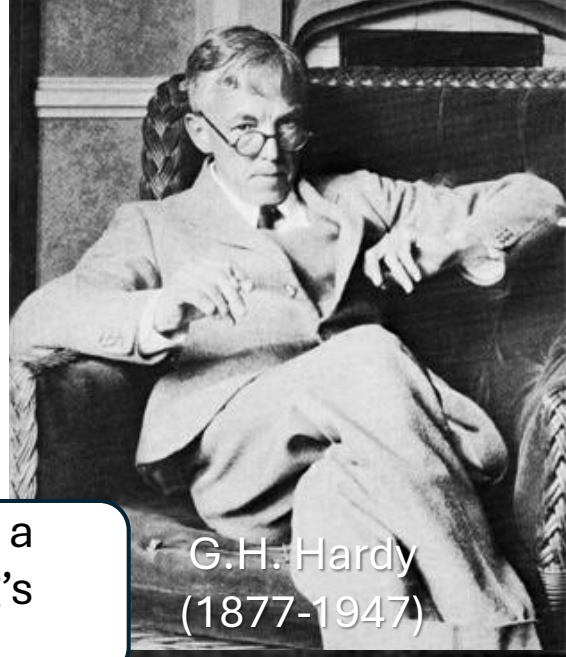




Bergeron 2023 et al., Nature

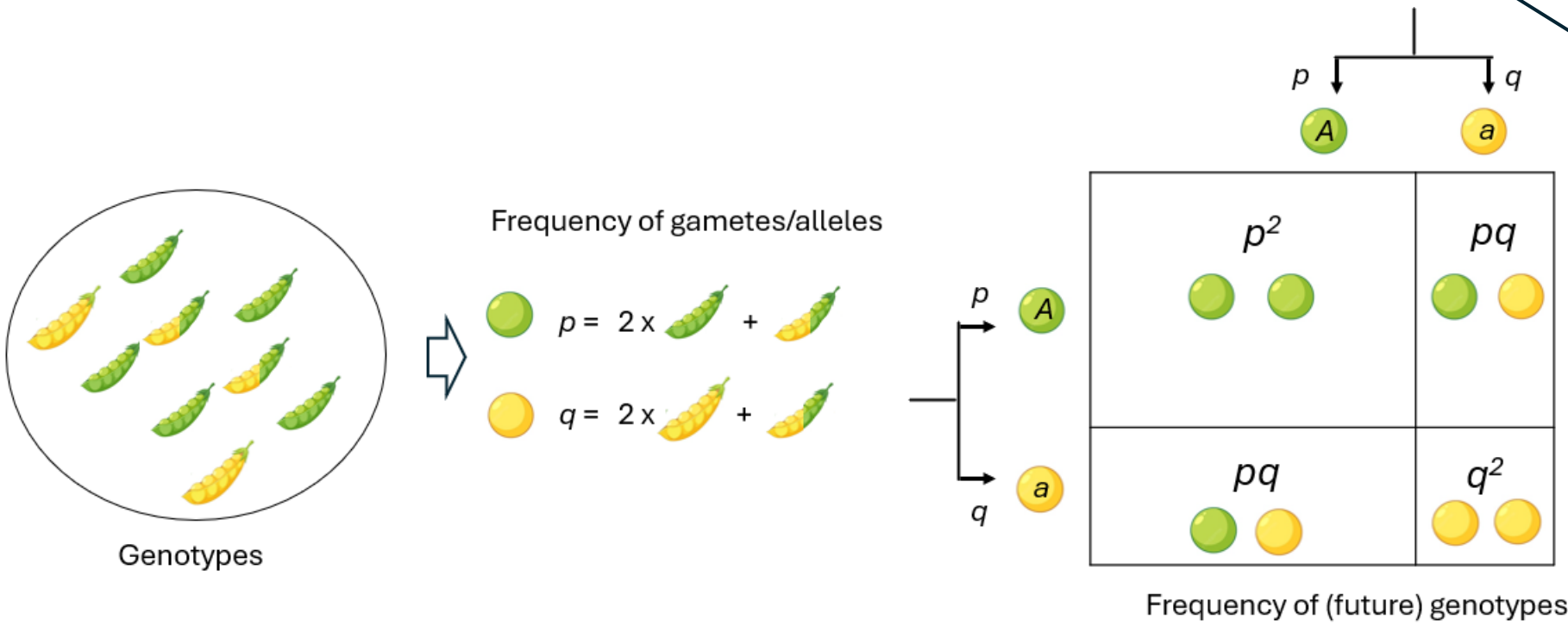
Hardy-Weinberg Equilibrium

For a randomly mating populations, where the transmission of gametes of an allele with frequency *p* is not affected by external forces (mutation, **selection**, migration, etc.), the expected proportion of genotypes in next generations is distributed as p^2, $2pq$, q^2


G.H. Hardy (1877-1947)

Can you give us an example of a population at Hardy-Weinberg's equilibrium?


Wilhelm Weinberg (1862-1937)



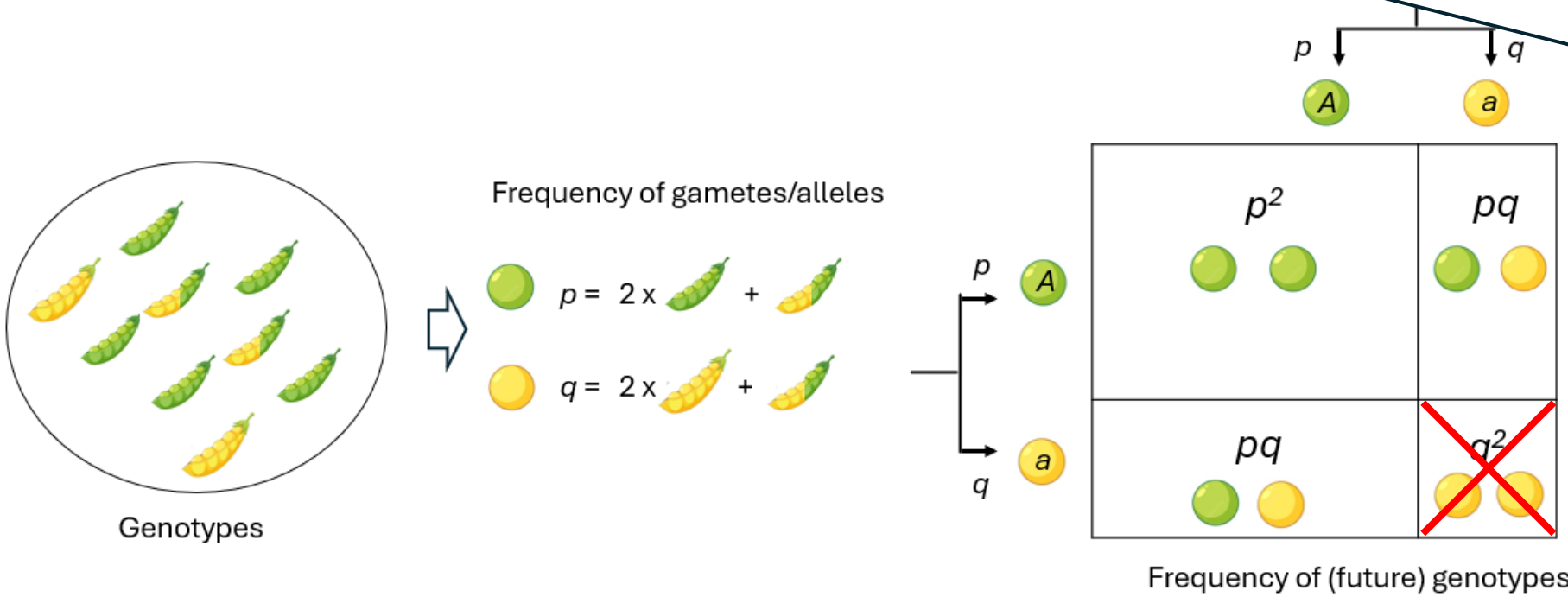
Hardy-Weinberg Equilibrium

For a randomly mating populations, where the transmission of gametes of an allele with frequency *p* is not affected by external forces (mutation, **selection**, migration, etc.), the expected proportion of genotypes in next generations is distributed as p^2, $2pq$, q^2

True in principle that it affects HWW. But selection is usually quite weak (selection coefficients about 0.001-0.01) so it mostly leads to transients effects on allele frequencies, not so much on genotype frequencies.

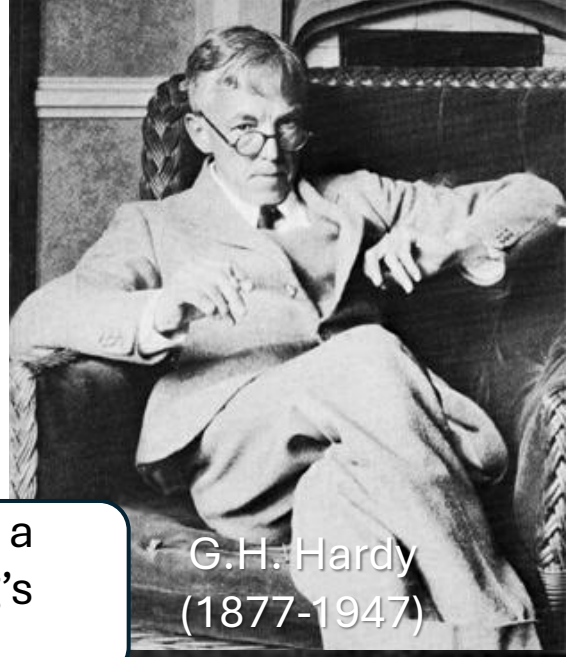
G.H. Hardy (1877-1947)

Wilhelm Weinberg (1862-1937)



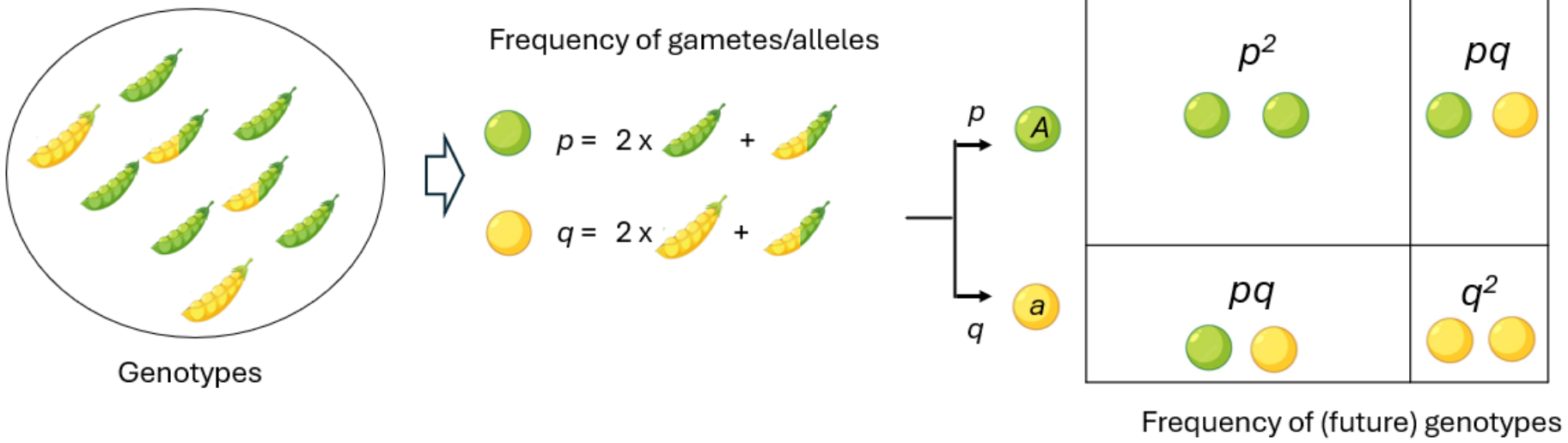
Hardy-Weinberg Equilibrium

For a randomly mating populations, where the transmission of gametes of an allele with frequency *p* is not affected by external forces (mutation, selection, **migration**, etc.), the expected proportion of genotypes in next generations is distributed as p^2, $2pq$, q^2

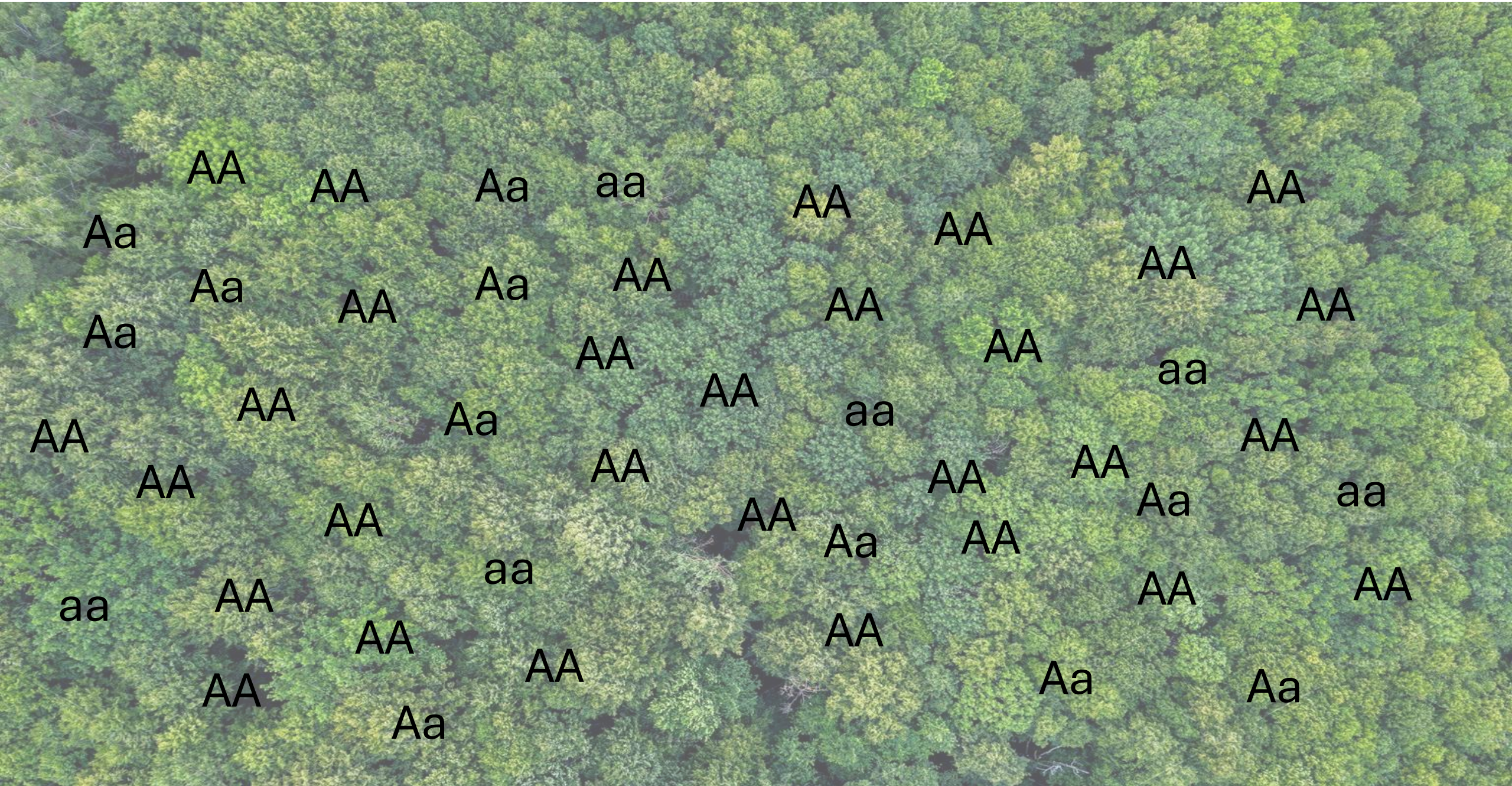

G.H. Hardy (1877-1947)

Can you give us an example of a population at Hardy-Weinberg's equilibrium?

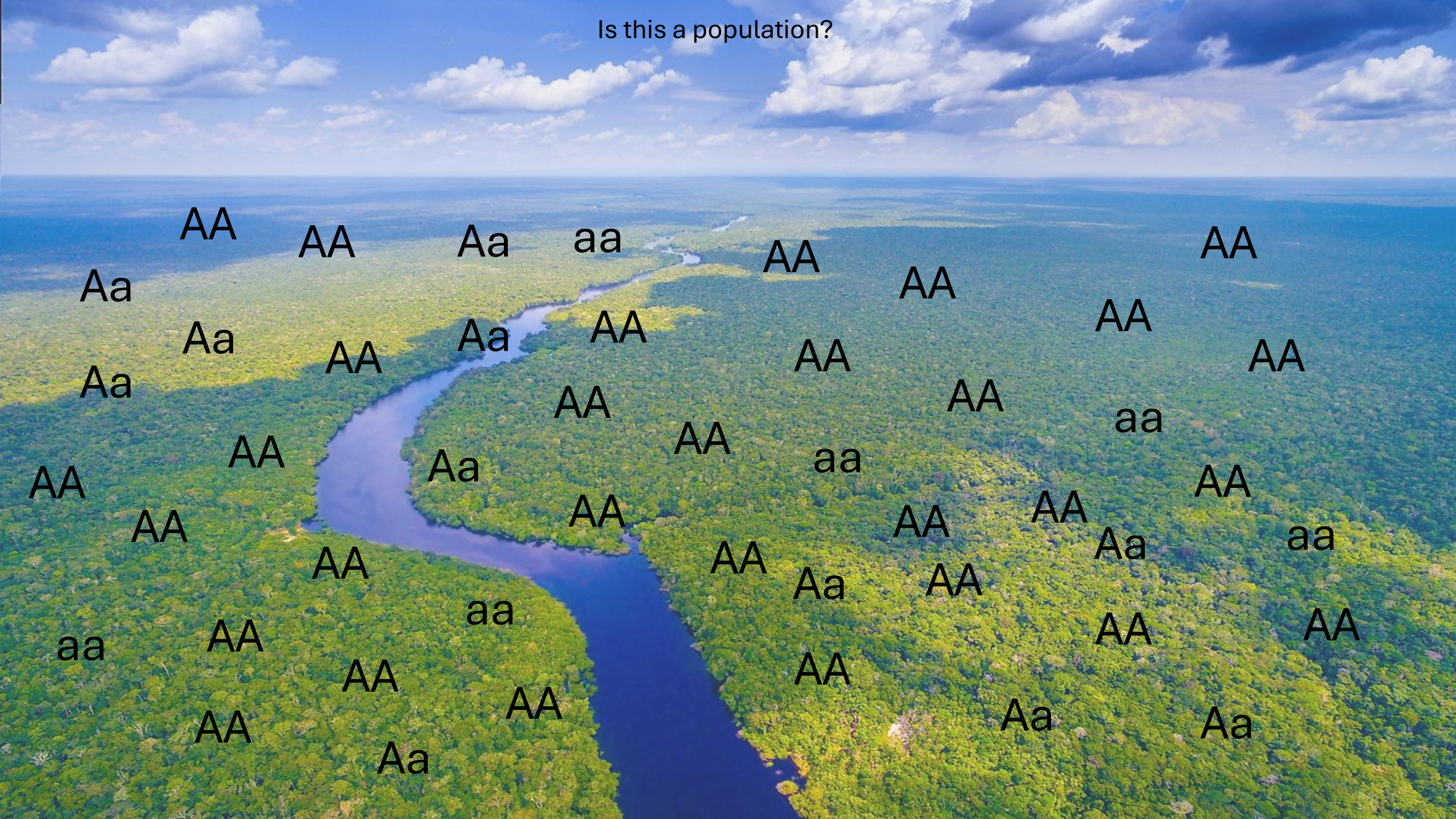

Wilhelm Weinberg (1862-1937)

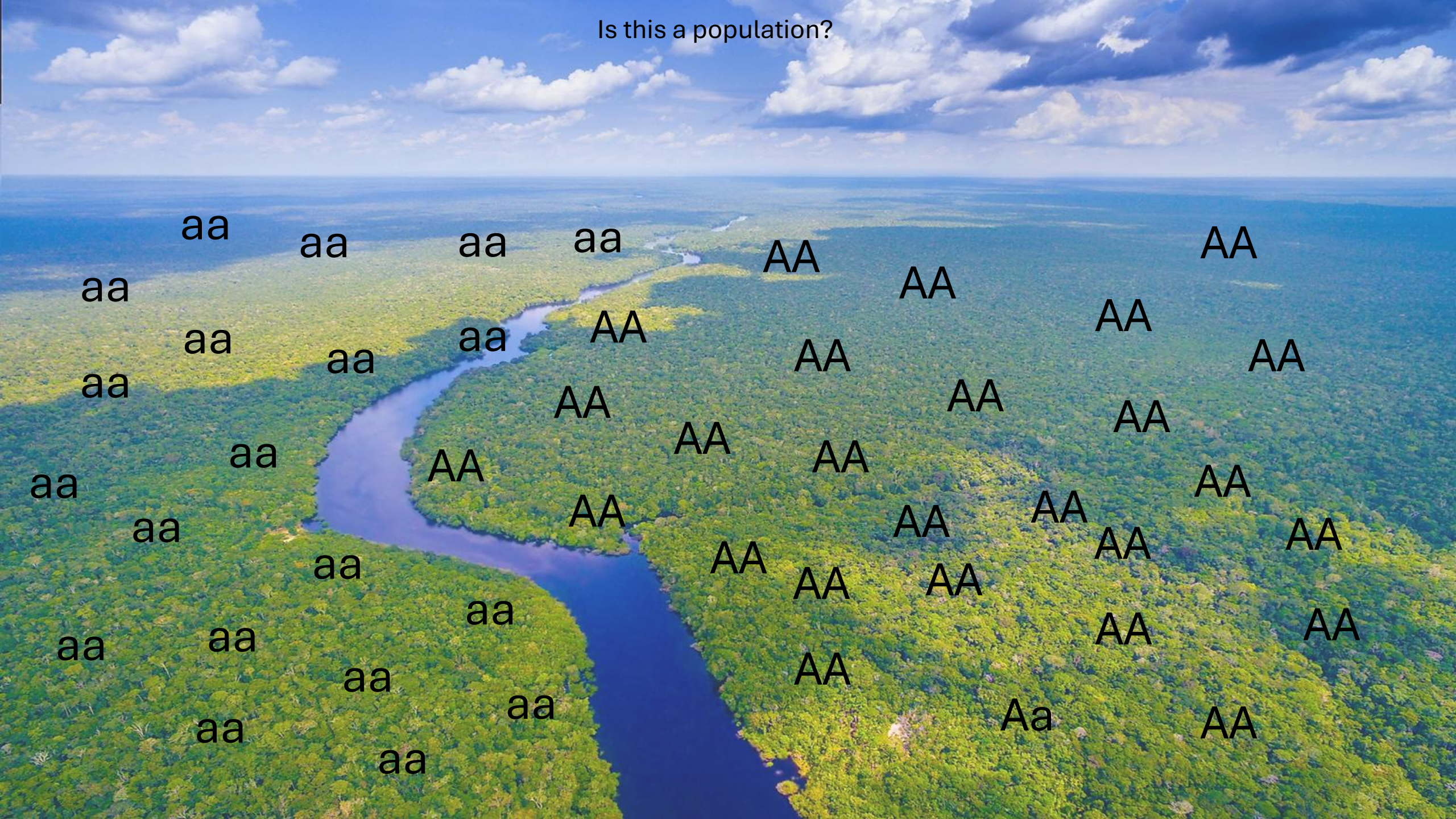


Is this a population?

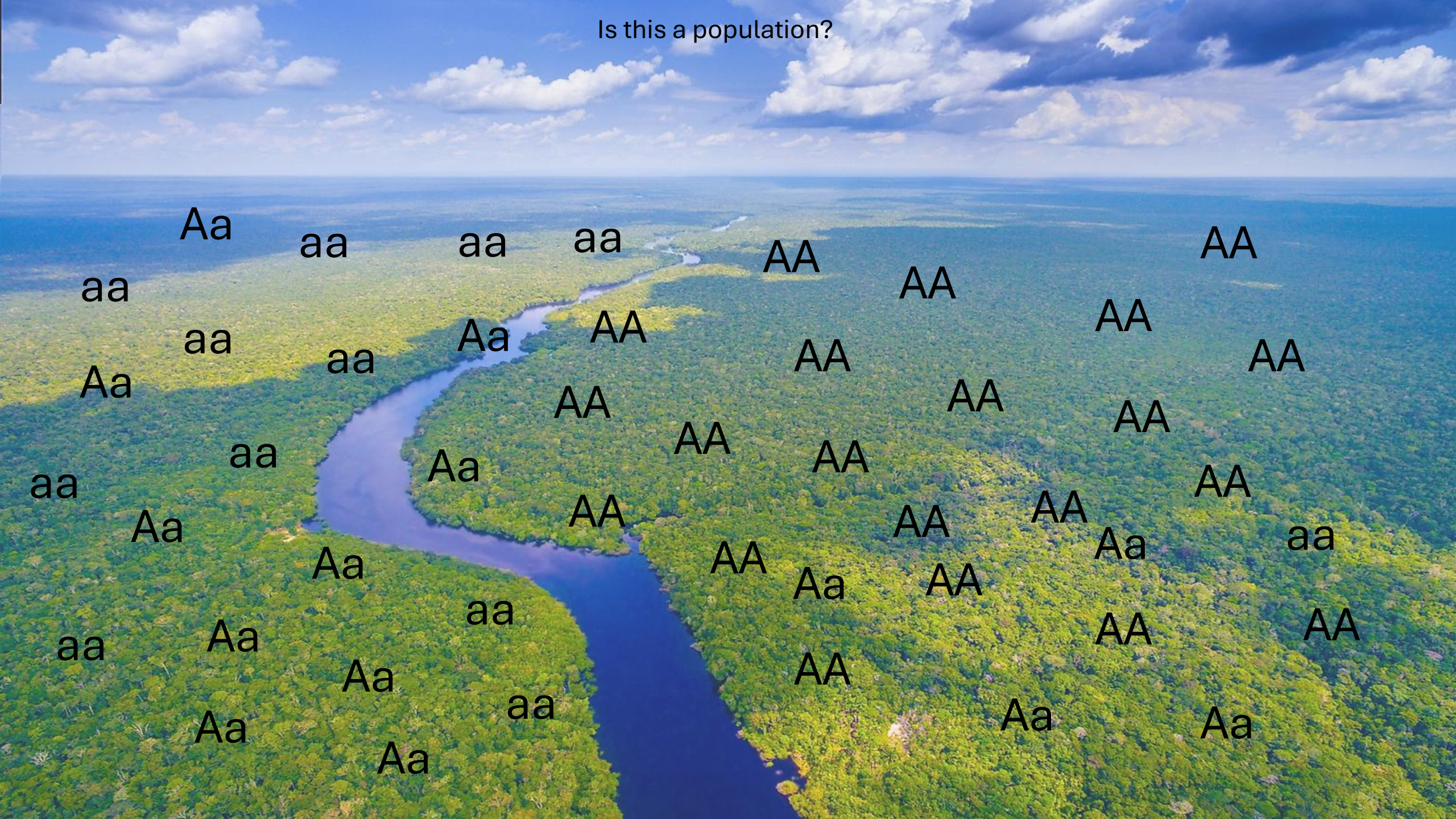


Is this a population?



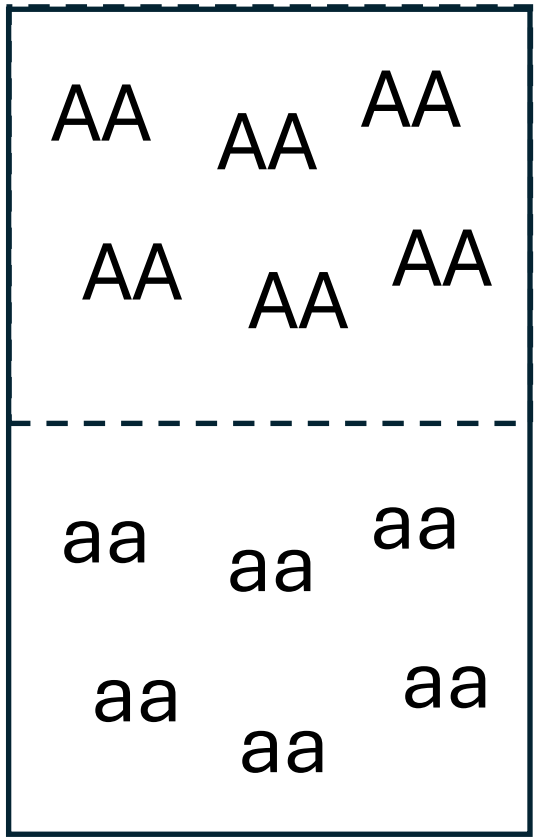

Is this a population?
aa
aa
aa
aa
aa
aa
aa
aa
aa
aa
aa
aa
aa
aa
aa
aa
aa
aa
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
AA
Aa

Is this a population?



Principio di Wahlund

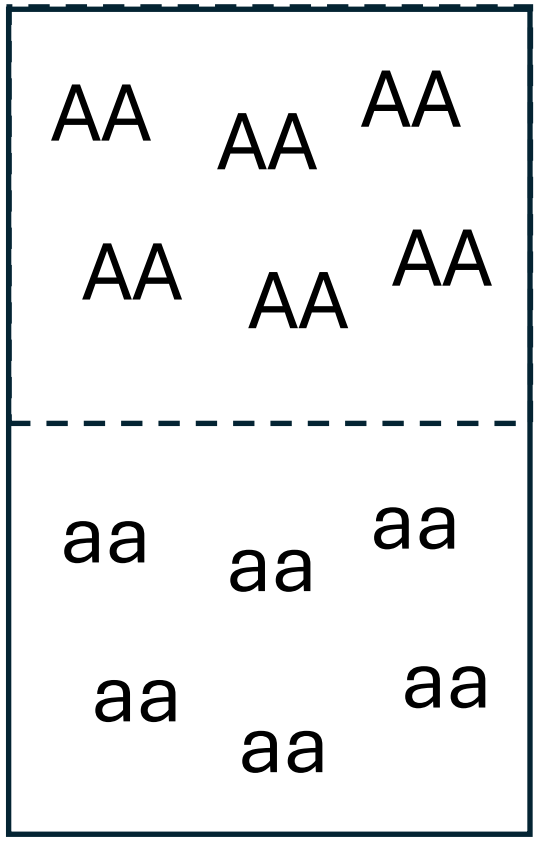
Popolazioni suddivise mostrano un eccesso di omozigoti/mancanza di eterozigoti rispetto a popolazioni non suddivise (in HWE).



H: eterozigosita’ (proporzione di genotipi eterozigosi)
Ho: eterozigosita’ osservata: 0% nell’esempio
He: eterozigosita’ attesa: 2x0.5x0.5=50% nell’esempio

Principio di Wahlund

Popolazioni suddivise mostrano un eccesso di omozigoti/mancanza di eterozigoti rispetto a popolazioni non suddivise (in HWE).



H: eterozigosita' (proporzione di genotipi eterozigosi)
Ho: eterozigosita' osservata: 0% nell'esempio
He: eterozigosita' attesa: 2x0.5x0.5=50% nell'esempio

Questa osservazione e' alla base di una delle piu importanti statistiche nella genetica di popolazione, l'**Fst.**

Testing for HWE

$(p+q)^2 =$

p^2 +	$2pq$	$+q^2$	Genotypes predicted by HWE
AA	Aa	aa	Observed genotypes

Contingency table

The equilibrium of Hardy-Weinberg gives us a theoretical prediction (null hypothesis) to test if the observed genotypes in a population emerge from random mating of alleles and are not influenced by other forces.

Fisher's exact test or χ2!

Box 3.2 Calculating Hardy–Weinberg equilibrium

Table 3.1 is an actual data set on scarlet tiger moths that was collected by the geneticist E. B. Ford.

The data in Table 3.1 tell us that in this sample there is a total of 2(1612) = 3224 alleles at this particular locus. Of these, 3076 are *A* alleles (2938 + 138), and 148 are *a* alleles (138 + 10). Therefore, the frequency *p* of the *A* allele in this population is:

$$p = \frac{3076}{3224} = 0.954$$

and the frequency *q* of the *a* allele can be calculated as either:

$$q = \frac{148}{3224} = 0.046$$

or, because $p + q = 1$, can alternatively be calculated as:

$$q = 1 - p = 1 - 0.954 = 0.046$$

If we know *p* and *q*, then we can calculate the frequencies of *AA* ($= p^2$), *Aa* ($= 2pq$) and *aa* ($= q^2$) that would be expected if the population is in HWE as follows:

$$p^2 = (0.954)^2 = 0.9101$$
$$2pq = 2(0.954)(0.046) = 0.0878$$
$$q^2 = (0.046)^2 = 0.002$$

We now need to calculate the number of moths in this population that would have each genotype if this population is in HWE. We can do this by multiplying the total number of moths (1612) by each genotype frequency:

$$AA = (0.9101)(1612) = 1467$$
$$Aa = (0.0878)(1612) = 142$$
$$aa = (0.002)(1612) = 3$$

Therefore the Hardy–Weinberg ratio expressed as the number of individuals with each genotype is 1467: 142: 3. This is very close to the actual ratio of genotypes within the population (from Table 3.1) of 1469 :138 : 5.

We can check whether or not there is a significant difference between the observed and expected genotype frequencies by using a chi-square (χ^2) test. This is based on the difference between the observed (O) number of genotypes, and the number that would be expected (E) under HWE, and is calculated as:

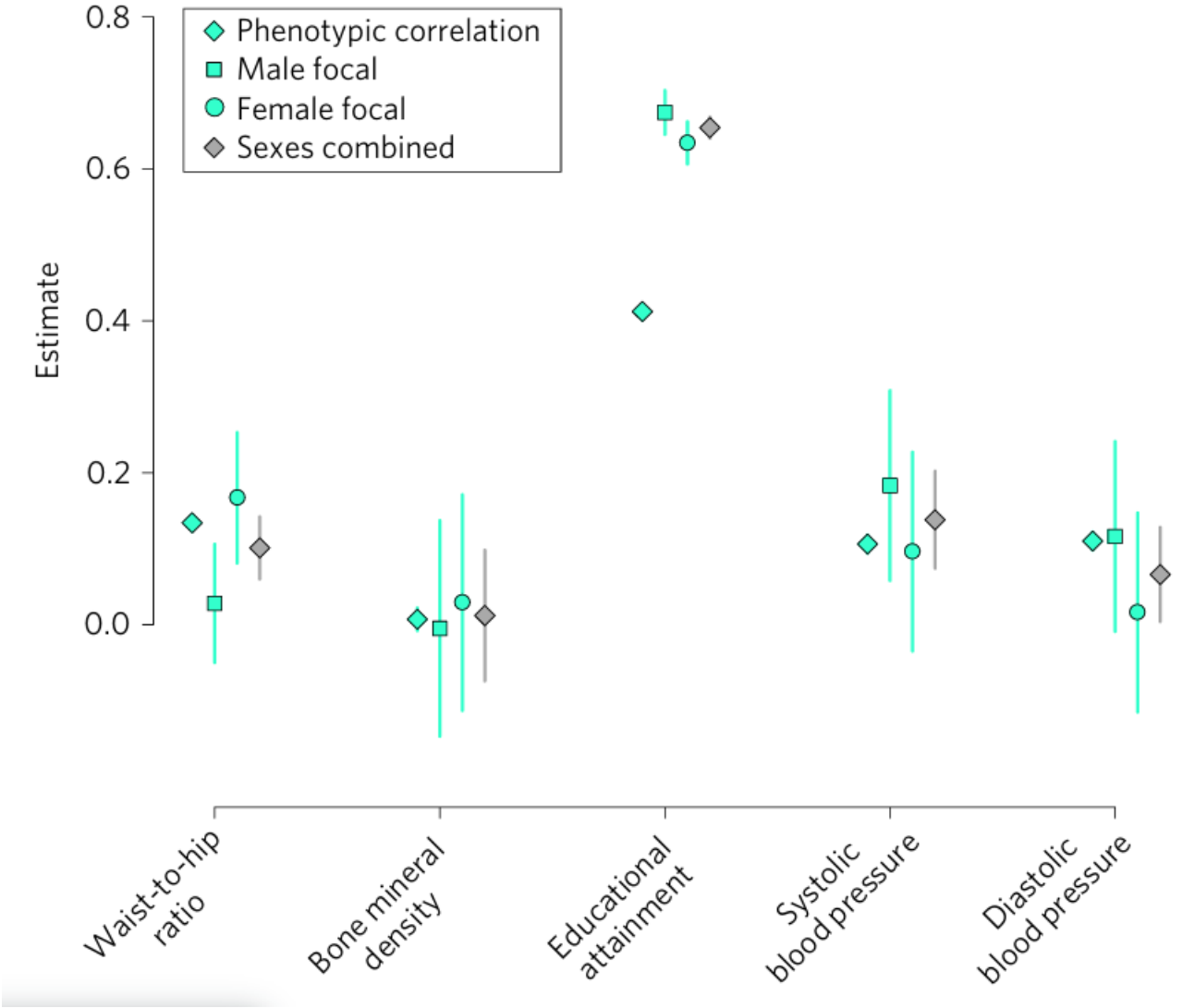
$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (3.5)$$

The χ^2 value of the scarlet tiger moth example is:

$$\chi^2 = \frac{(1469 - 1467)^2}{1467} + \frac{(138 - 142)^2}{142} + \frac{(5 - 3)^2}{3}$$
$$= 0.0027 + 0.11 + 1.33$$
$$= 1.44$$

The number of degrees of freedom (df) is determined as 3 (the number of genotypes) minus 1 (because if we know two of the expected genotype frequencies we automatically know the third) minus 1 (the number of independent values we calculated from our observed data to determine our expected values), which leaves df = 1. By using a statistical table, we learn that a χ^2 value of 1.44, in conjunction with one df, leaves us with a probability of $P = 0.230$. This means that there is no significant difference between the observed genotype frequencies in the scarlet tiger moth population and the genotype frequencies that are expected under HWE. We would therefore conclude that this population is in HWE.

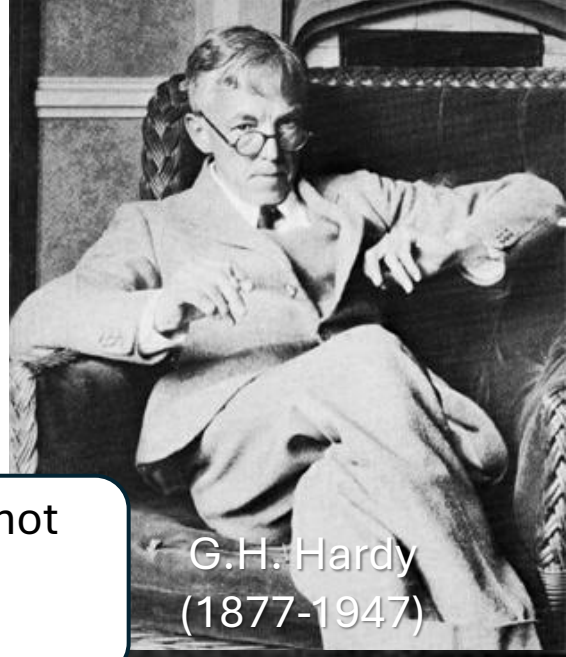
Assortative mating also determines departures from the Hardy-Weinberg equilibrium



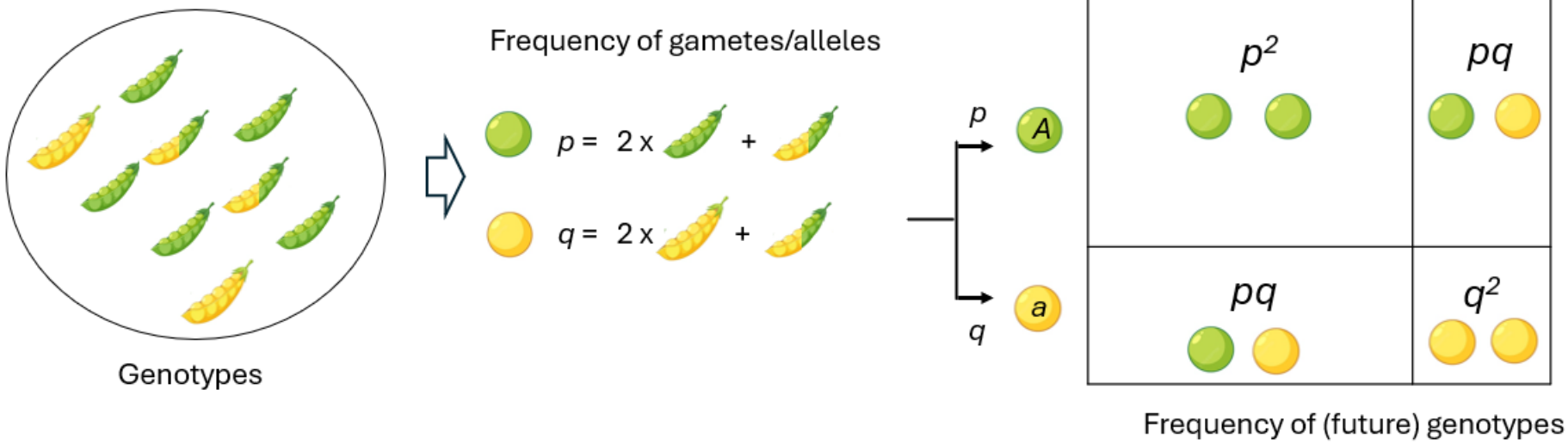
Robinson et al.,2017, Nat Hum Behavior

Hardy-Weinberg Equilibrium

For a randomly mating populations, where the transmission of gametes of an allele with frequency *p* is not affected by external forces (mutation, selection, **migration**, etc.), the expected proportion of genotypes in next generations is distributed as p^2, $2pq$, q^2



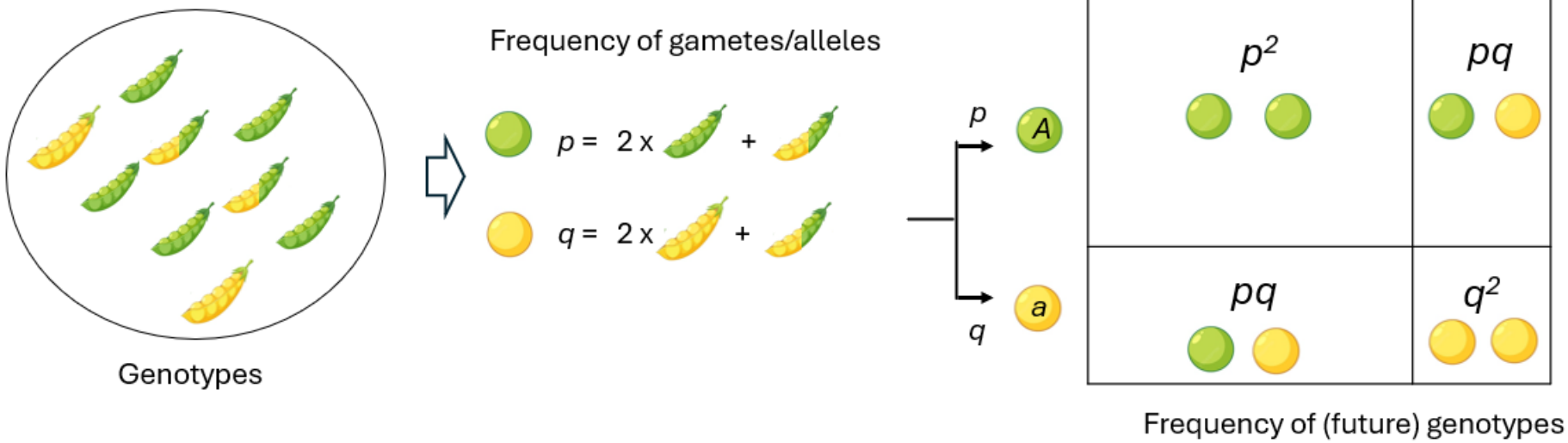
OK you are right, HW usually does not hold for real populations due space/substructure



Hardy-Weinberg Equilibrium

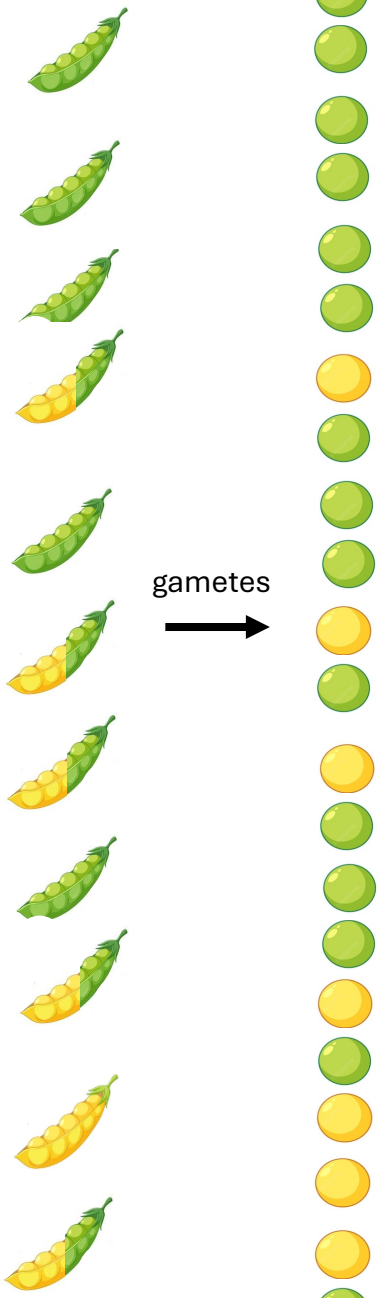
For a randomly mating populations, where the transmission of gametes of an allele with frequency *p* is not affected by external forces (mutation, selection, **migration**, etc.), the expected proportion of genotypes in next generations is distributed as p^2, $2pq$, q^2

But there is an even bigger problem!



Genotypes
p^2, $2pq$, q^2

Allele/gametes p,q



gametes

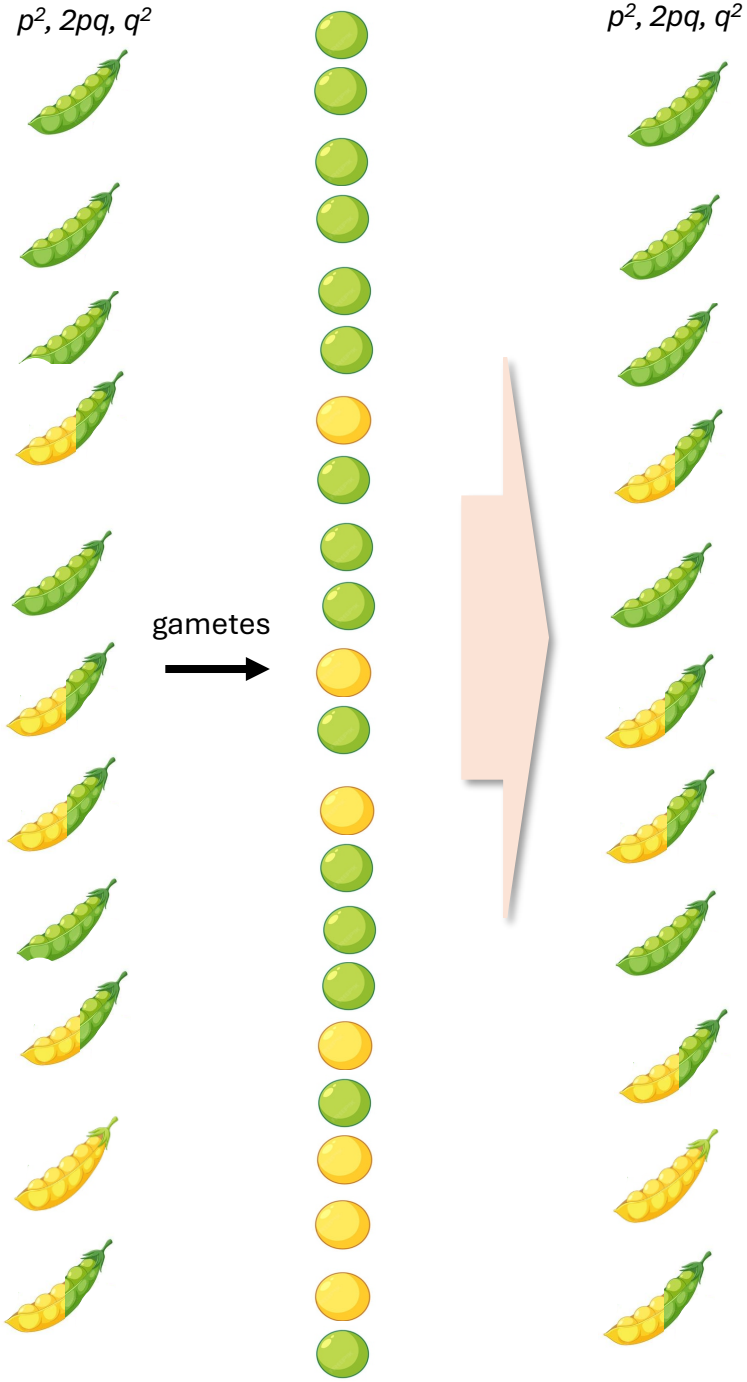
There is an even bigger problem!

G.H. Hardy
(1877-1947)

Genotypes
p^2, $2pq$, q^2

Allele/gametes p,q

Genotypes
p^2, $2pq$, q^2



There is an even bigger problem!

Genotypes
p^2, $2pq$, q^2

Allele/gametes p,q

gametes

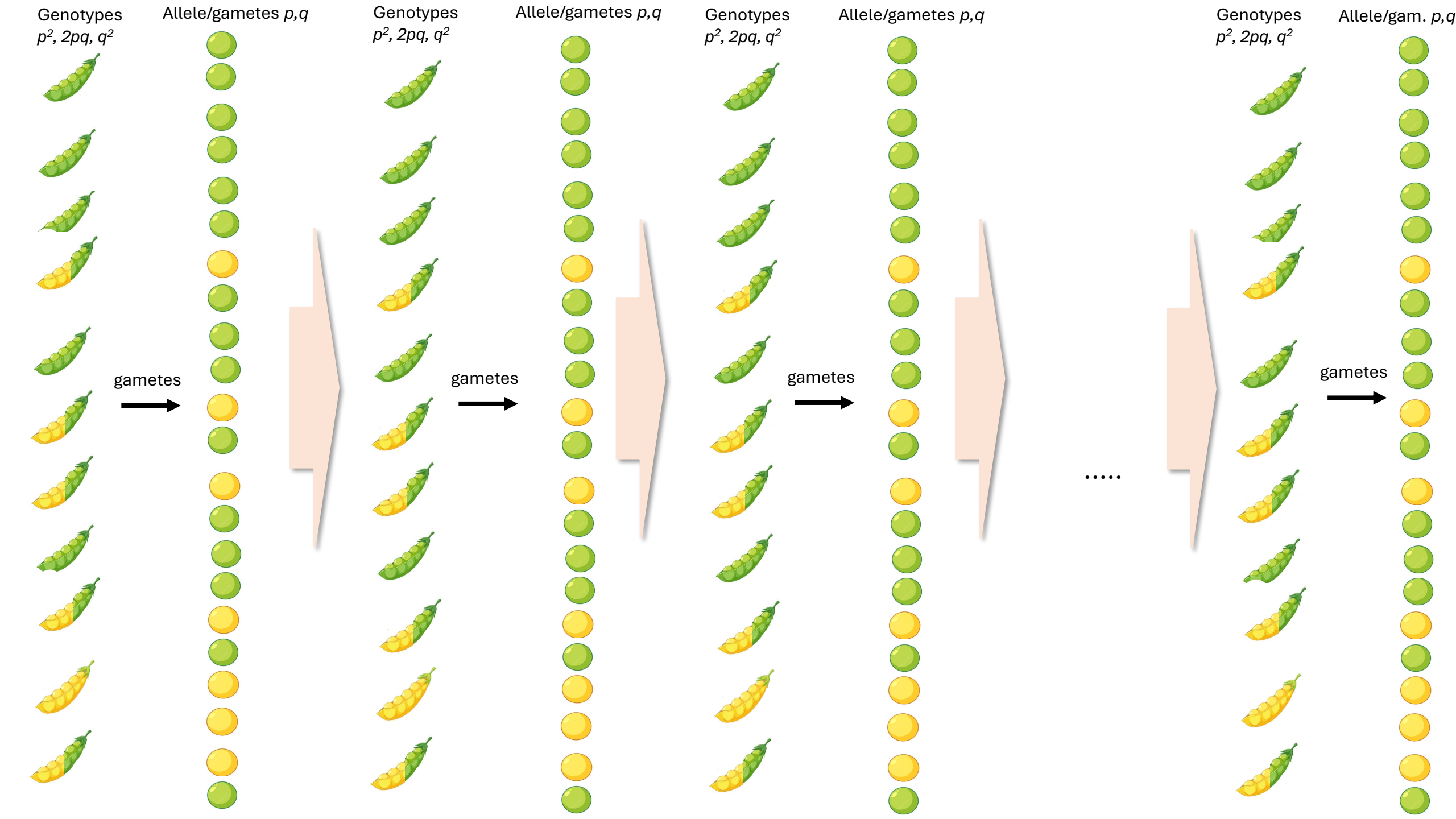
Genotypes
p^2, $2pq$, q^2

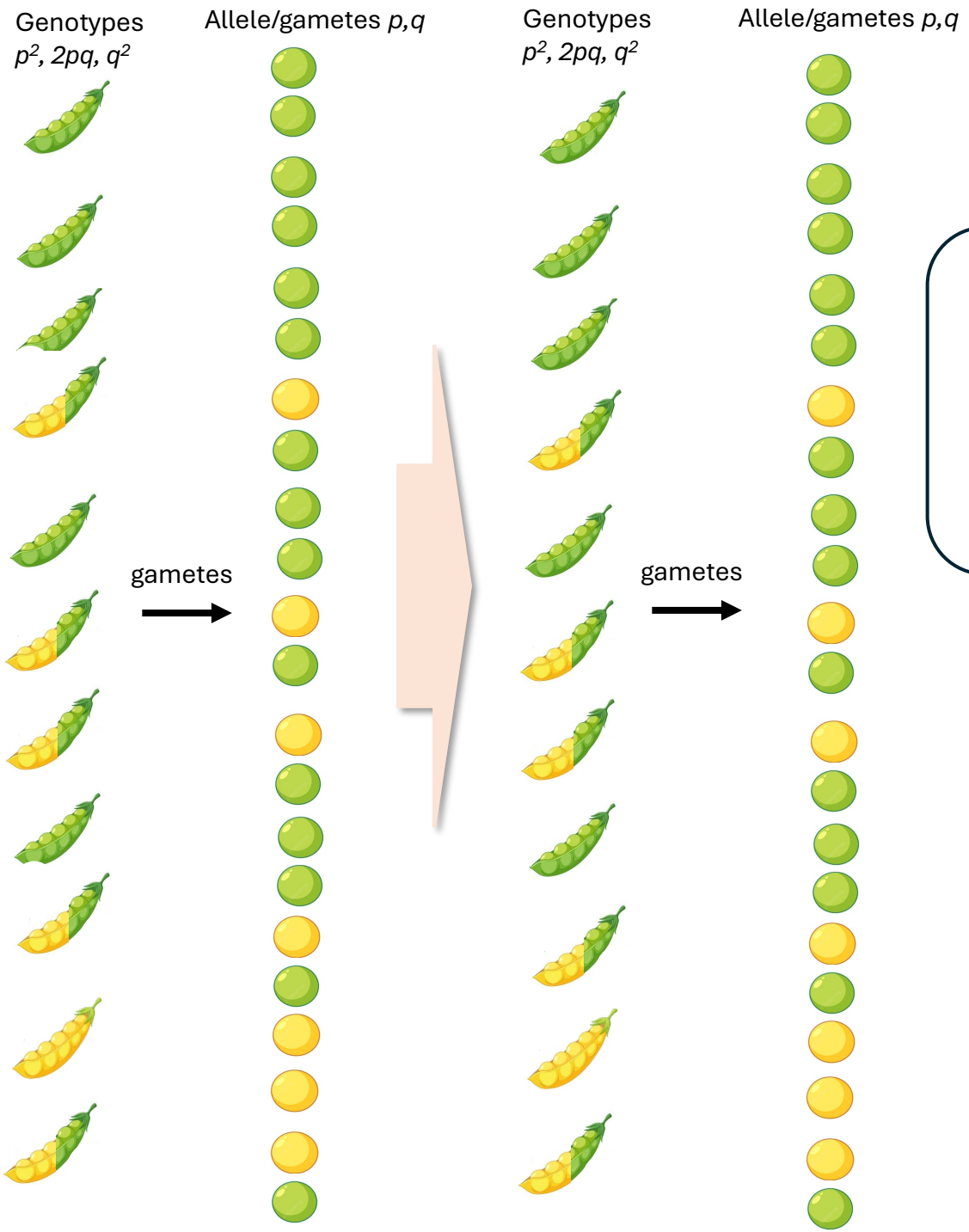
Allele/gametes p,q

gametes

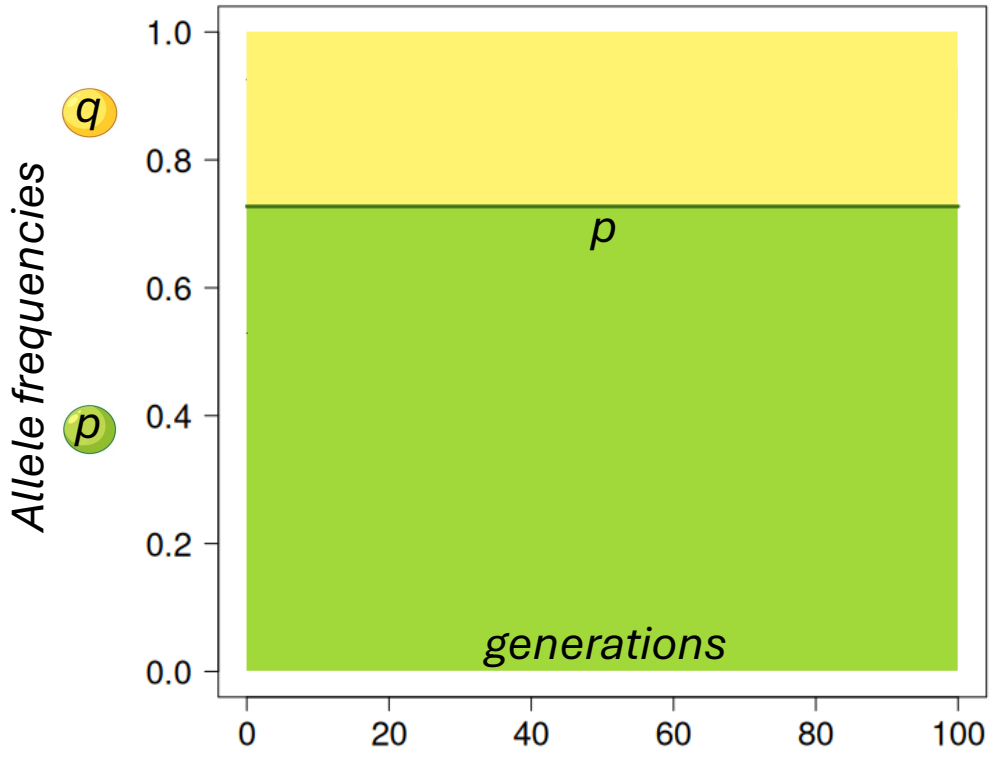
G.H. Hardy
(1877-1947)

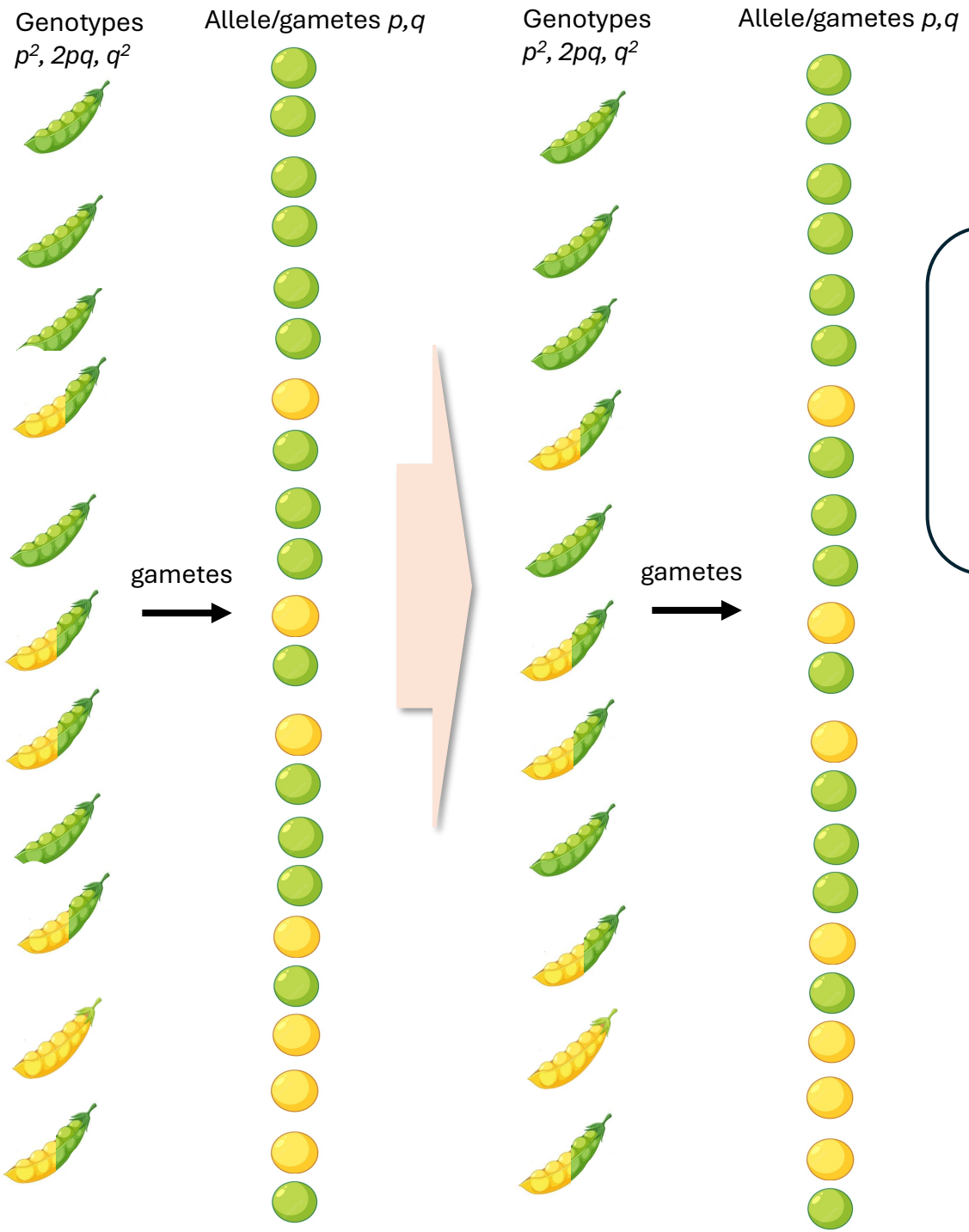
There is an even bigger problem!


Genotypes
p^2, 2pq, q^2
Allele/gametes p,q
gametes
Genotypes
p^2, 2pq, q^2
Allele/gametes p,q
gametes
Genotypes
p^2, 2pq, q^2
Allele/gametes p,q
gametes
.....
Genotypes
p^2, 2pq, q^2
Allele/gam. p,q
gametes

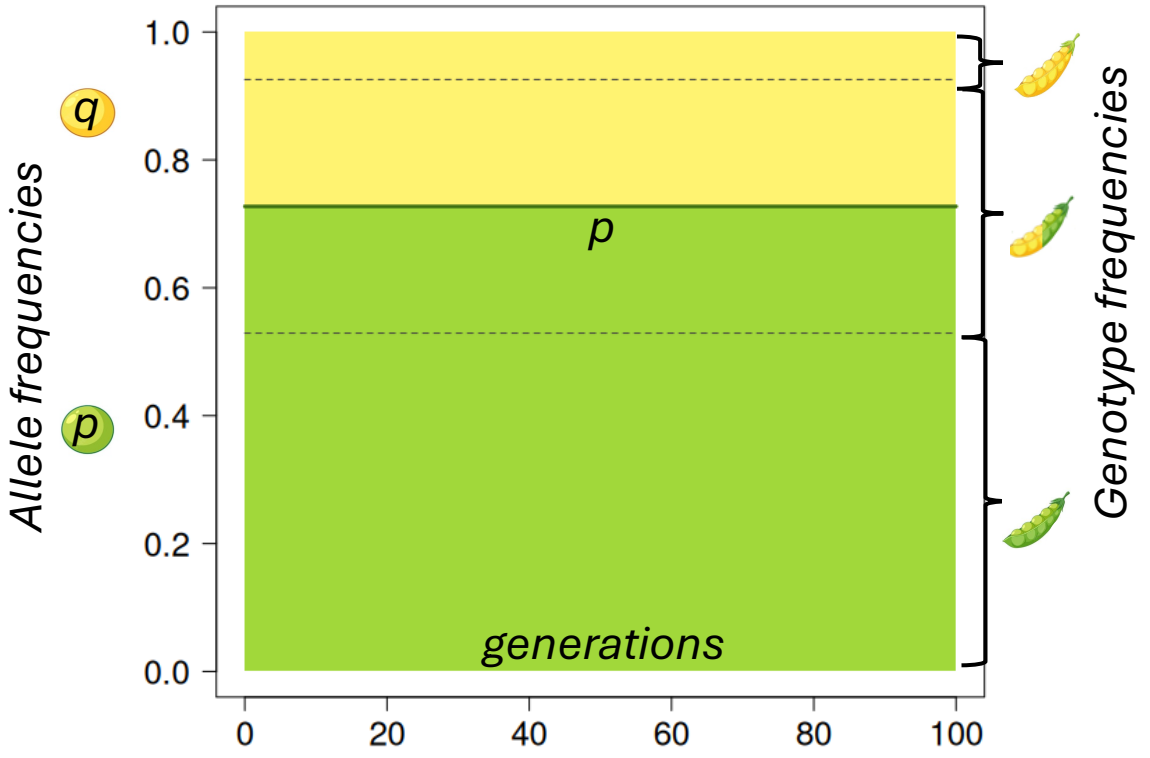


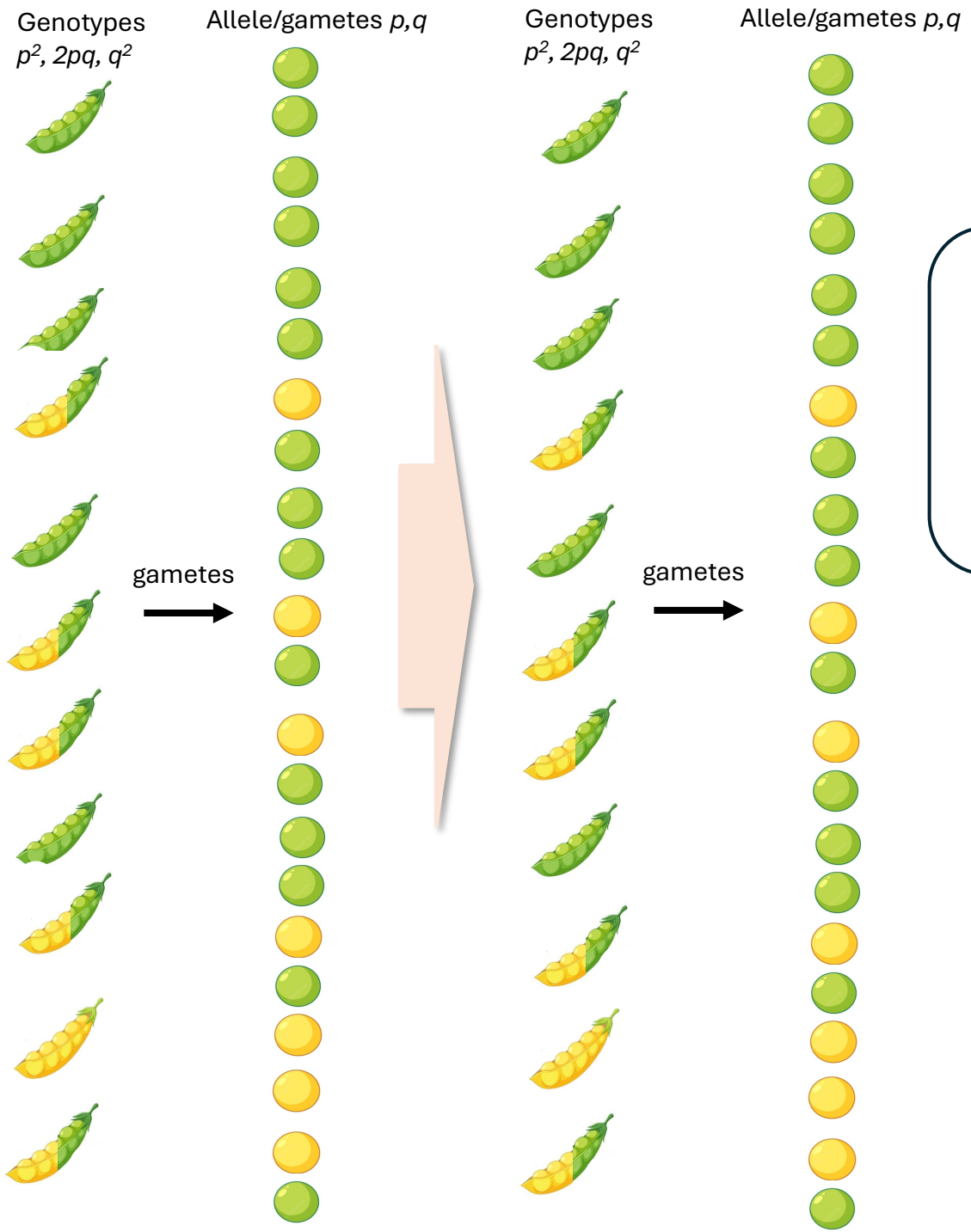
The allele frequencies do not change over time..



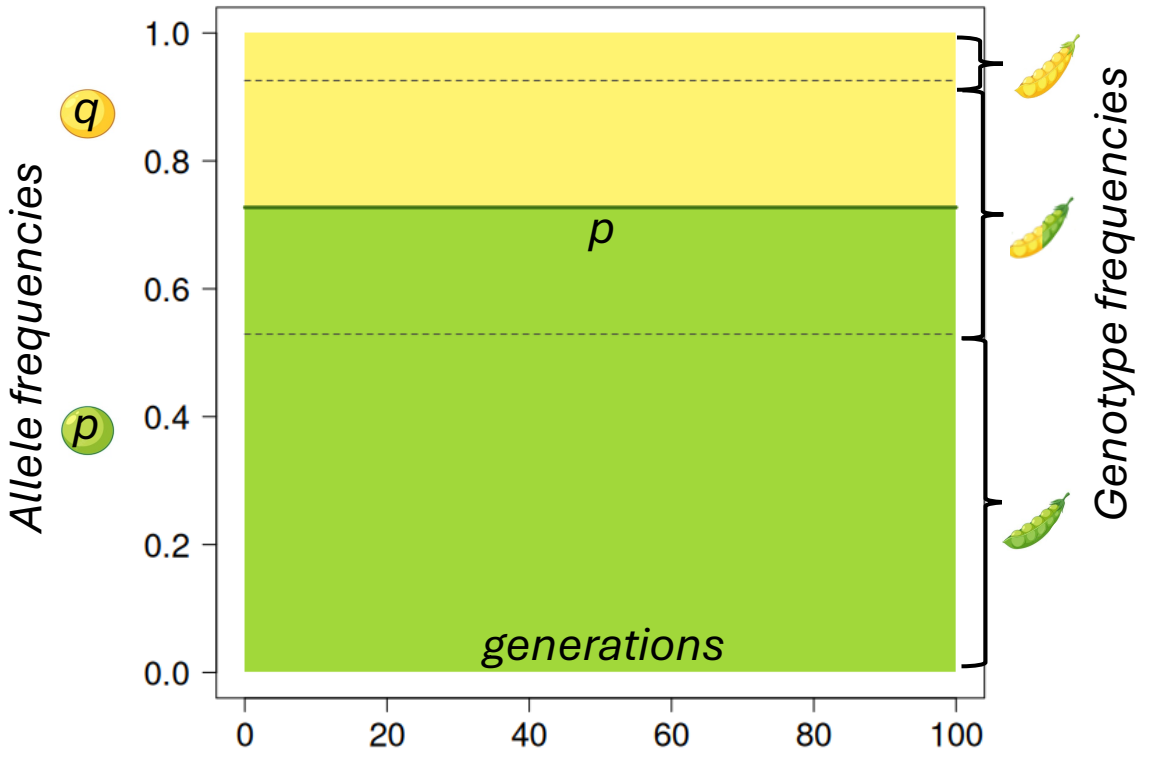


..and neither the genotype frequencies..





However, this is only the «expected value» (average): that is why sometimes people say that the «Hardy-Weinberg equilibrium» only works for infinite populations!



Genotypes
p^2, $2pq$, q^2

Allele/gametes p,q

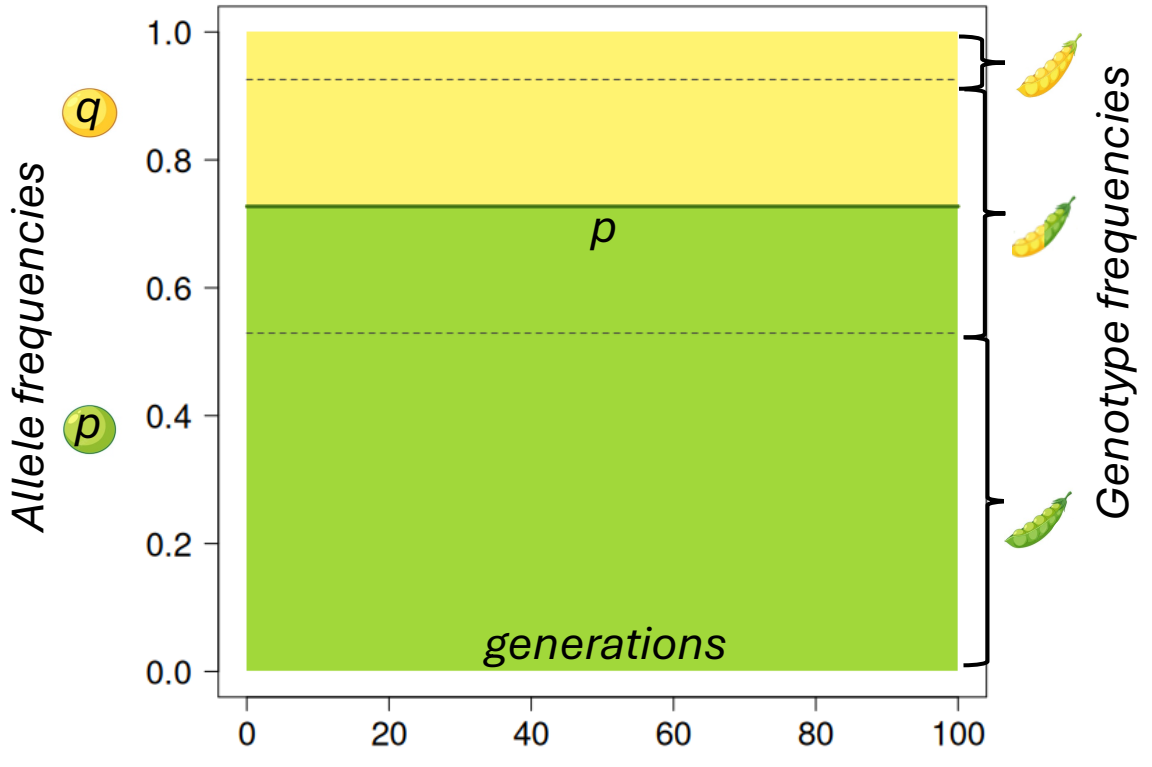
Genotypes
p^2, $2pq$, q^2

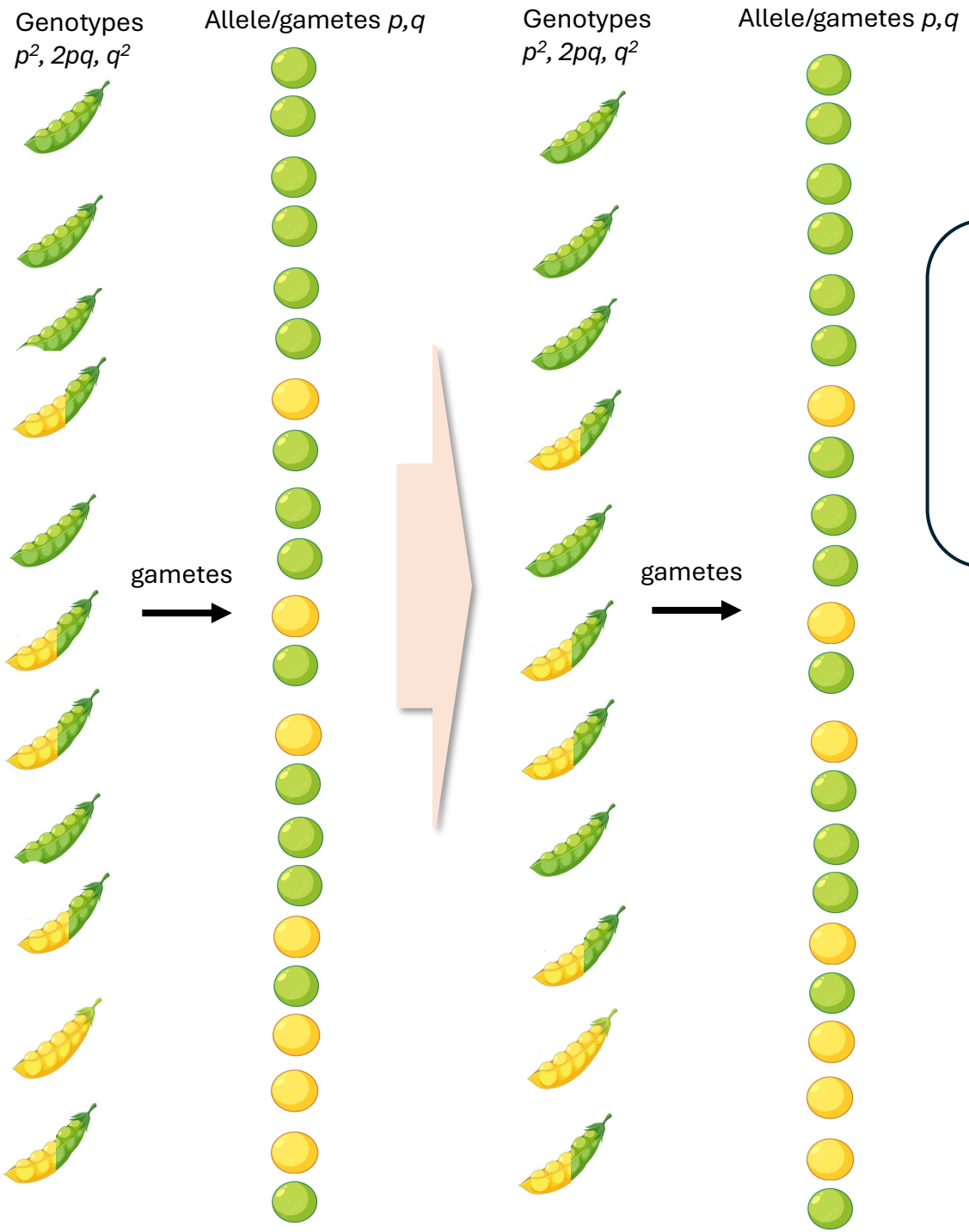
Allele/gametes p,q

Note that I explicitly wrote this in my 1908 letter to Science (see Moodle)!

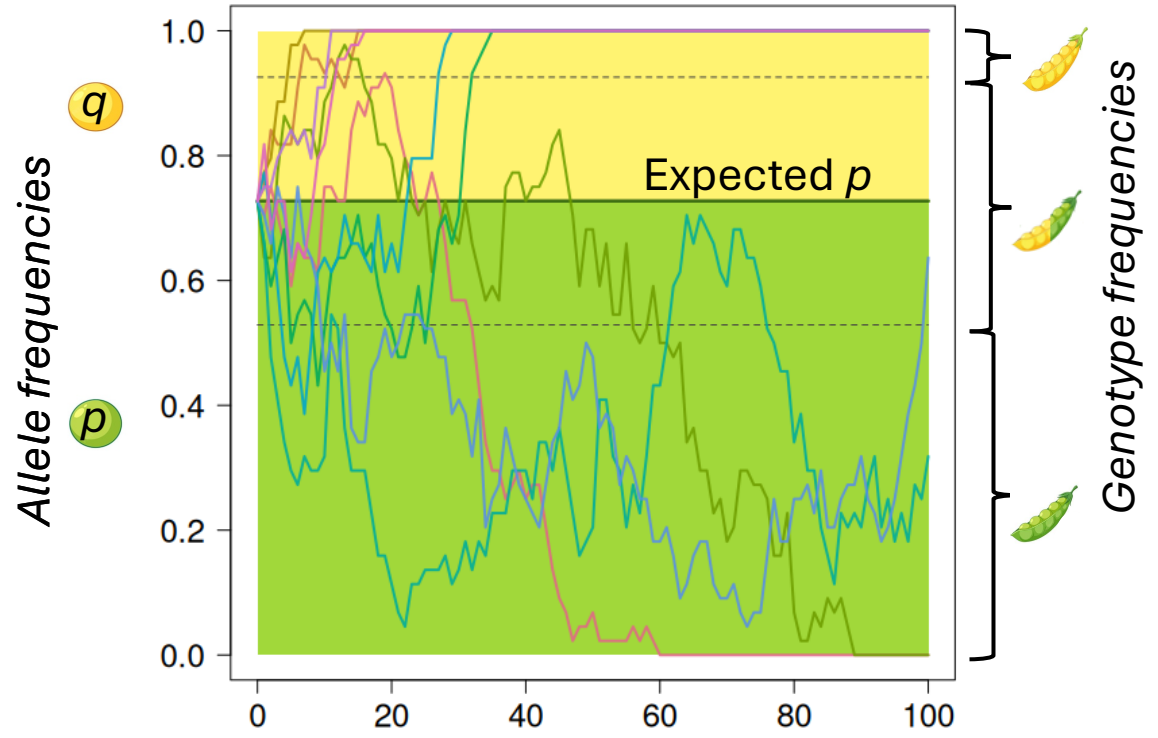
I just wanted to make the point that on average allele genotype frequencies stay the same regardless of dominance and that Yule was wrong: brachydactyly can be dominant and yet not spread in a population!

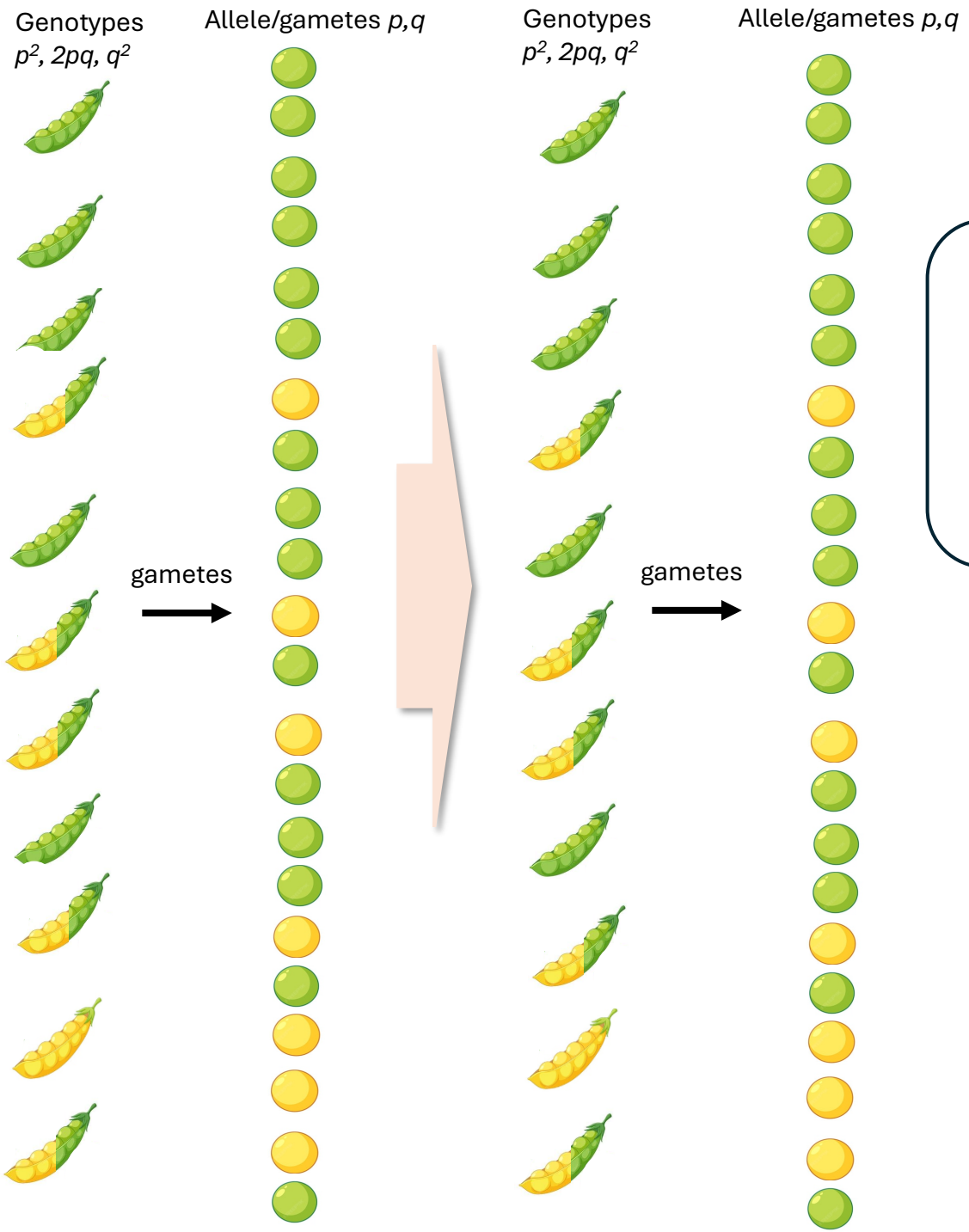
However, this is only the «expected value» (average): that is why sometimes people say that the «Hardy-Weinberg equilibrium» only works for infinite populations!





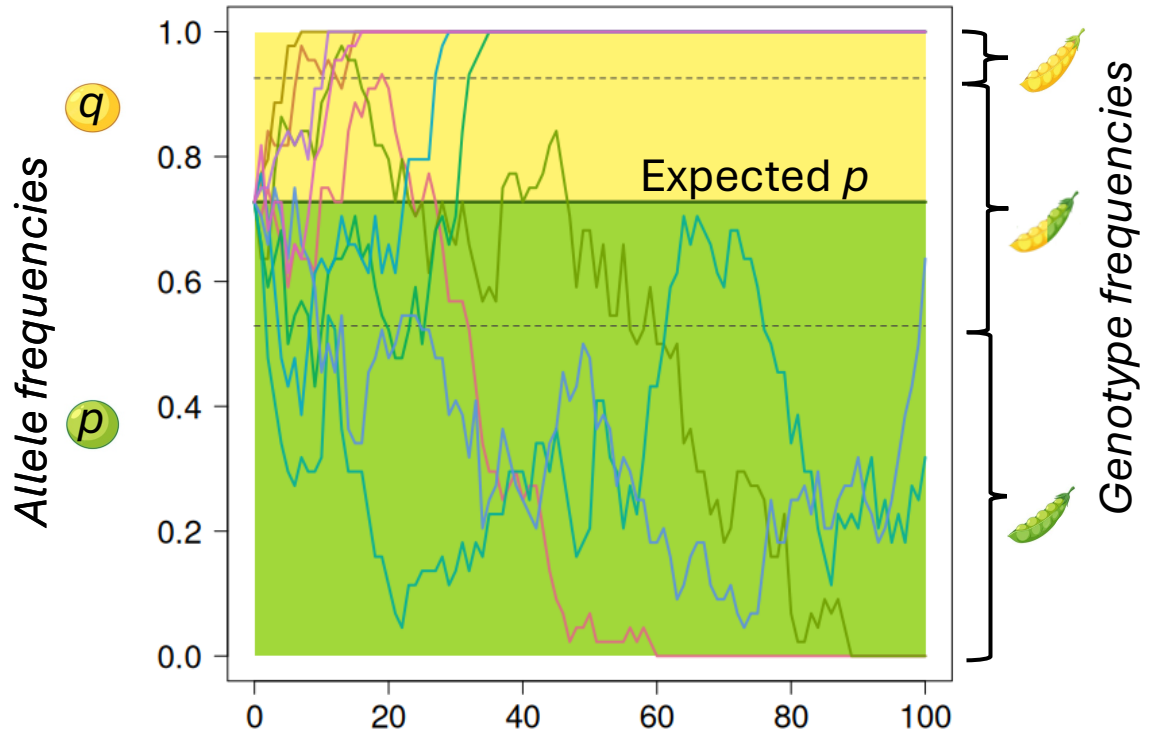
In real (finite) populations there is sampling noise!





In real (finite) populations there is sampling noise!

Such sampling noise is also called «**random drift**» or «**genetic drift**»!



Allele frequencies change even just because of random sampling

Let’s sample a population of 10 gametes

```
s = sample(c("A", "a"), size = 10, prob = c(p,q), replace = T)
```

Allele frequencies change even just because of random sampling

Let's sample a population of 10 gametes

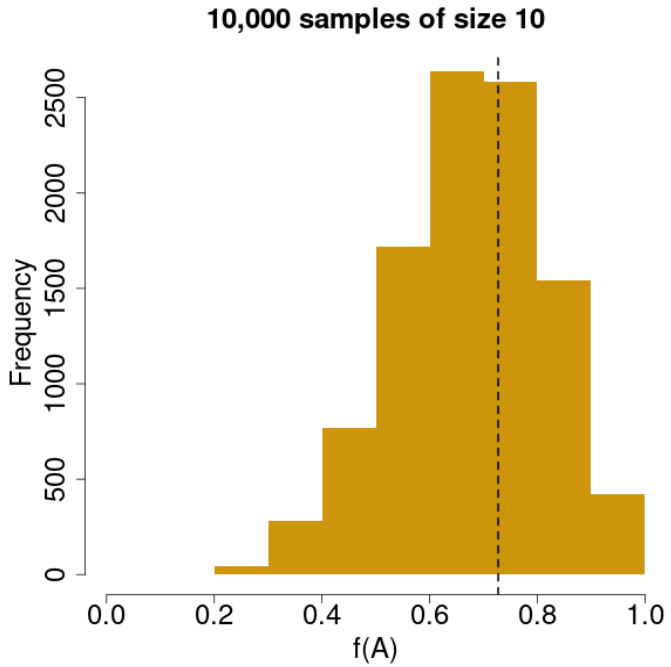
```
s = sample(c("A", "a"), size = 10, prob = c(p,q), replace = T)
```

The option `replace=T` indicates that the «sampling» of gametes does not remove them from the gene pool.

It is like saying that even if there is a finite number of individuals, anybody can reproduce more than once.

Allele frequencies change even just because of random sampling

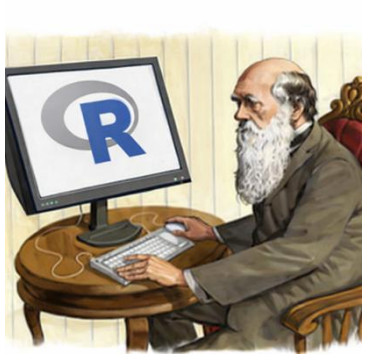
Let's sample a population of 10 gametes 10.000 times and explore the sampling variance

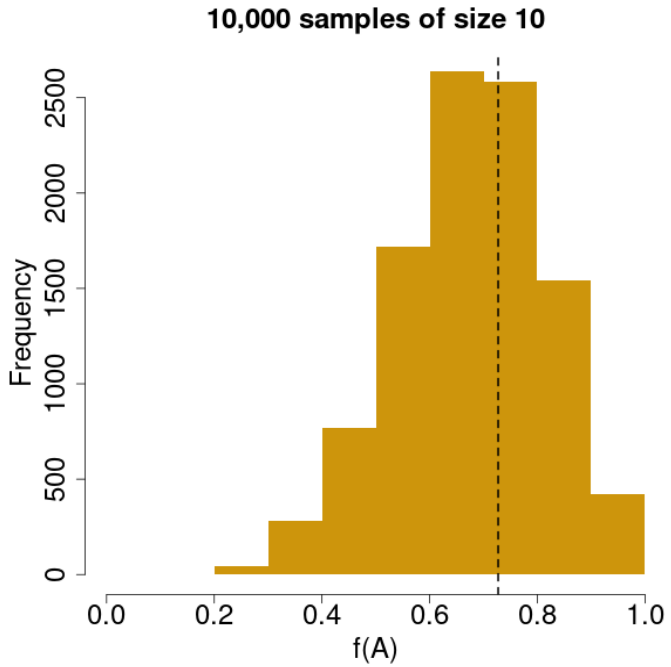


In a very small population even after one generation the frequency of *A (p)* can change drastically

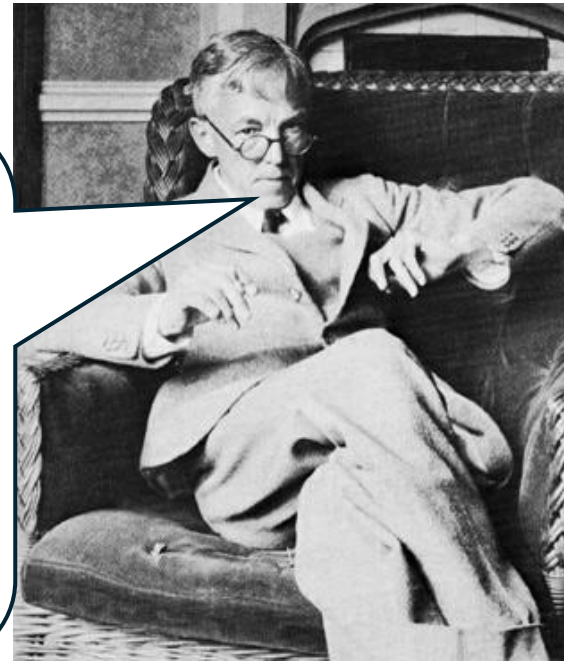
Allele frequencies change even just because of random sampling

Let's sample a population of 10 gametes 10.000 times and explore the sampling variance





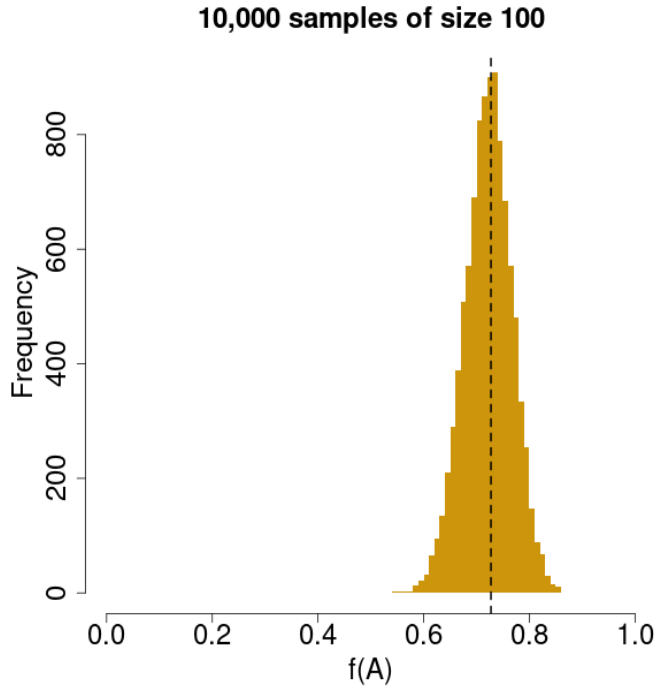
In a very small population even after one generation the frequency of *A (p)* can change drastically



```
npop=10; y=c()
for (i in 1:10000)
{
    x<-sample(c("A", "a"), size = npop, prob = c(p,q), replace = T)
    x<-sum(x=="A")/npop
    y=c(y,x)
}
hist(y)
```

Allele frequencies change even just because of random sampling

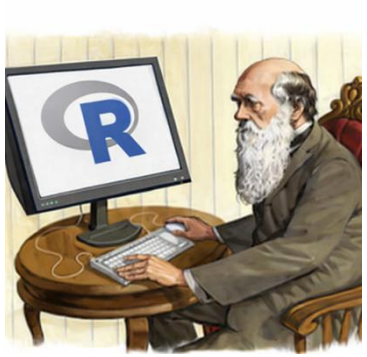
The sampling error is smaller as the population gets bigger

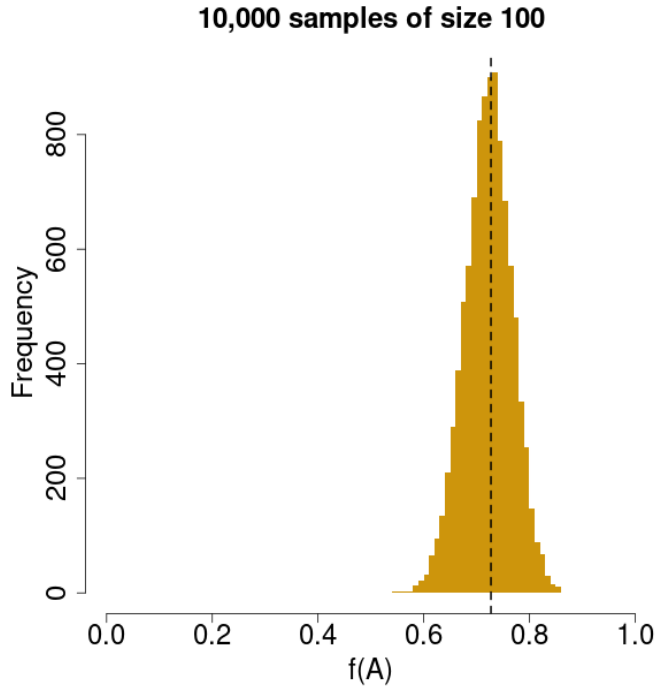


Even with 100 gametes the sampling error is quite large even after just 1 generation

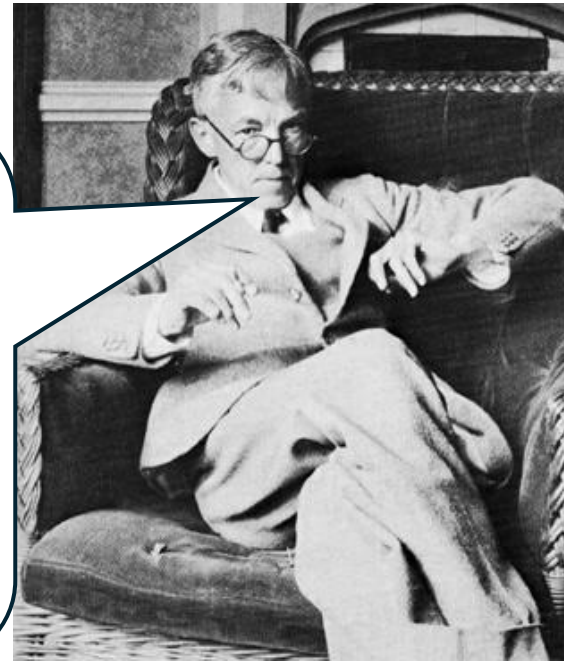
Allele frequencies change even just because of random sampling

The sampling error is smaller as the population gets bigger





Even with 100 gametes the sampling error is quite large even after just 1 generation



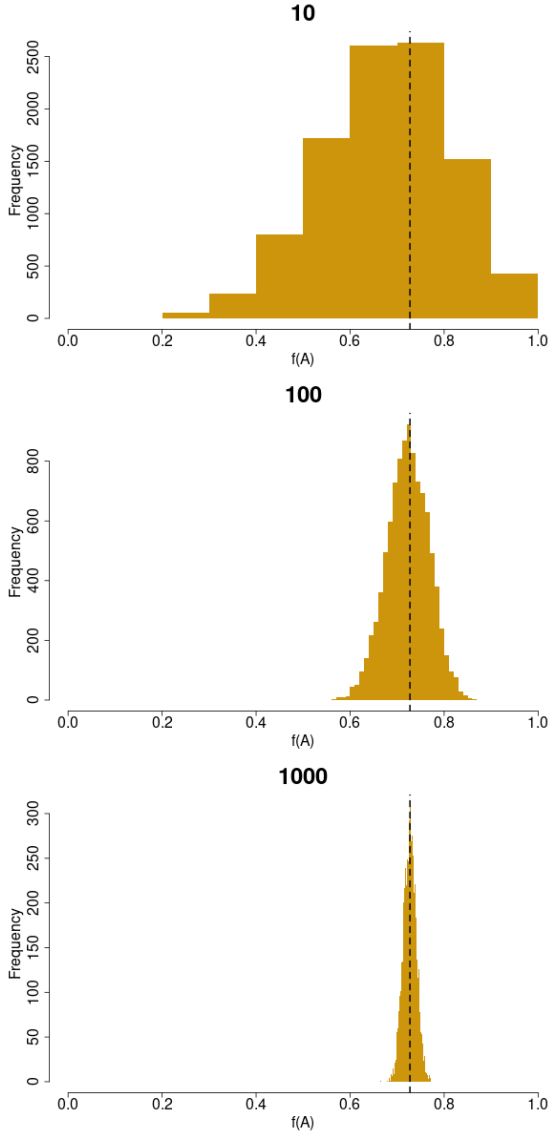
```
npop=100; y=c()
for (i in 1:10000)
{
    x<-sample(c("A", "a"), size = npop, prob = c(p,q), replace = T)
    x<-sum(x=="A")/npop
    y=c(y,x)
}
hist(y)
```

Random sampling is stronger in small populations

The larger the size of the sample, the smaller the variation in respect to the original allele frequencies.

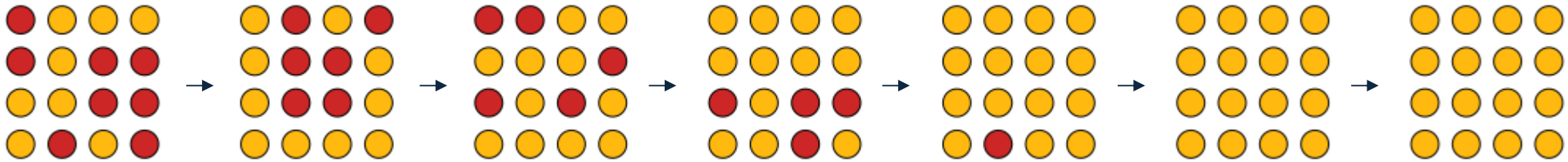
However, even for moderately large populations, random drift cannot be neglected!

We will see that drift is arguably the strongest force in evolution by far.

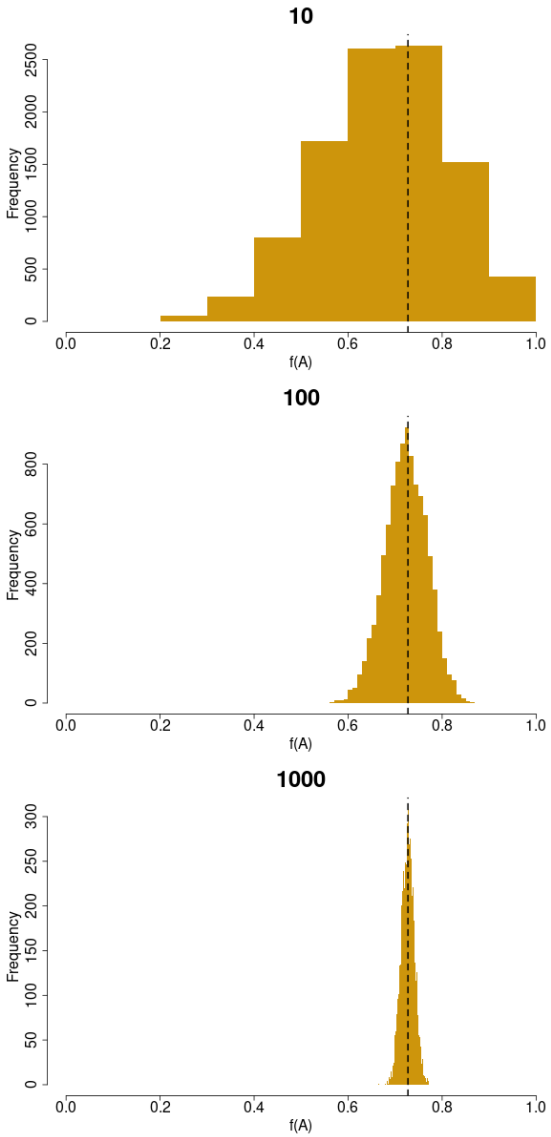


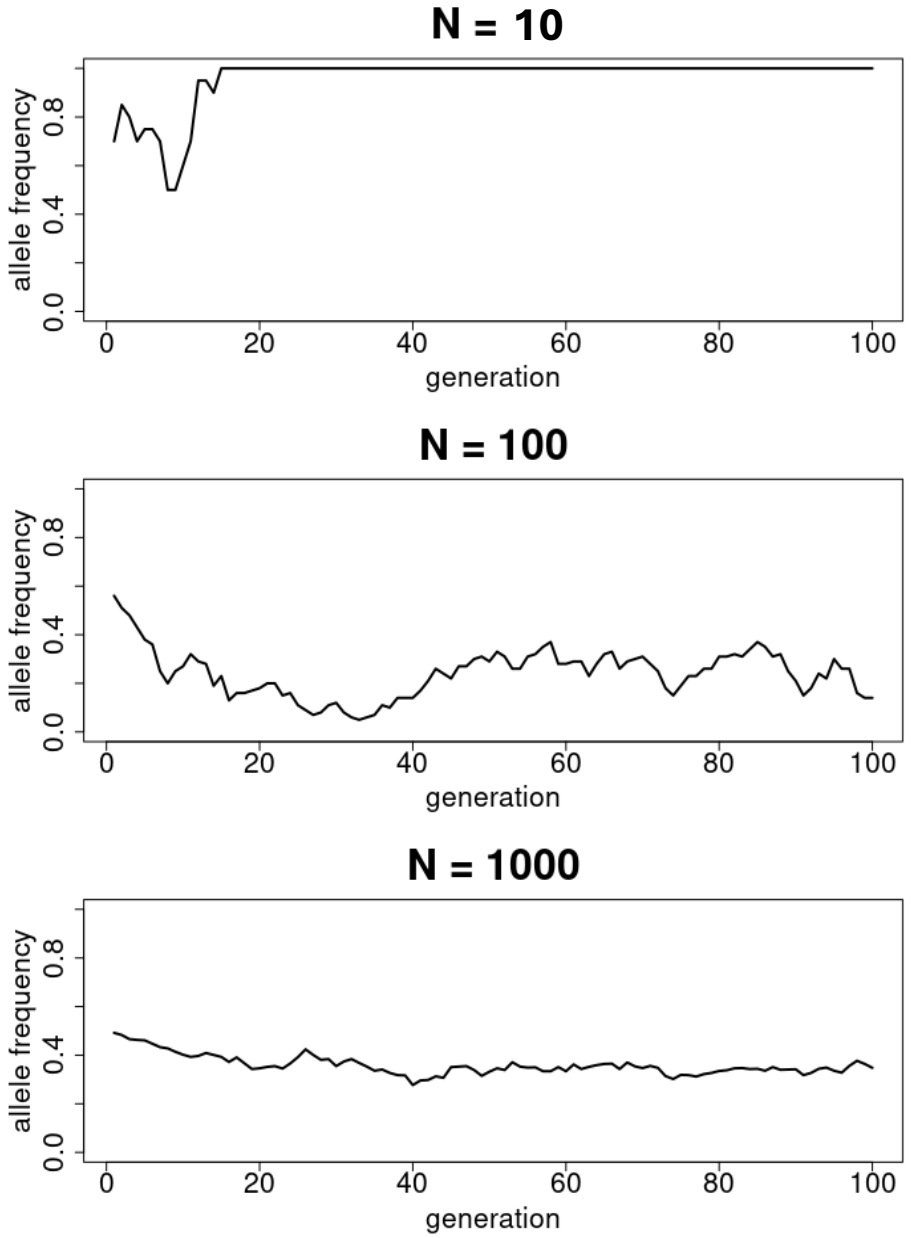
Random *genetic drift*

Sampling error propagates **across generations**!



Random genetic drift is stronger in small populations

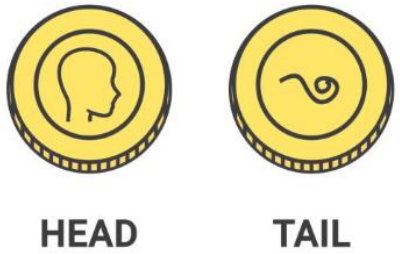




That’s why we need *Probability*!

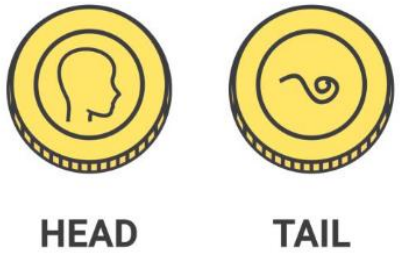
That's why we need *Probability*!

- **Random variables**:
  - In *Pr(X=a)*, *X* is called a ***random variable***, *a* is called value.
  - e.g. For a balanced coin *Pr(X=Head)=0.5*, X can take the values Head or Tail



That's why we need *Probability*!

- **Random variables**:
  - In *Pr(X=a)*, *X* is called a ***random variable***, *a* is called value.
  - e.g. For a balanced coin *Pr(X=Head)=0.5*, X can take the values Head or Tail

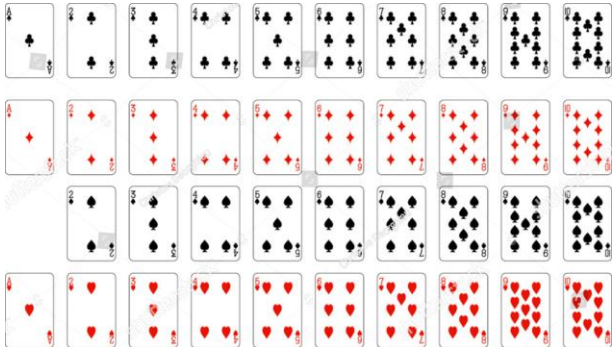


- **Independent events**:
  - the joint probability of two independent events A and B equals
    - *Pr(A,B)=Pr(A)Pr(B)*
  - e.g. The probability of obtaining six twice is 1/6 x 1/6

That's why we need *Probability*!

- **Independent events**:
  - the joint probability of two independent events A and B equals
    - $Pr(A,B)=Pr(A)Pr(B)$
  - e.g. The probability of obtaining six twice is 1/6 x 1/6

- **Dependent events**
  - The joint probability that A and B occur, if A depends on B, equals:
    - $Pr(A,B)=Pr(A|B)Pr(B)$ where $Pr(A|B)$ is the «conditional probability of A given B»
  - e.g. The joint probability of extracting two aces from a deck of 52 cards, Pr(A,B) is given by the probability of extracting a first ace Pr(A)=4/52 and that of extracting a second one given that a first was already sampled, or Pr(B|A)=3/51



That's why we need *Probability*!

- Law of total probability:
  - The probability of an event A is the sum of the conditional probabilities for all possible conditioning events
    - e.g. The probability of a day of rain is the sum of conditional probabilities of rain given that the season is summer, autumn, winter or spring

$$P(A) = \sum_{n} P(A \mid B_n)P(B_n)$$

Pr(rain)=Pr(rain|summer)Pr(summer)+Pr(rain|autumn)Pr(autumn)+Pr(rain|winter)Pr(winter)+Pr(rain|spring)Pr(spring)

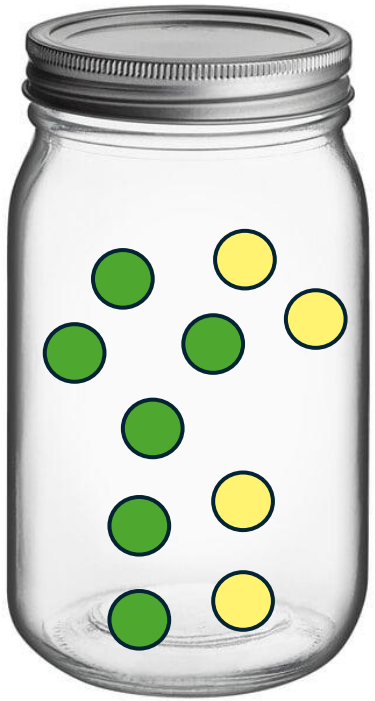
= [Pr(rain|summer)+Pr(rain|autumn)+Pr(rain|winter)+Pr(rain|spring)]/4

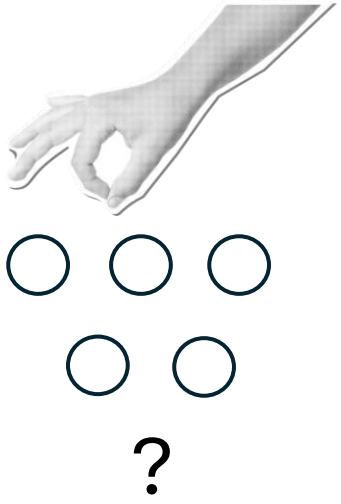
Probability distributions

Functions that describe the probability of an event as a function of one or more parameters (e.g. mean, rate, variance, shape)

The binomial distribution describes the probability to observe *k* successes over *n* total events

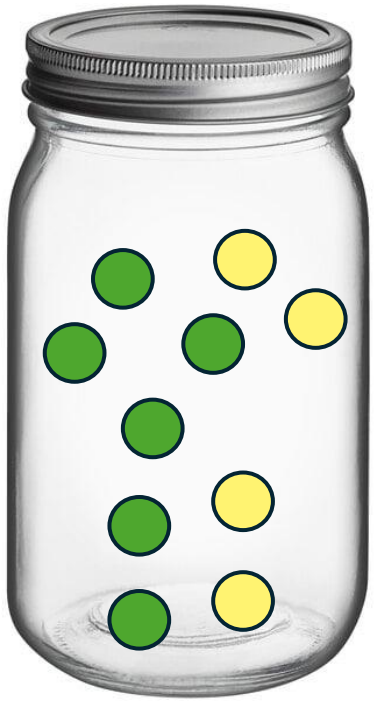
- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (5) marbles, what is the probability of extracting ***k*** green marbles?

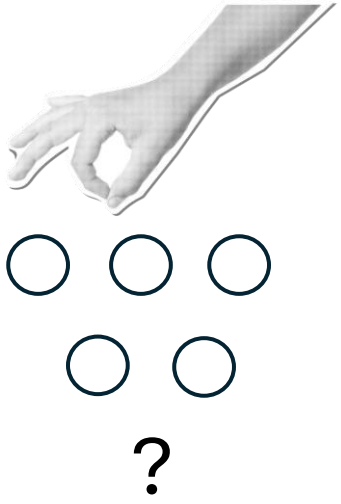




The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (5) marbles, what is the probability of extracting ***k*** green marbles?

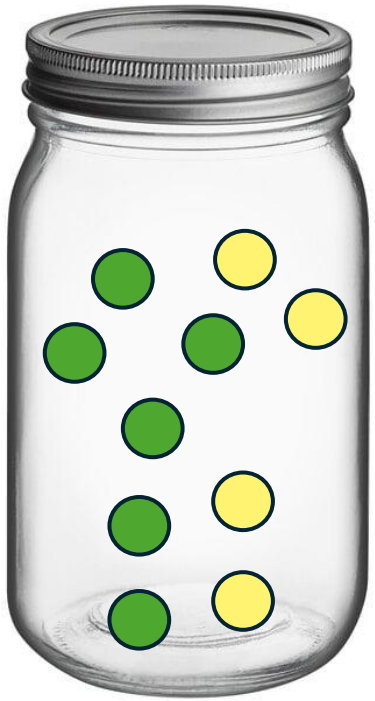


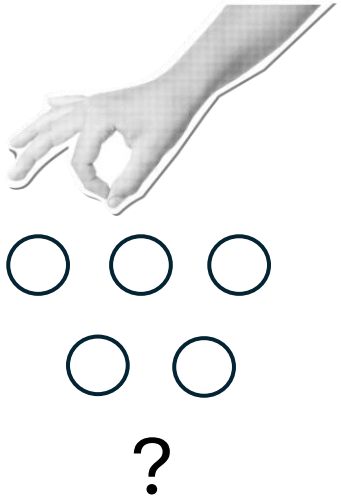


$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (5) marbles, what is the probability of extracting ***k*** green marbles?



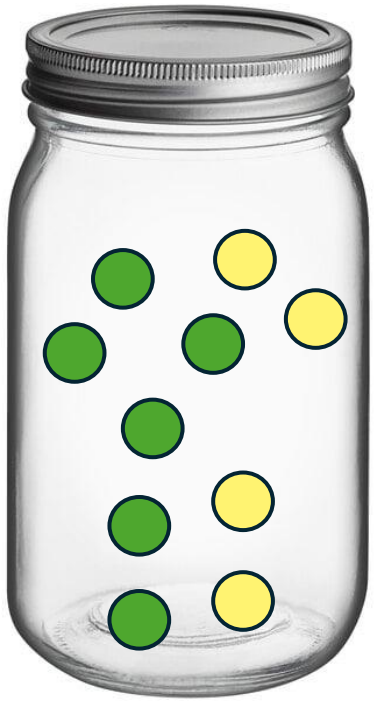


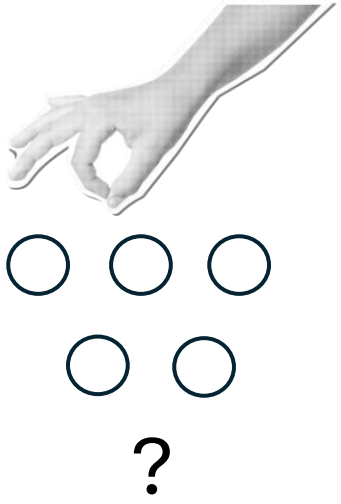
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

Note that we ask for the probability to observe k given the parameters p and n

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (5) marbles, what is the probability of extracting ***k*** green marbles?

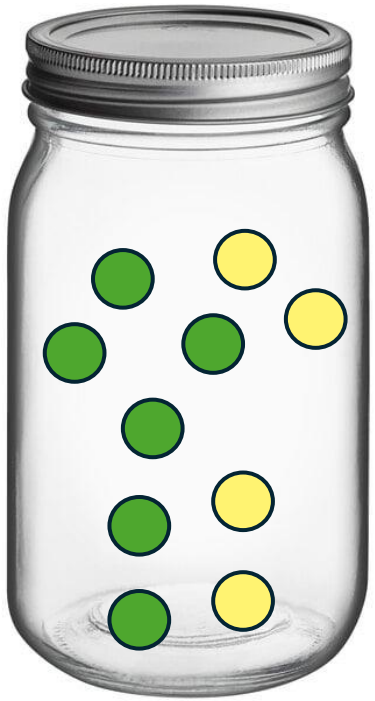


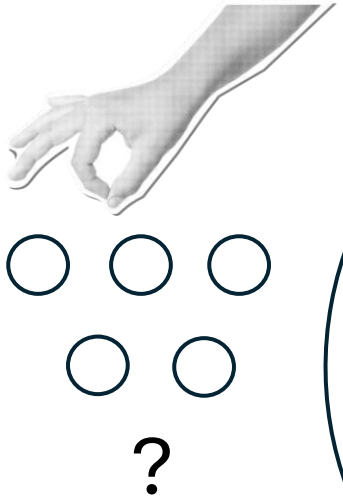


Note that *p* (the proportion of green marbles in the jar) changes every time I extract a marble. When does it stay the same?

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (5) marbles, what is the probability of extracting ***k*** green marbles?





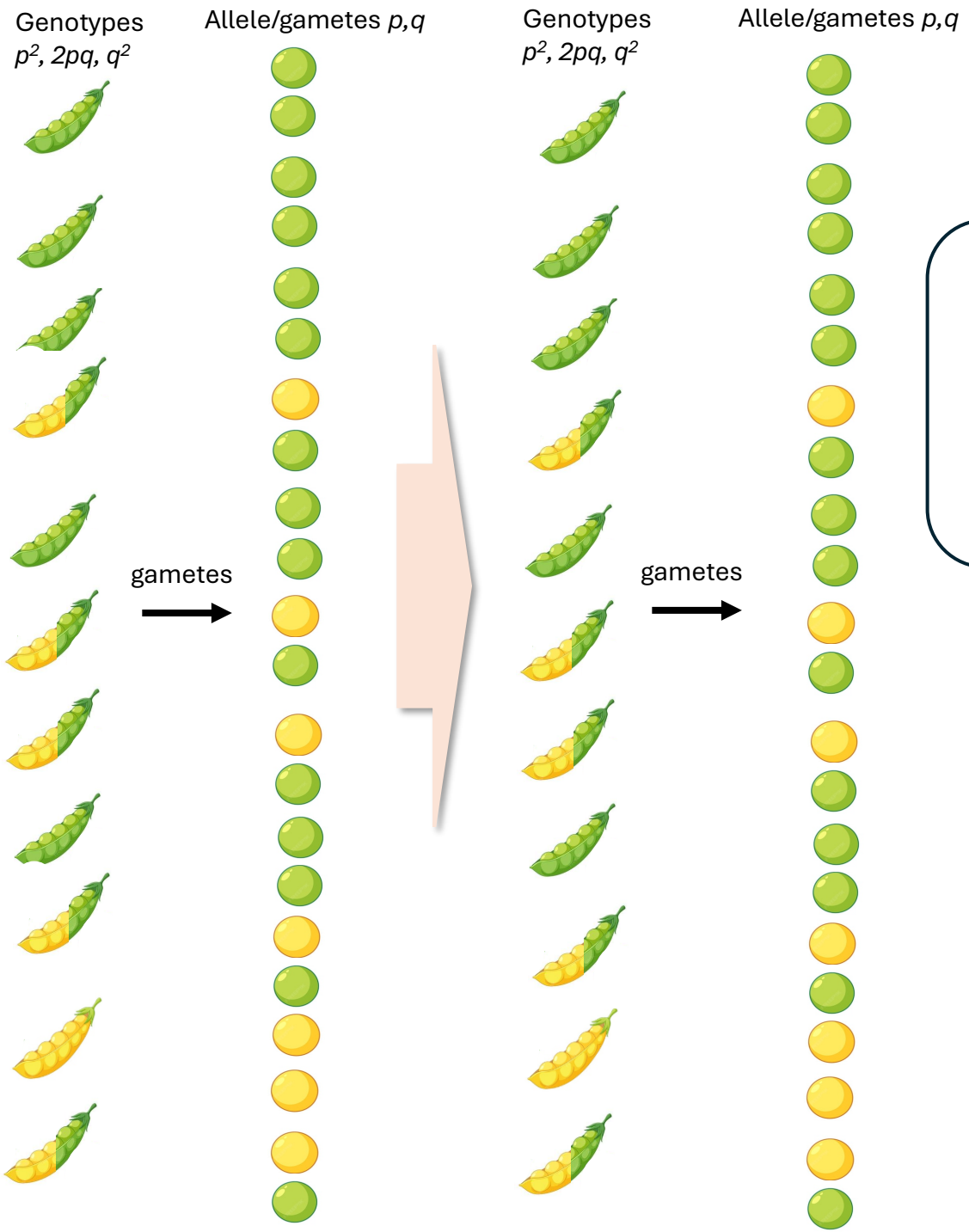
The binomial distribution describes either an infinite jar, or sampling WITH replacement from a population (I put back the marble in the jar after extracting it). When did we see this?

Allele frequencies change even just because of random sampling

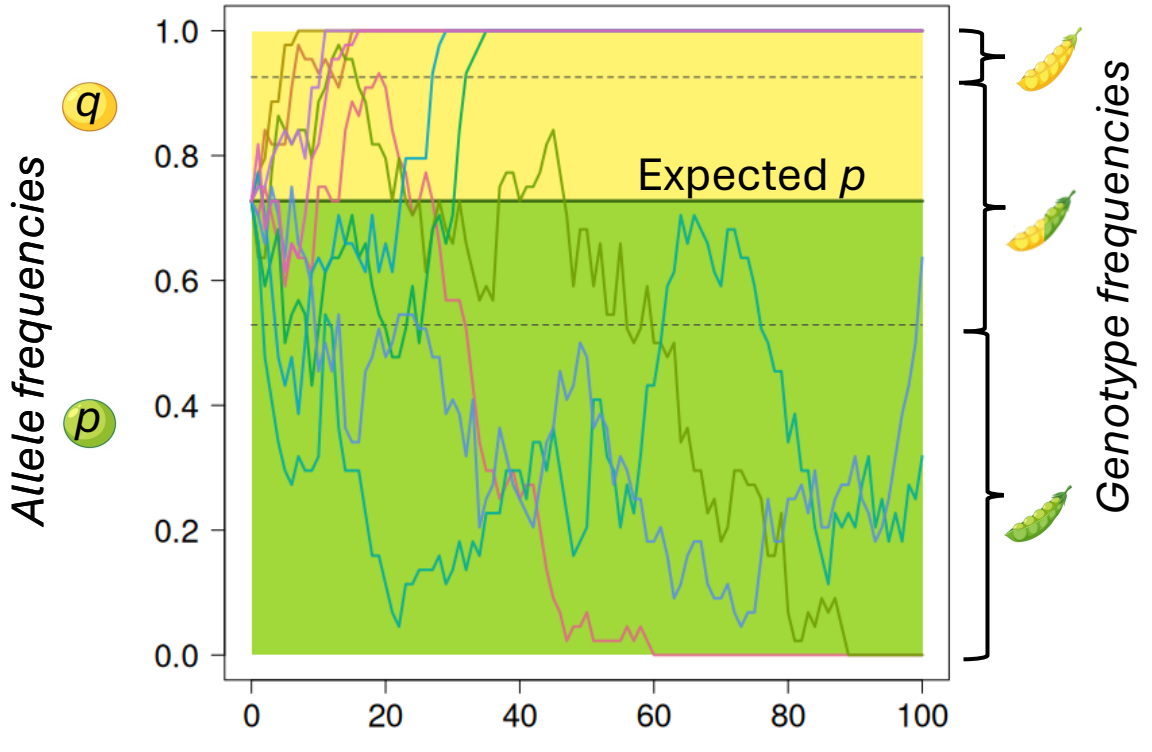
Let’s sample a population of 10 gametes

```
s = sample(c("A", "a"), size = 10, prob = c(p,q), replace = T)
```

The binomial distribution describe «sampling error» in our population model!

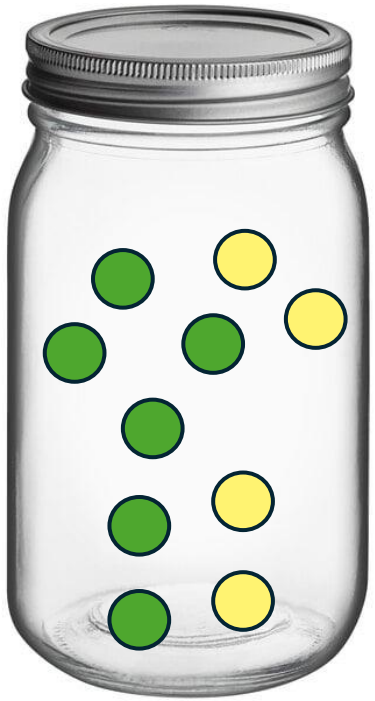


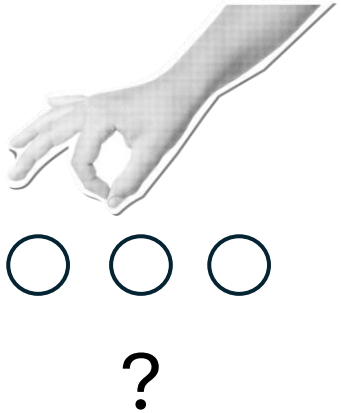
We will see soon in fact that the Wright-Fisher model (the most fundamental population genetic model capturing drift) is nothing else than a binomial distribution!



The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?





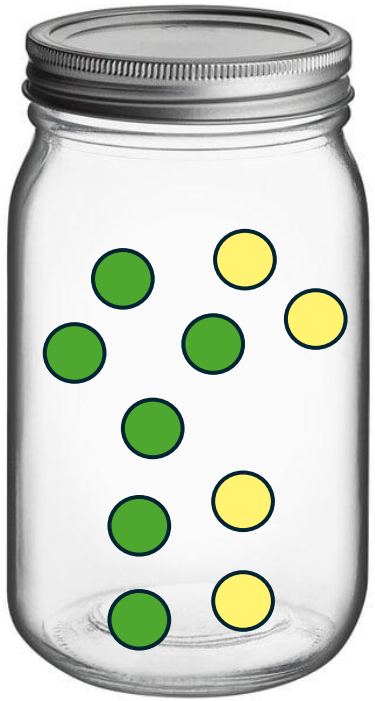
?

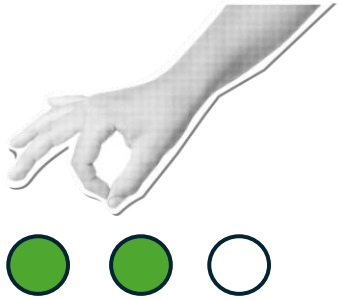
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

What are these two terms?

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?



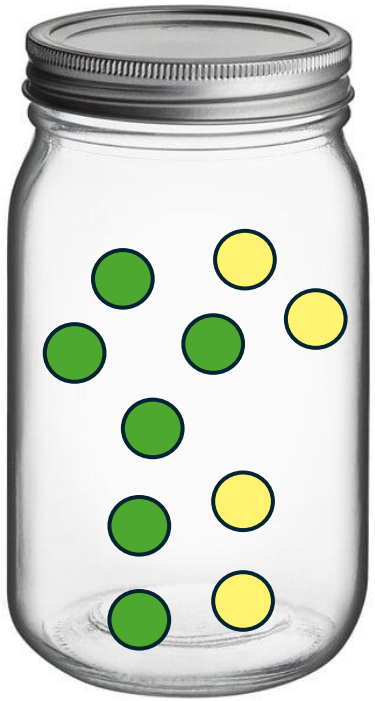


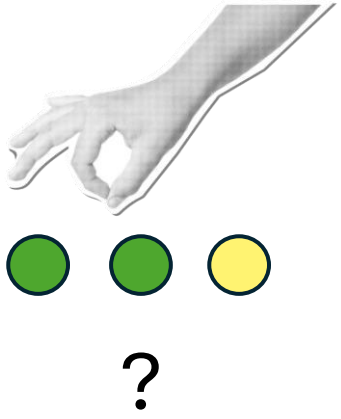
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

e.g. for k=2,
I need to sample twice the green marble $p^k=0.6^2$

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?





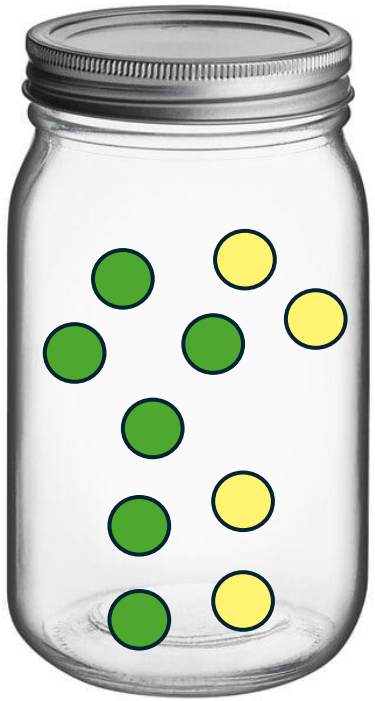
?

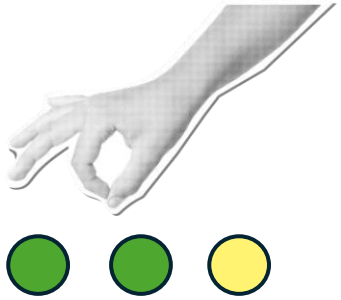
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

e.g. for k=2,
I need to sample twice the green marble $p^k=0.6^2$
and then once a yellow marble $(1-p)^{n-k}=0.4^1$

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?





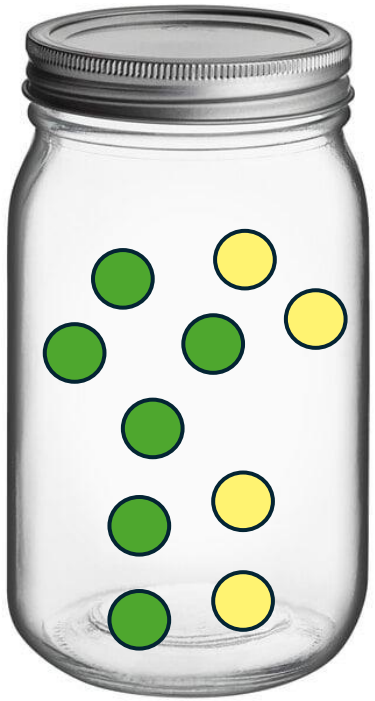
What is missing?

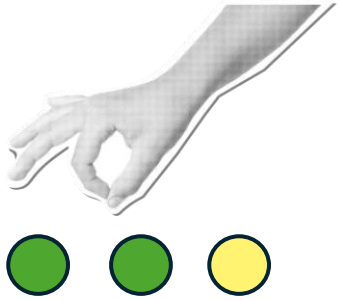
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

e.g. for k=2,
I need to sample twice the green marble $p^k=0.6^2$
and then once a yellow marble $(1-p)^{n-k}=0.4^1$
Thus the probability of an exact sequence of 2 green marbles and a yellow one is:
$p^k=0.6^2$

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?





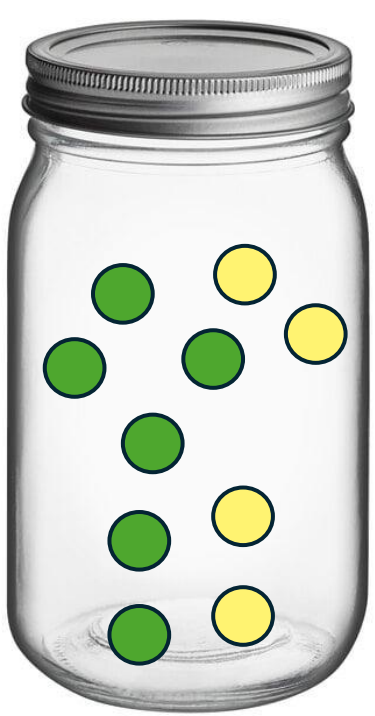
What is missing?

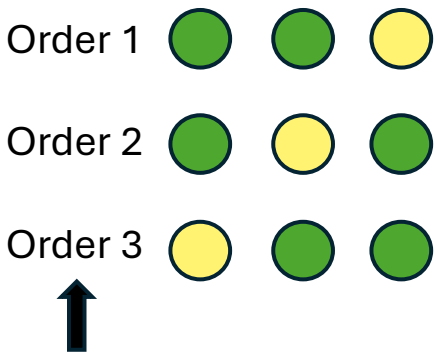
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

e.g. for k=2,
I need to sample twice the green marble $p^k=0.6^2$
and then once a yellow marble $(1-p)^{n-k}=0.4^1$
Thus the probability of an exact sequence of 2 green marbles and a yellow one is:
$p^k (1-p)^{n-k}=0.6^2\ 0.4=0.144$

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?





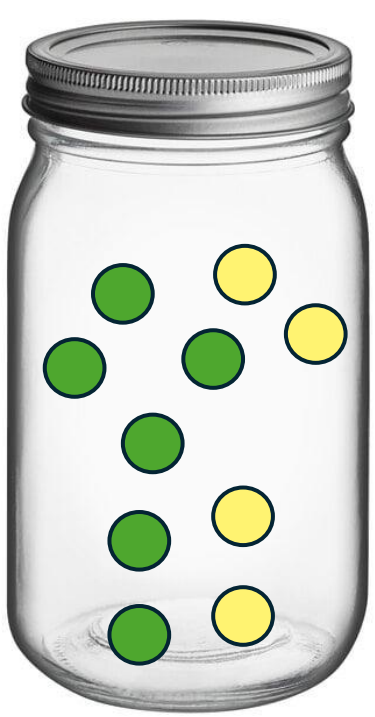
There can be more than one order to sample the colored marbles!

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

The $\binom{n}{k}=n!/(k!(n-k)!)$ is called «binomial coefficient» and describes the number of possible combinations of sampled marbles.

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?



Order 1

Order 2

Order 3

There can be more than one order to sample the colored marbles!

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

The $\binom{n}{k}=n!/(k!(n-k)!)$ is called «binomial coefficient» and describes the number of possible combinations of sampled marbles.

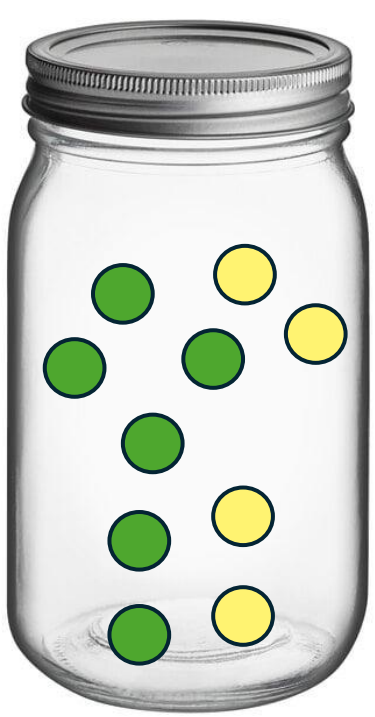
The factorial $n!=n(n-1)(n-2)....1$

e.g.

$\binom{3}{2} = 3!/(2!1!)=3*2/2=3$

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?



Order 1

Order 2

Order 3

Note that all order have the same proportions of green, so they have all the same probability!

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

The $\binom{n}{k}=n!/(k!(n-k)!)$ is called «binomial coefficient» and describes the number of possible combinations of sampled marbles.

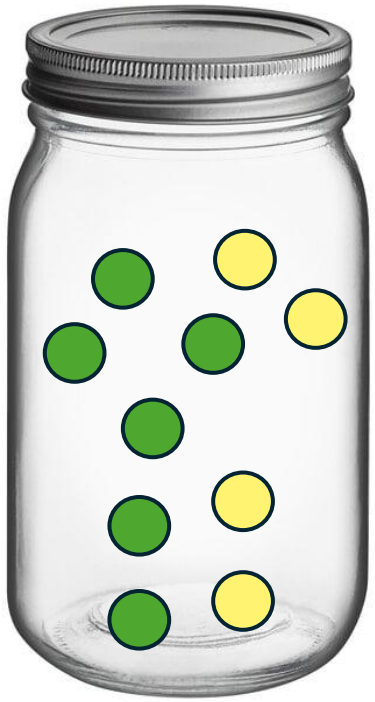
The factorial $n!=n(n-1)(n-2)....1$

e.g.

$\binom{3}{2} = 3!/(2!1!)=3*2/2=3$

The binomial distribution describes the probability to observe *k* successes over *n* total events

- E.g. If I have a jar with a proportion ***p*** (0.6) of green marbles and I extract a ***n*** (3) marbles, what is the probability of extracting ***k*** green marbles?



Order 1 0.6*0.6*0.4

Order 2 0.6*0.4*0.6

Order 3 0.4*0.6*0.6

Note that all order have the same proportions of green, so they have all the same probability!

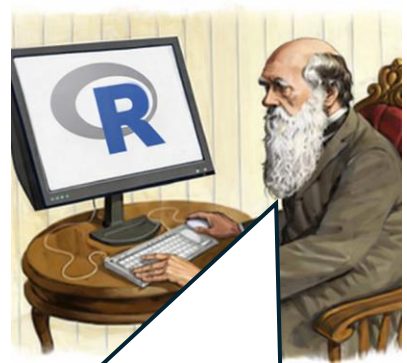
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

Potential order of sampled marbles

Probability of each order

$$Pr(2)=3*0.6^2\ 0.4=0.432$$

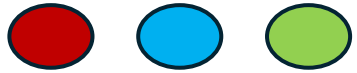
Our first basic “evolutionary model”



- Variability among individuals
- Differential contribution to successive generations
- Mechanism of inheritance

Let’s put together our toy model

```
N=100
z=rnorm(N)
```

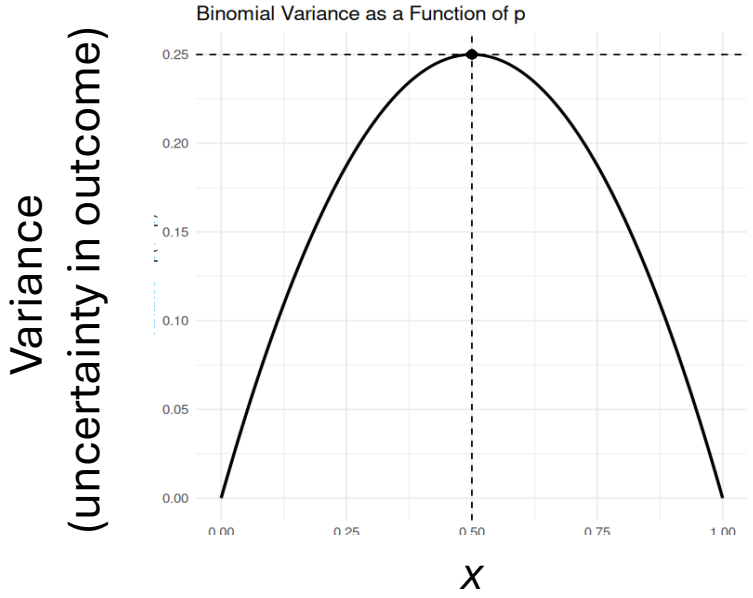


```
survivors=z[z>median(z)]
```

```
z=sample(survivors, N, replace = TRUE)
```

The general structure of 2-players' games

	A	B
A	a	b
B	c	d



The variance of a Binomial Distribution is **~p(1-p)**.

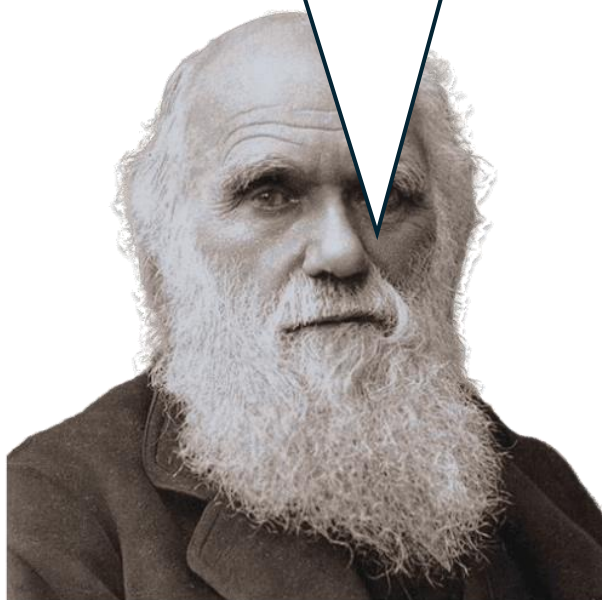
Its mean is simply **~p.**

Do you recognize this?

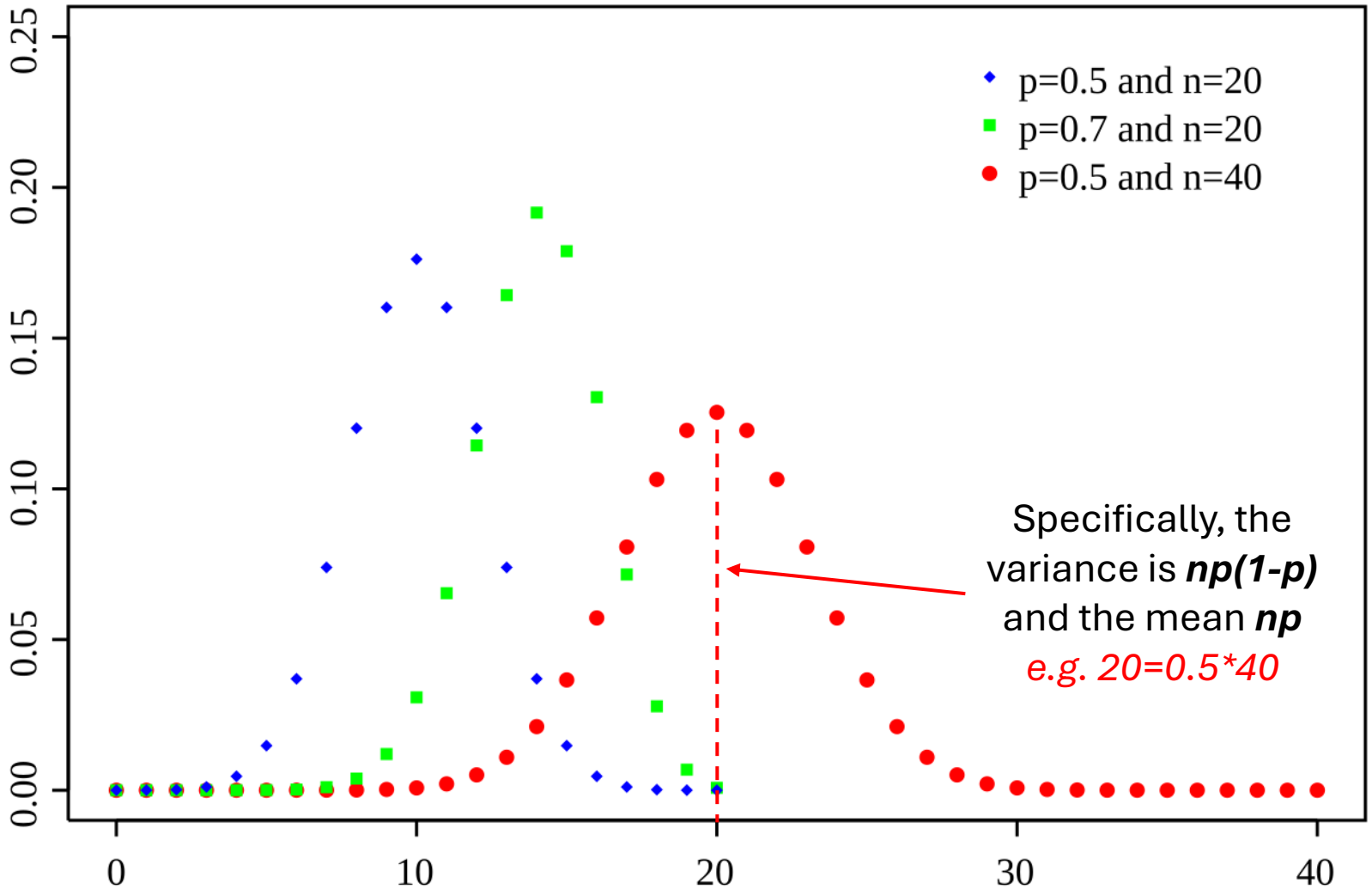
Payoffs

$$\dot{x} = x(1-x)[(a-b-c+d)x+b-d].$$

Replicator's equation



The binomial distribution



Specifically, the variance is ***np(1-p)*** and the mean ***np***

*e.g. 20=0.5*40*

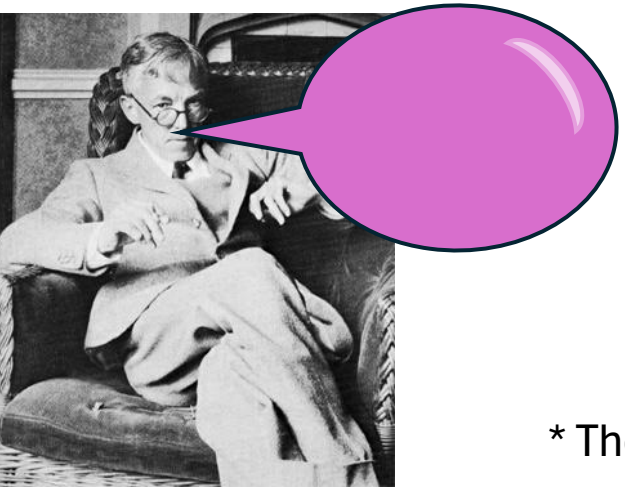
How are chewingum distributed over tiles?

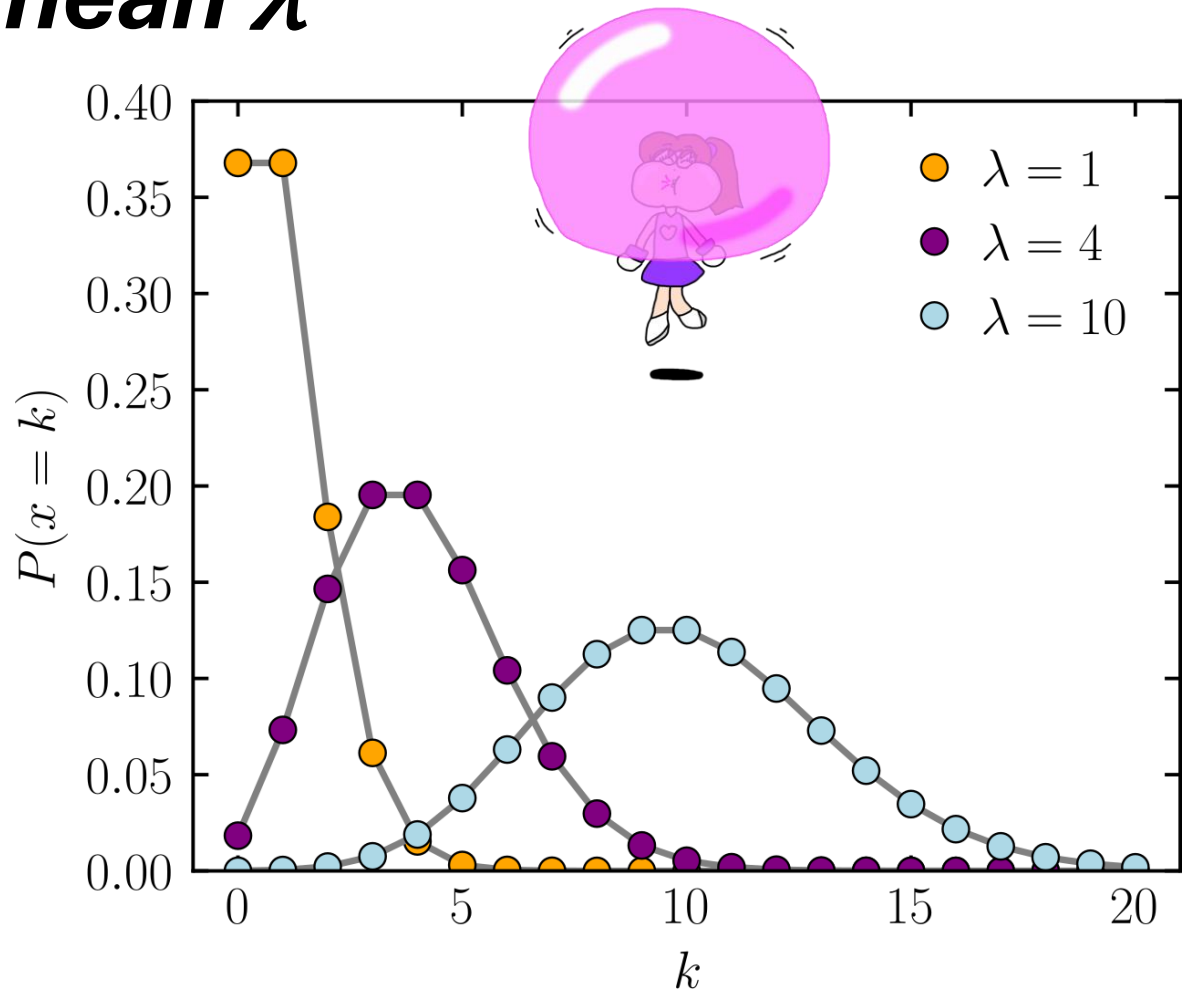
How many chewingums on each tile?

The Poisson distribution describes the probability to **count *k*** successes in an interval given a process with ***(constant)* mean λ***

$$\Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

e.g. probability of k chewingums in a tile, given an average of λ chewingums per tile



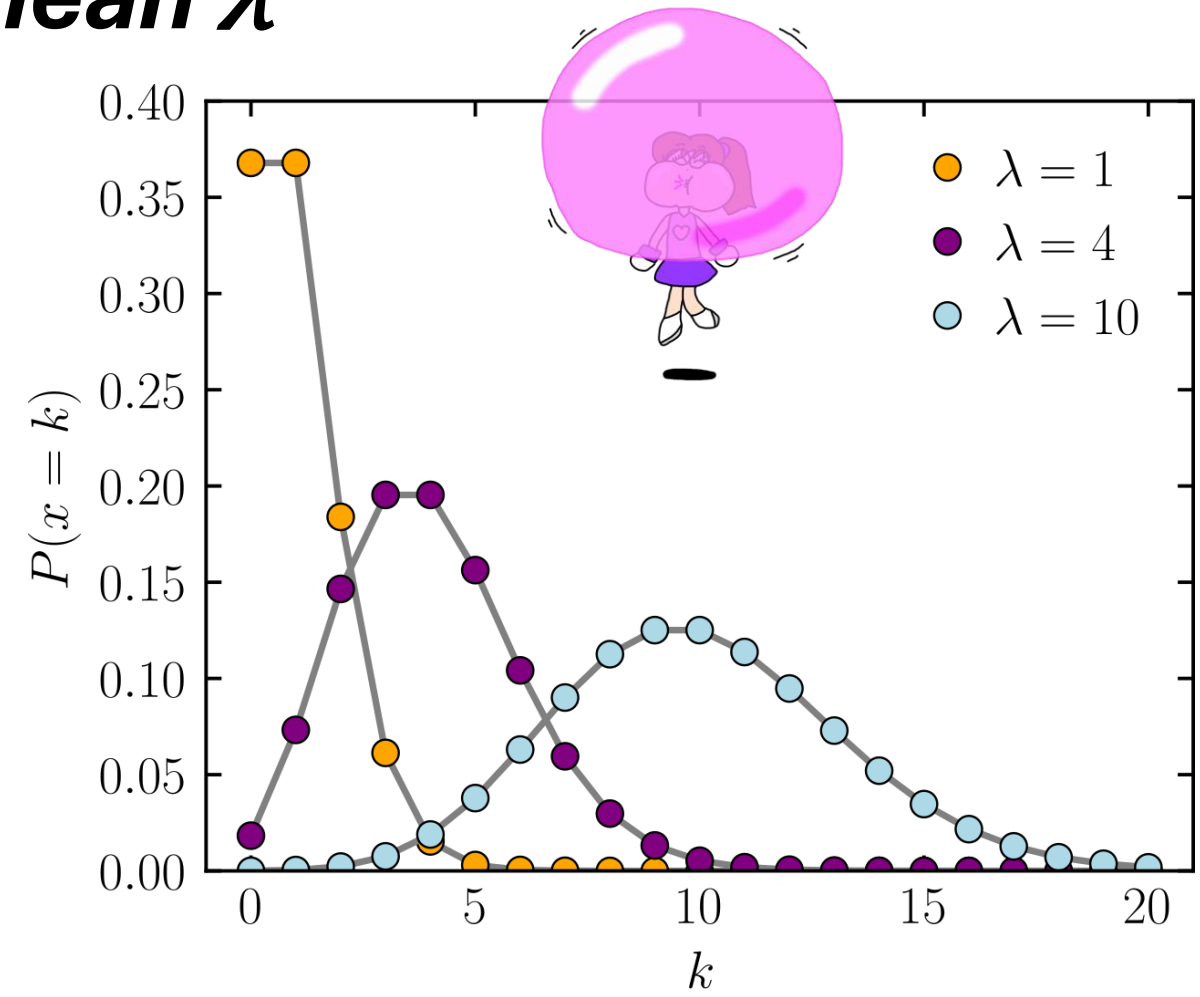


* The key assumption of the Poisson is that the rate at which the process occurs (1/ λ) is constant.

The Poisson distribution describes the probability to **count *k*** successes in an interval given a process with ***(constant) mean λ***

$$\Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

e.g. probability of k chewingums in a tile, given an average of λ chewingums per tile

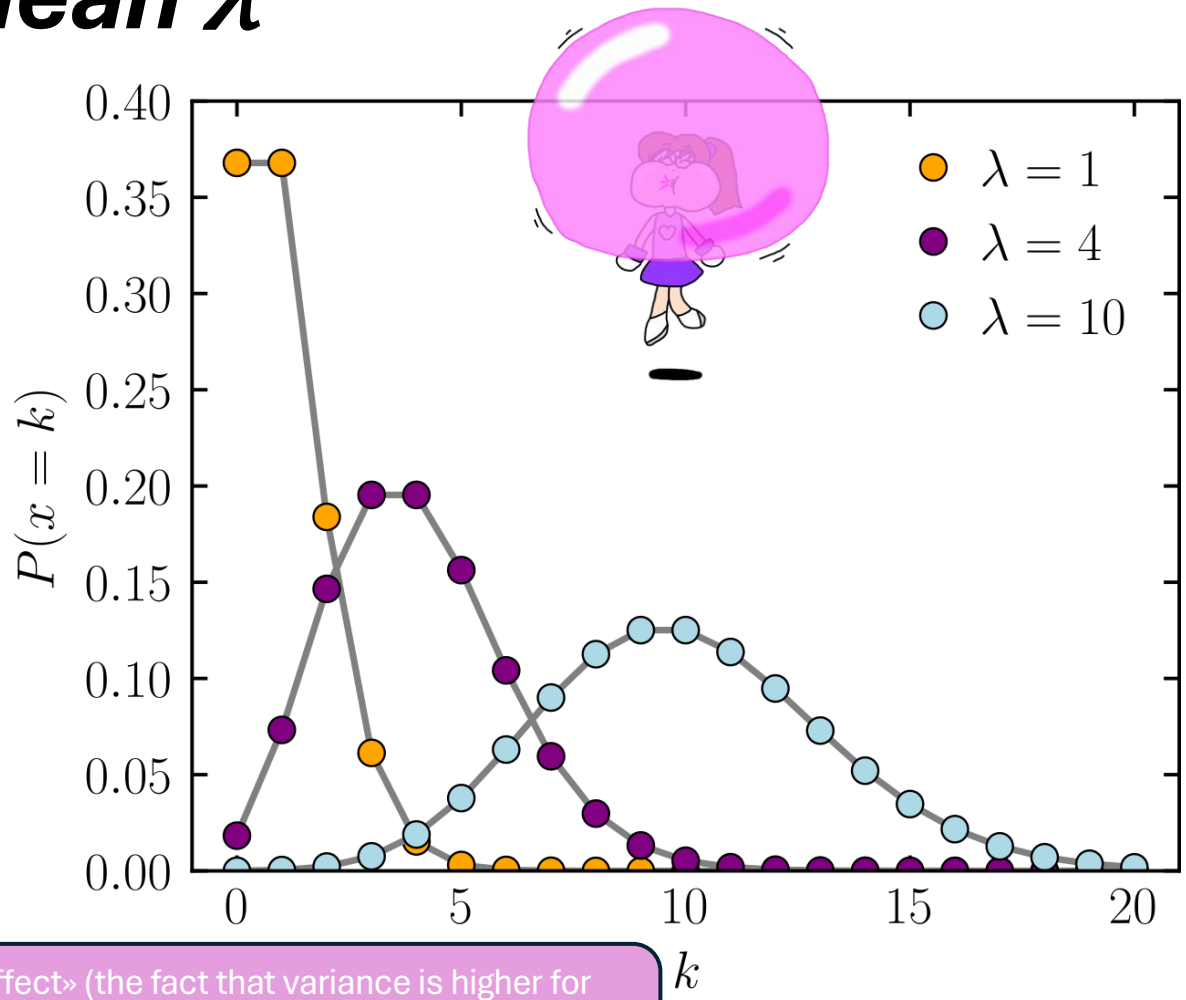


The Poisson distribution describes the probability to **count *k*** successes in an interval given a process with ***(constant) mean λ***

$$\Pr(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

e.g. probability of k chewingums in a tile, given an average of λ chewingums per tile

Awesome distribution because both mean and standard deviation are just λ

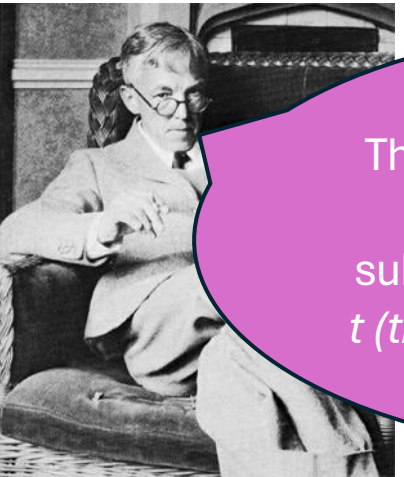


This naturally captures the «variance around the mean effect» (the fact that variance is higher for higher mean, not specified independently as for a Gaussian). Note that for low λ it behaves like an «exponential decay», while for higher λ it is bell-shaped

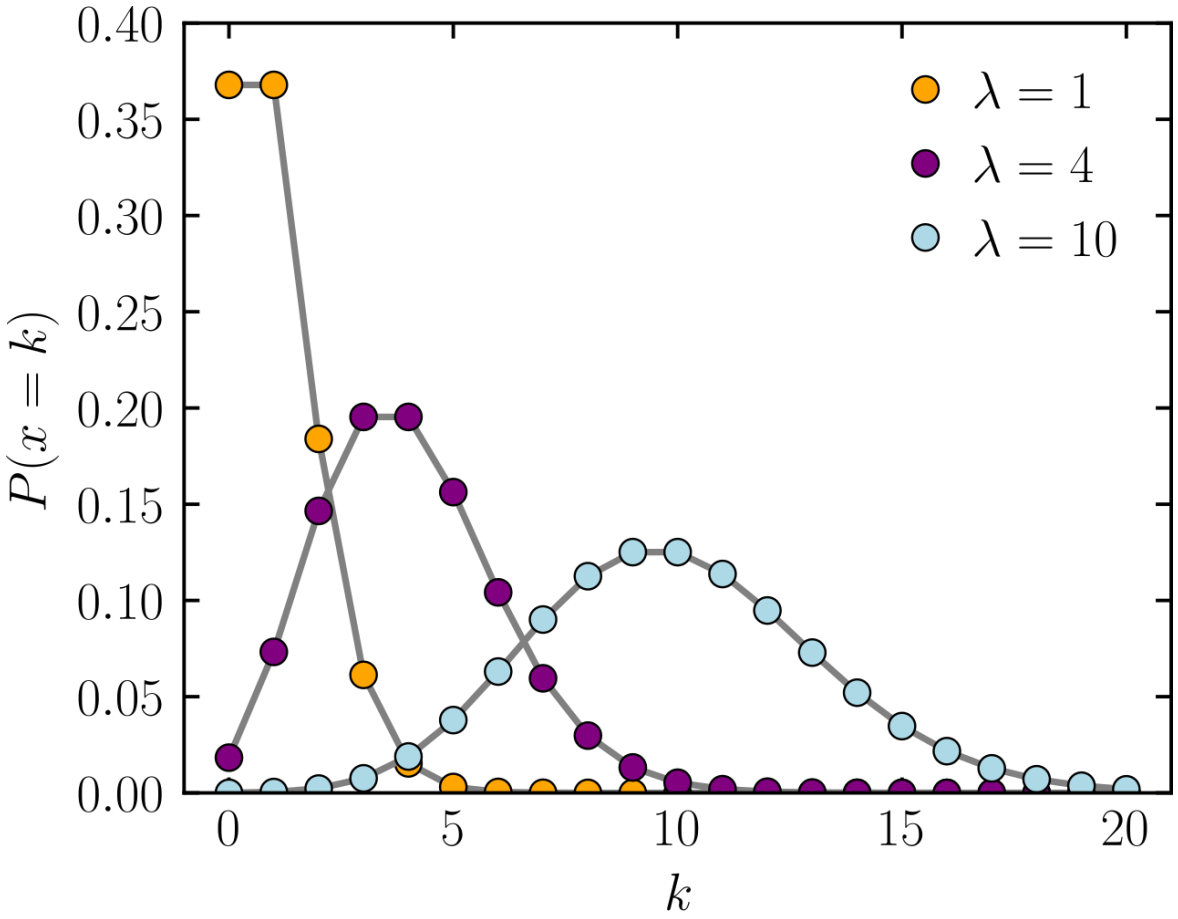
Alternative formulation: The Poisson distribution describes the probability to **count *k*** successes in an interval of length ***t*** given a process with ***constant rate r=1/λ***

$$P(k \text{ events in interval } t) = \frac{(rt)^k e^{-rt}}{k!}$$

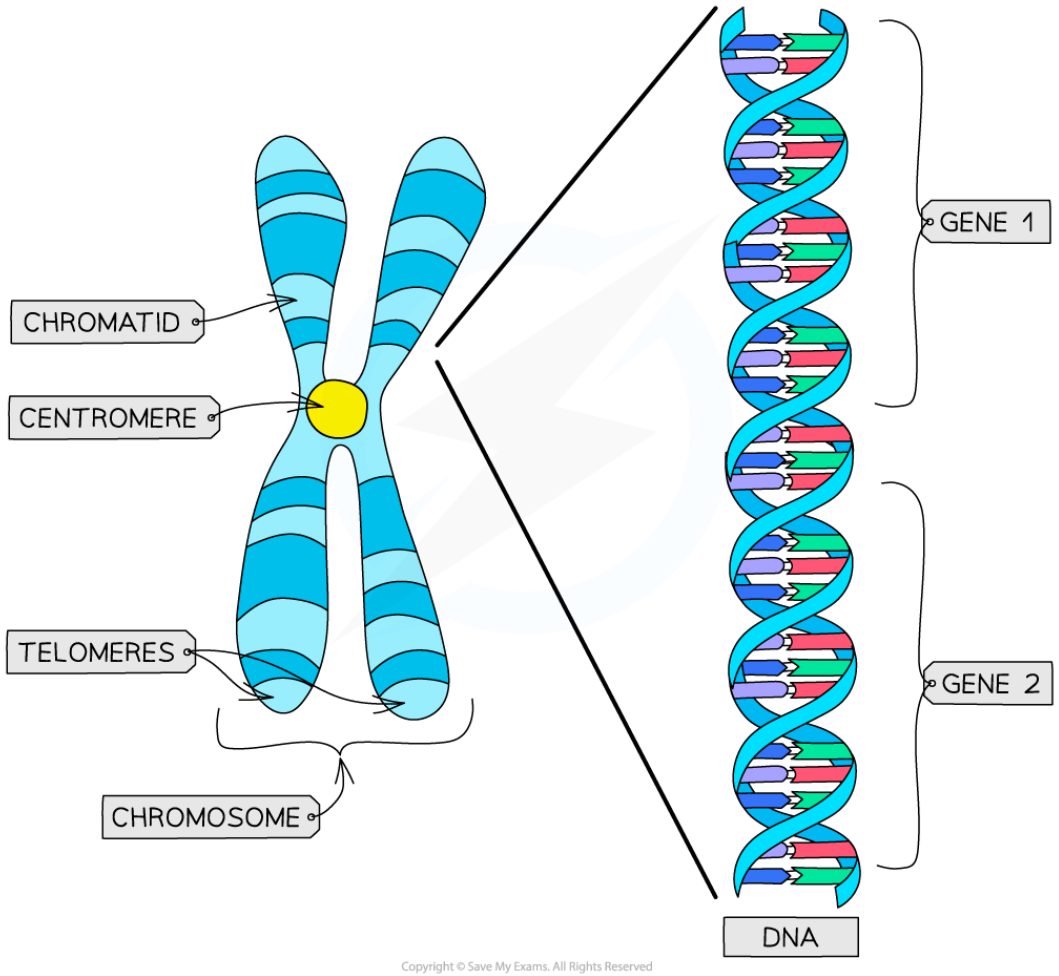
$$\lambda = rt$$



This formulation is analogous to the previous one, only substitute ***rt*** to ***λ***. *Useful when t (time or space) is continuous.*



Some examples of Poisson distribution

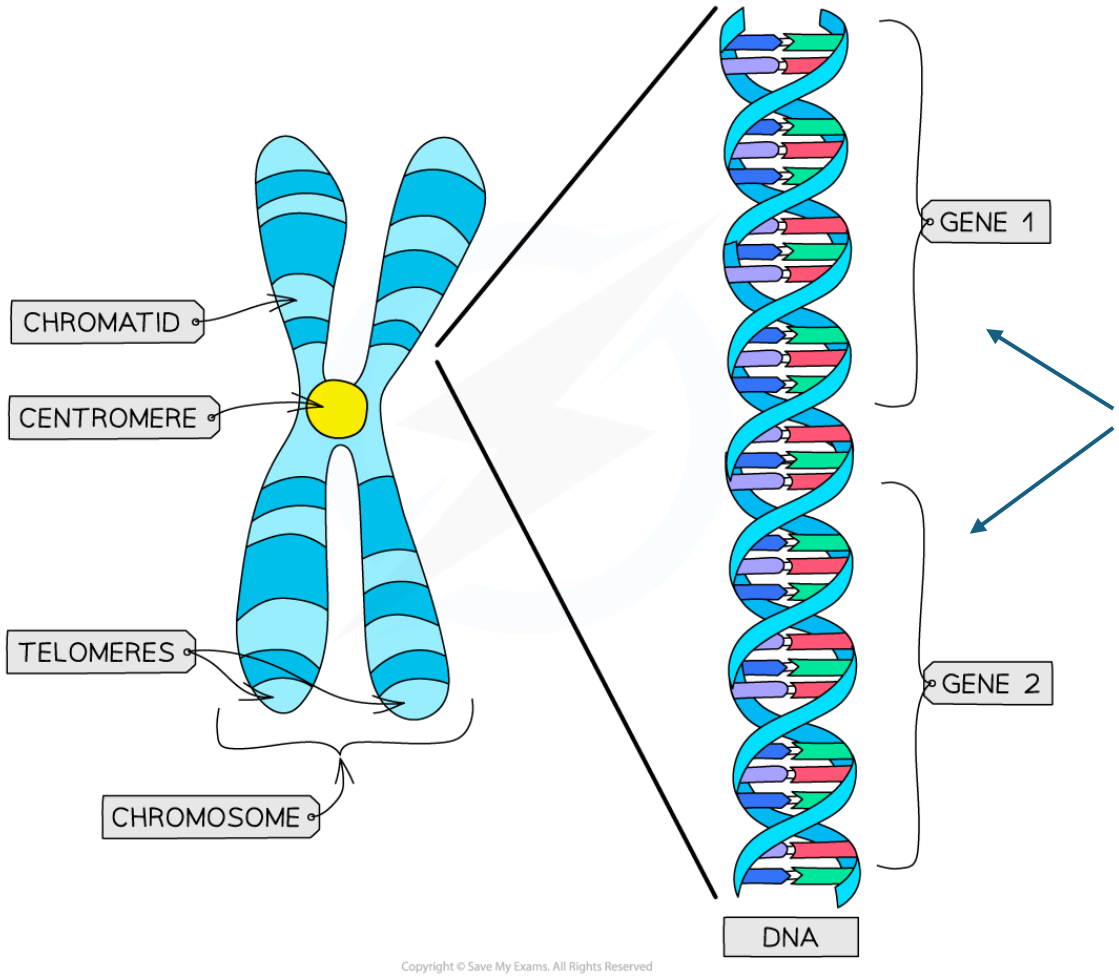


Our chewingums ↓

Our tiles ↓

Number of mutations occurring in a given genomic regions

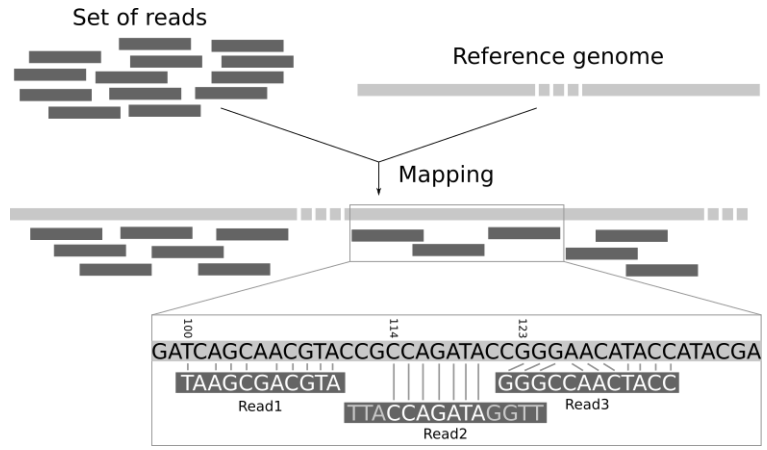
Some examples of Poisson distribution



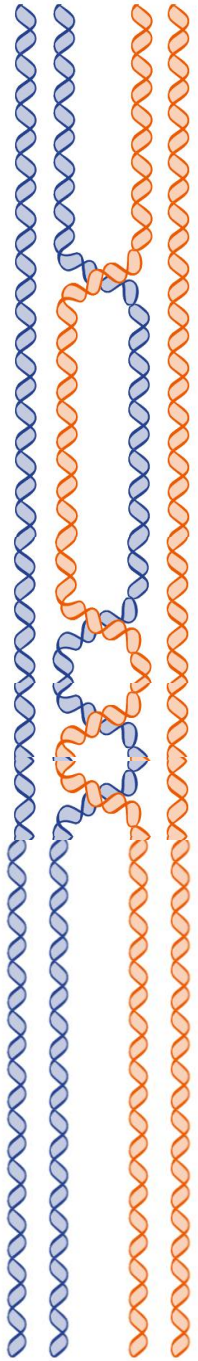
Our chewingums → Number of mutations occurring in a given genomic regions

Our tiles →

Number of sequencing reads (coverage) mapping along the genome

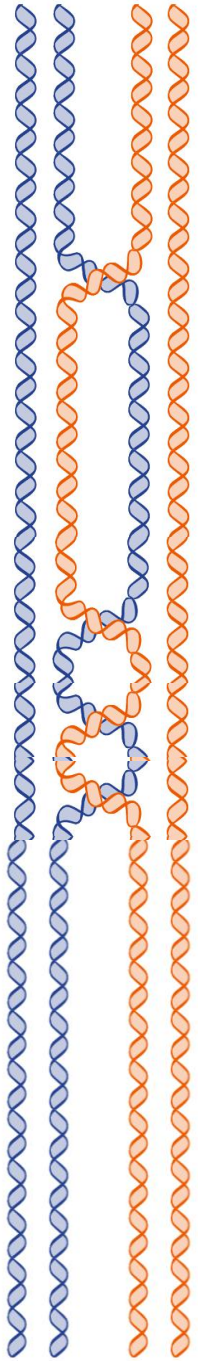


Some examples of exponential distribution



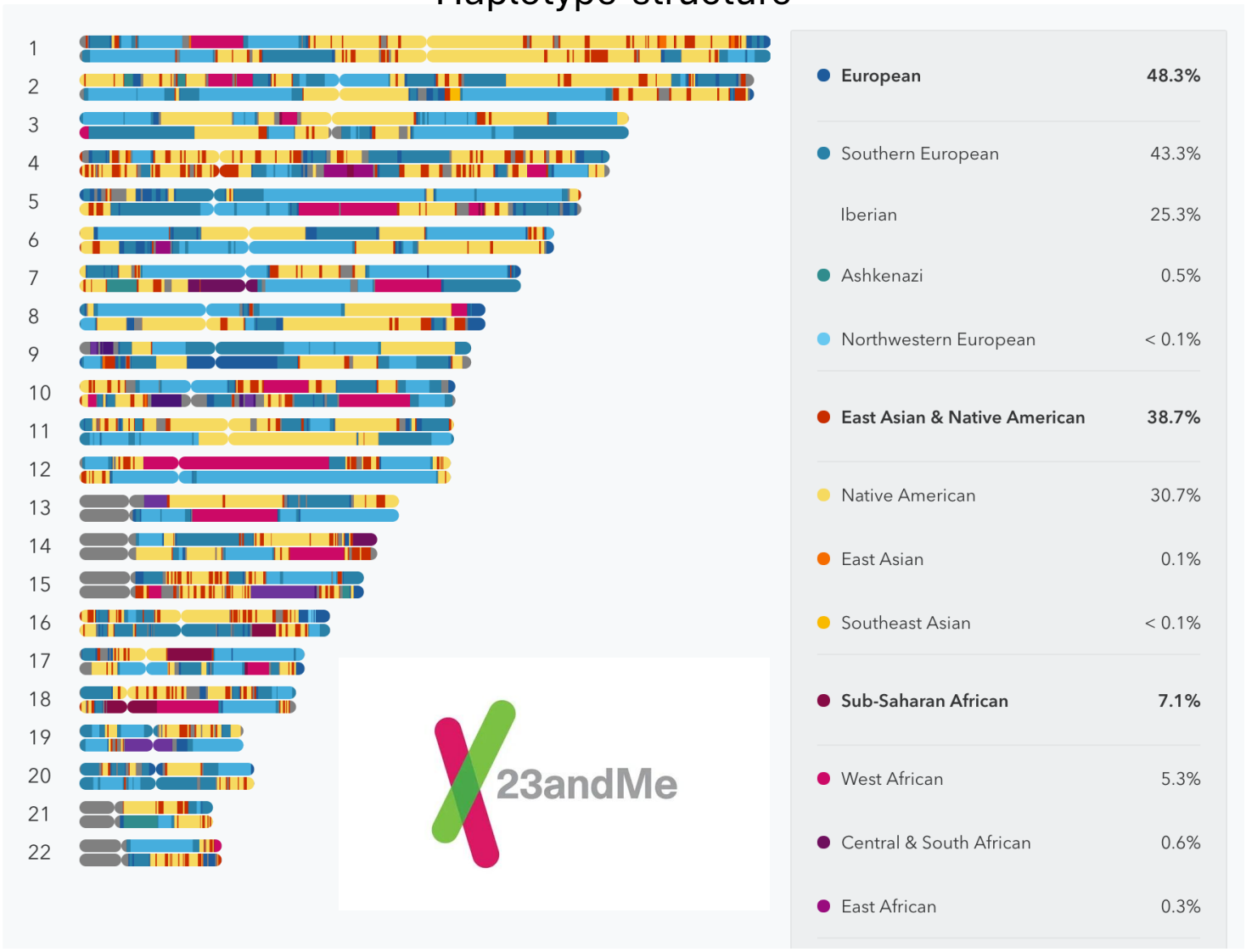
Distance between two recombination events in base-pairs

Some examples of exponential distribution



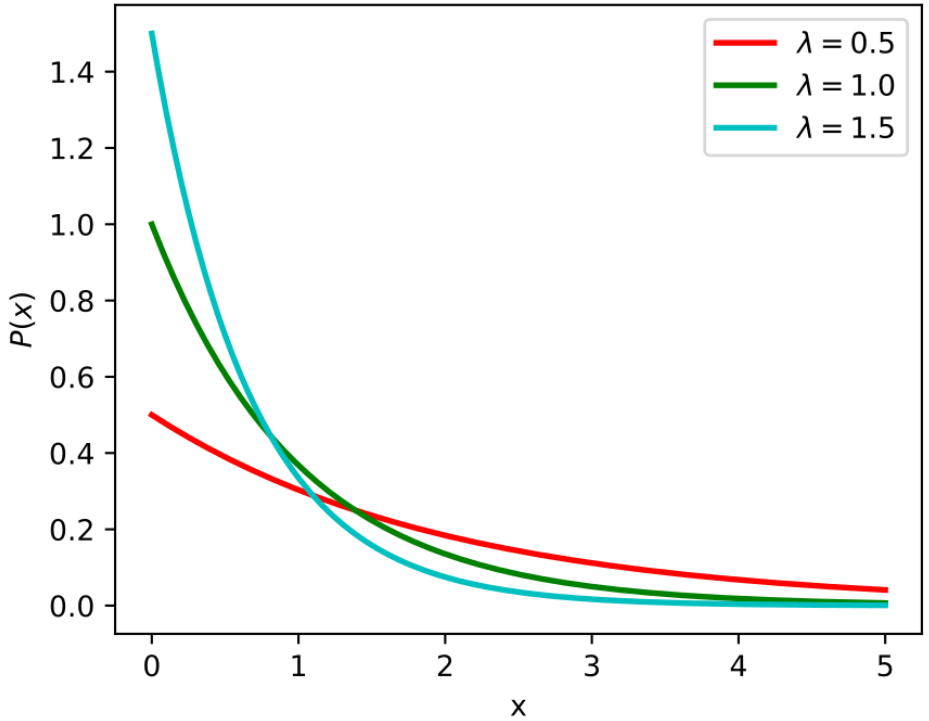
Distance between two recombination events in base-pairs

Haplotype-structure

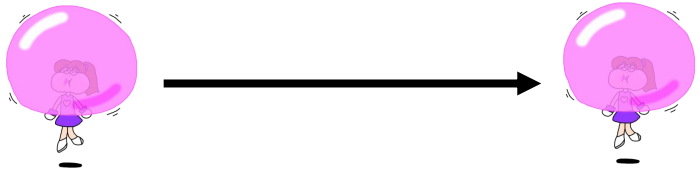


The exponential distribution describes the probability that a given amount of **time**/space **x** occurs before an event occurring with **constant rate λ**

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$



It is the «distance between chewingums» (in linear space)*!!!



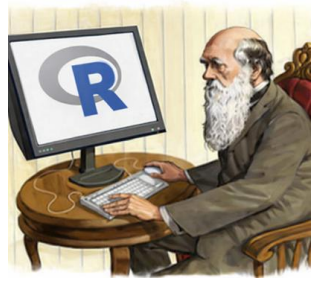
* or the time passing before somebody walking on the street spits another one on the floor (assuming people passing and spitting at a constant rate!)

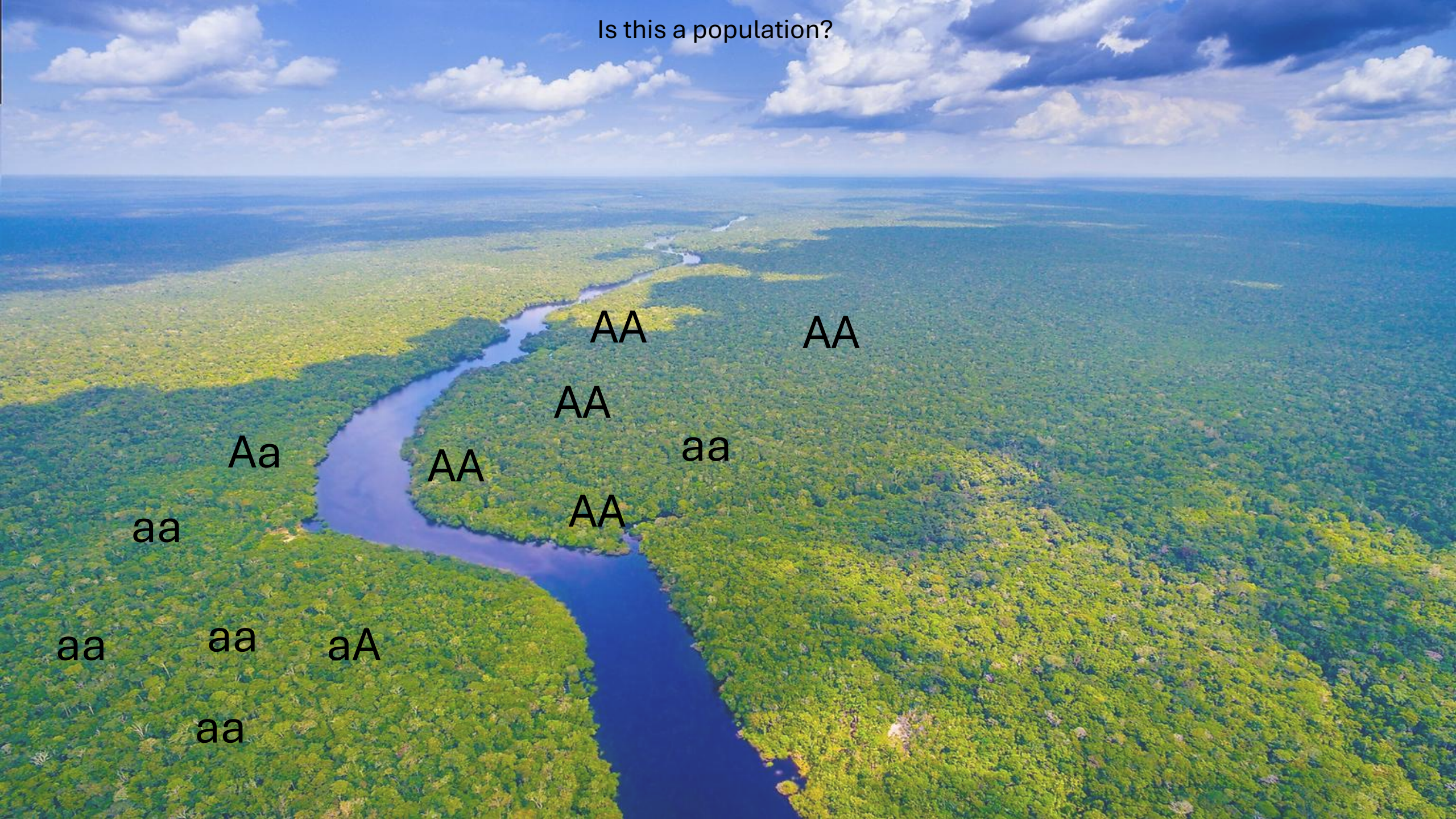
Appendix: clarification and questions asked by students answered on the black board (so not visible on the recordings)

- **Can we use these distributions to estimate deviations from the HWE similarly to the χ^2?**
- Yes we can! For example, let's assume we have two populations as in the forest example we had before. Rather than using «statistical Golems» like the χ^2 we can use probability theory to build a small model of population structure with two populations, and one with just one population and check which «explains better the data».
- For instance, let's imagine two subpopulations with 10*A* and 2*a*, and the other with 10*a* and 2° *(see next slide)*. we can first think of the whole forest as two population separated by the river. In that case we can calculate the maximum likelihood for each of the two populations, which is Pr_{pop1}=Binomial(k=2,n=12,p_1=2/12)=0.296 and Pr_{pop2}=Binomial(k=10,n=12,p_1=10/12)=0.296. Since these are independent events, we can think of our two population model as a «complex» model with two parameters (two degrees of freedom, p_1 and p_2) and likelihood Pr_{two_pops}=Pr_{pop1}* Pr_{pop2}=0.296*0.296= 0.0877.
- We can also imagine simpler population model with just 1 panmictic population. In this case the maximum likelihood model is Pr_{one_pop}=Binomial(k=2,n=12,p_1=12/24=0.5)*Binomial(k=10,n=12,p=12/24=0.5)=0.0003, but this simple model has just one free parameter – p, rather than p1 and p2.
- Now what do we do? More complex models have always «higher likelihood» – since they have more degrees of freedom. But the simpler is more parsimonious. We can simply compare them with a likelihood ratio test. Remarkably, this concides with a **χ^2** distribution with degrees of freedom equal to the difference in free parameter between the complex and the simple model (in our case 1). In R, in our case this can be done:

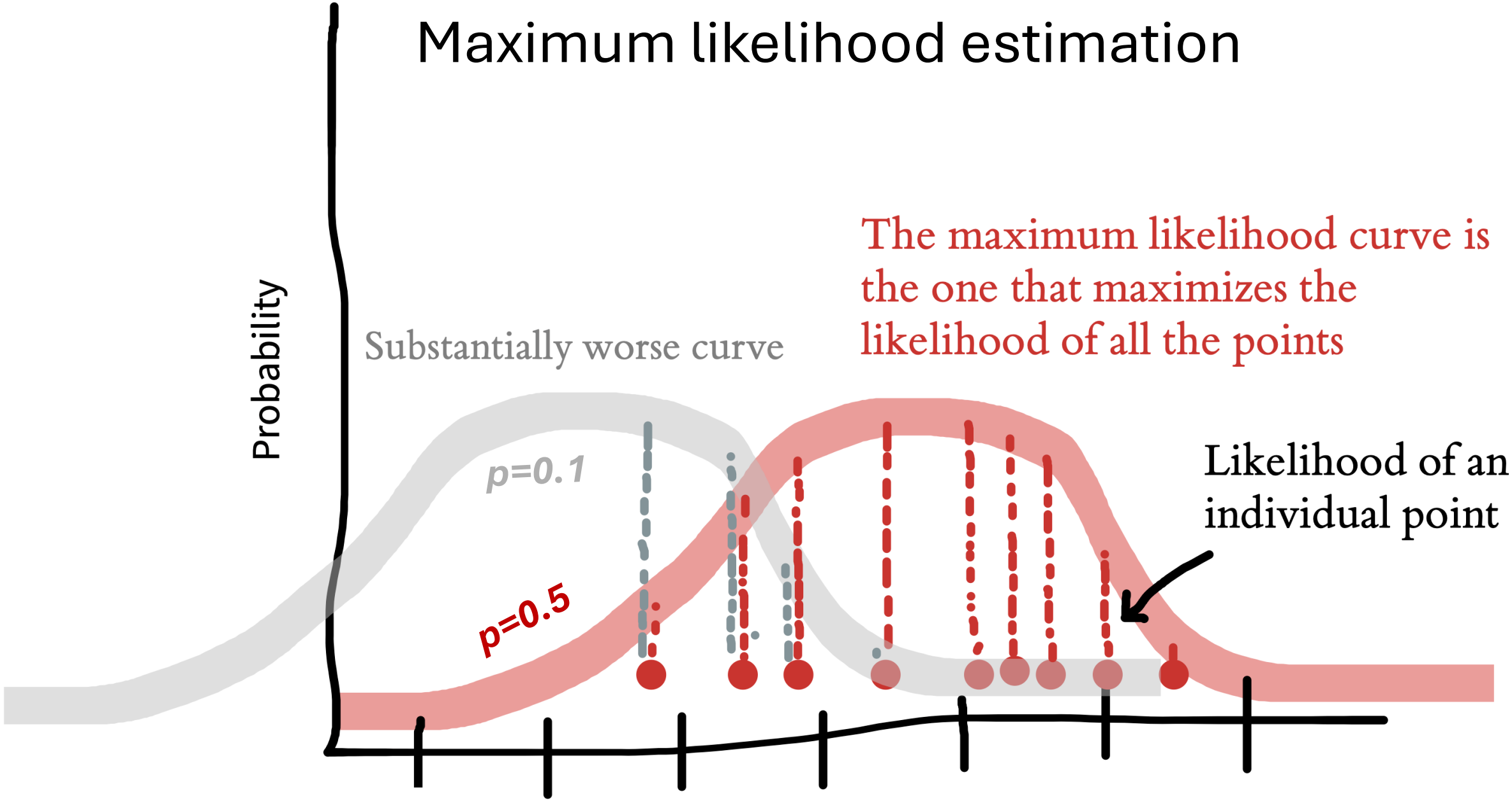
```
logL0 <- log(dbinom(2,12,1/2))+log(dbinom(10,12,1/2)) # null model
logL1 <- log(dbinom(10,12,5/6))+log(dbinom(2,12,1/6)) # alternative model
#note that the logs are used to sum the log-likelihoods rather than multiplying
them
k0 <- 1 #number of free parameters
k1 <- 2
LR <- 2 * (logL1 - logL0)
df <- k1 - k0
pval <- pchisq(LR, df = df, lower.tail = FALSE)
```

- pvalue is 0.0006440557




Is this a population?
AA
AA
AA
aa
AA
AA
Aa
aa
aa
aa
aA
aa

Maximum likelihood estimation



The binomial distribution