More Computational Statistics and Data Visualisation

Key concepts

- Statistical inference is the process of drawing conclusions about an underlying population based on a sample or subset of the data
- Hypothesis testing is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis
- Exploratory data analysis is an approach to analysing datasets to summarize their main characteristics, often through visual methods

VISUALISE

HYPOTHESISE

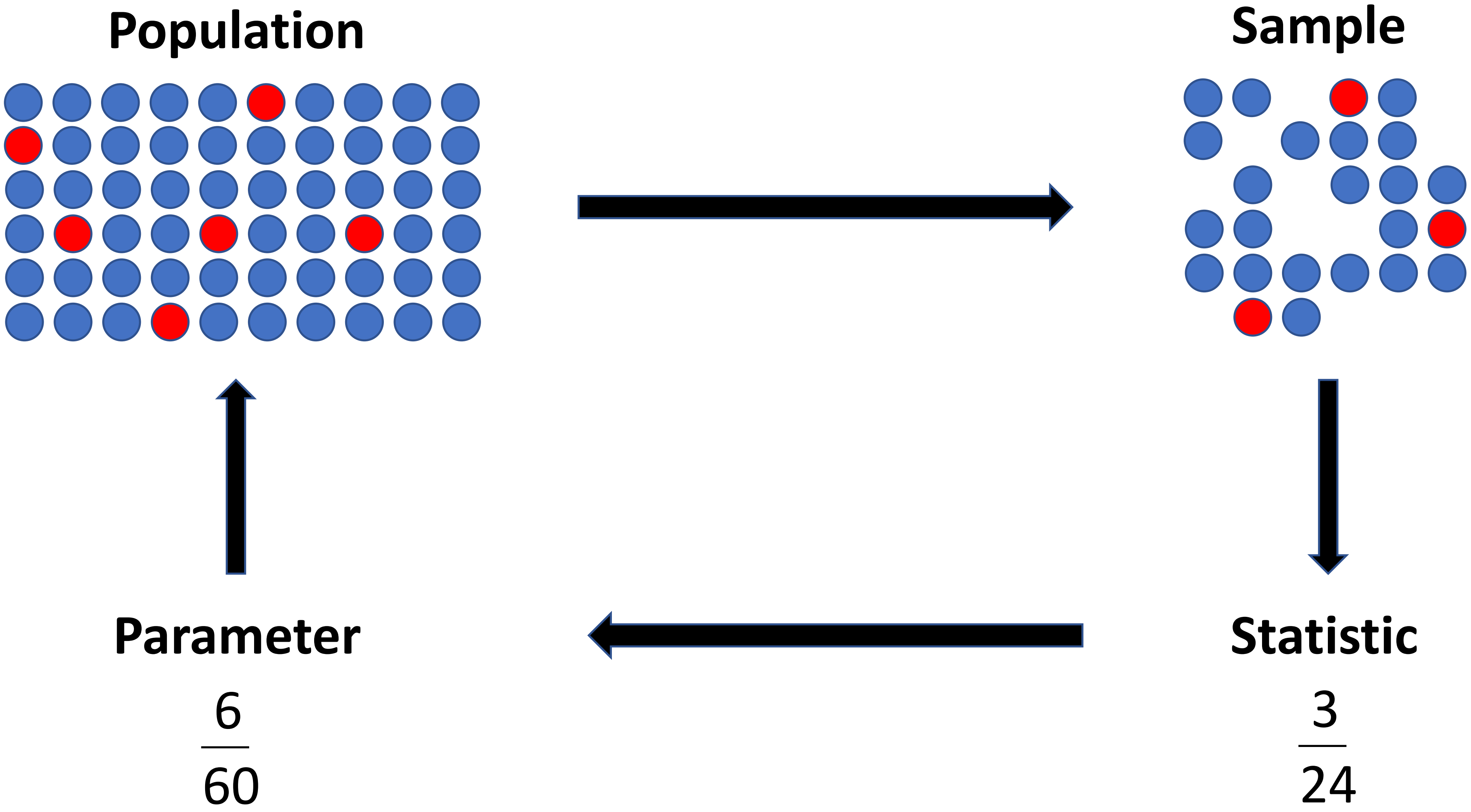
SUMMARISE

ANALYSE

Key concepts

- Statistical inference is the process of drawing conclusions about an underlying population based on a sample or subset of the data

Statistical inference



Population

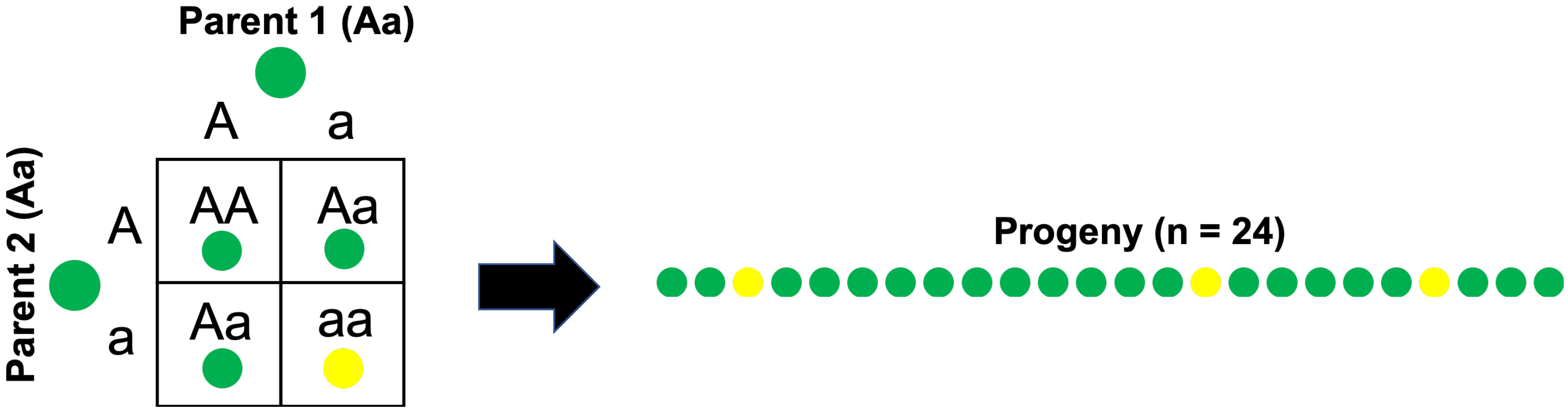
Sample

Parameter

$$\frac{6}{60}$$

Statistic

$$\frac{3}{24}$$



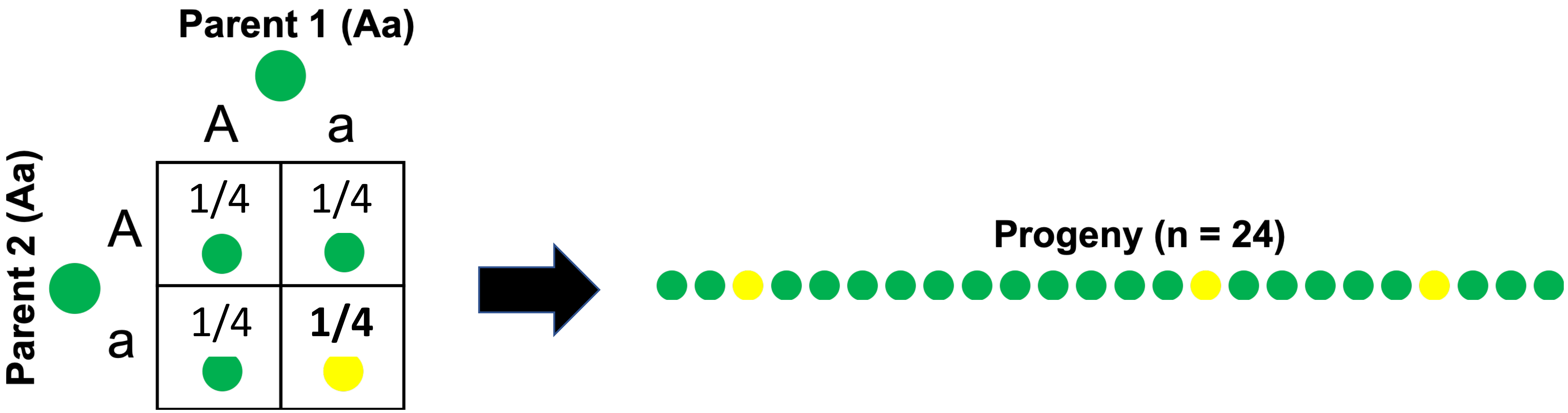
Parent 1 (Aa)

Parent 2 (Aa)

	A	a
A	AA 🟢	Aa 🟢
a	Aa 🟢	aa 🟡

Progeny (n = 24)

🟢🟢🟡🟢🟢🟢🟢🟢🟢🟢🟢🟢🟢🟢🟡🟢🟢🟢🟢🟢🟡🟢🟢🟢



Parent 1 (Aa)

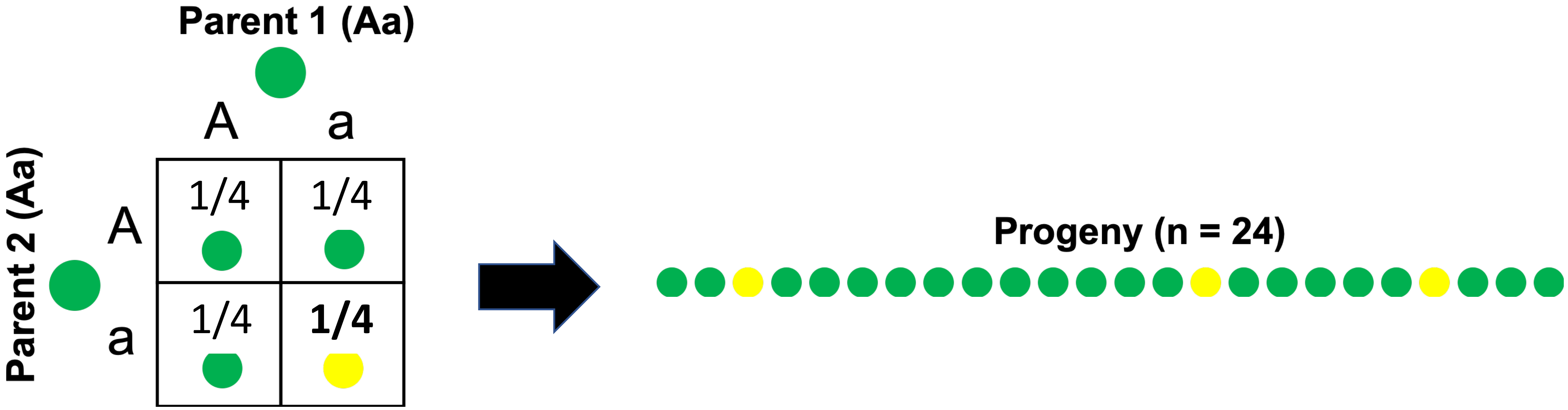
Parent 2 (Aa)

	A	a
A	1/4	1/4
a	1/4	**1/4**

Progeny (n = 24)

Null hypothesis

H_0: The proportion of yellow progeny is 1/4



Parent 1 (Aa)

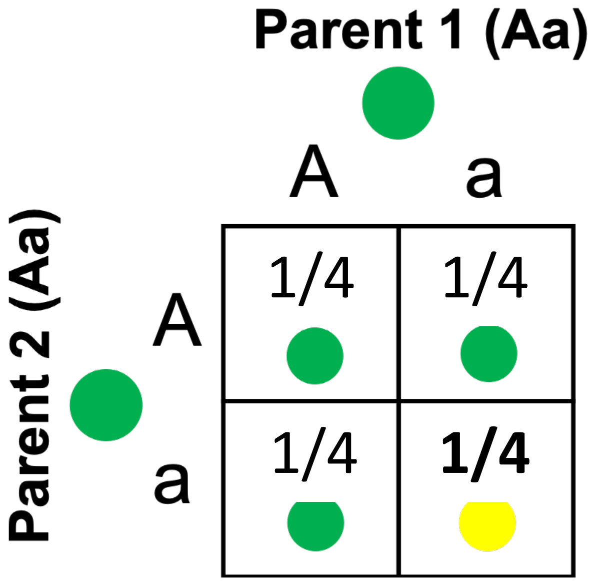
Parent 2 (Aa)

	A	a
A	1/4	1/4
a	1/4	**1/4**

Progeny (n = 24)

Null hypothesis

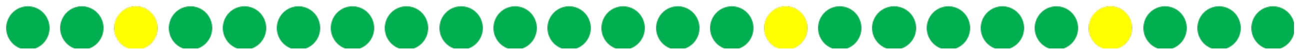
$$H_0 : p_y = \frac{1}{4}$$



Parent 1 (Aa)

	A	a
Parent 2 (Aa) A	1/4	1/4
a	1/4	**1/4**

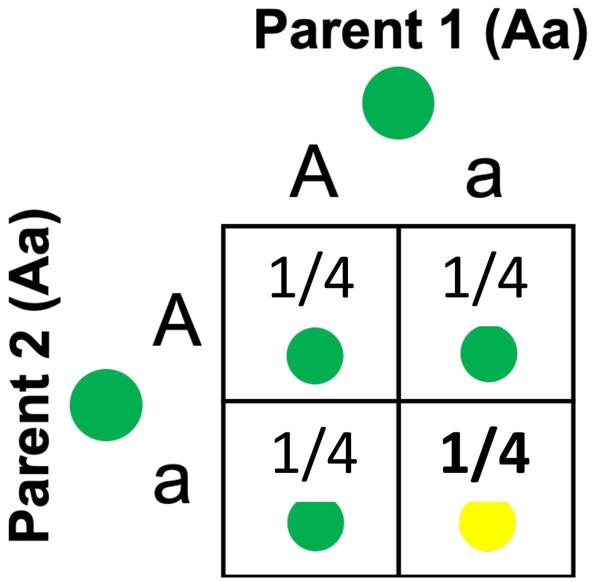
Progeny (n = 24)

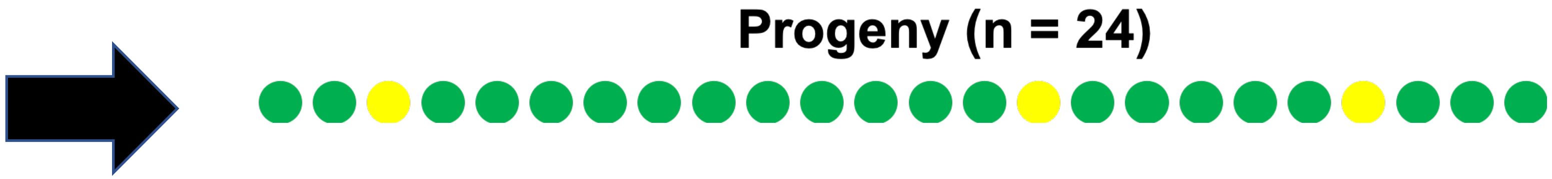


$$\widehat{p_y} = \frac{3}{24}$$

Null hypothesis

$$H_0 \colon p_y = \frac{1}{4}$$



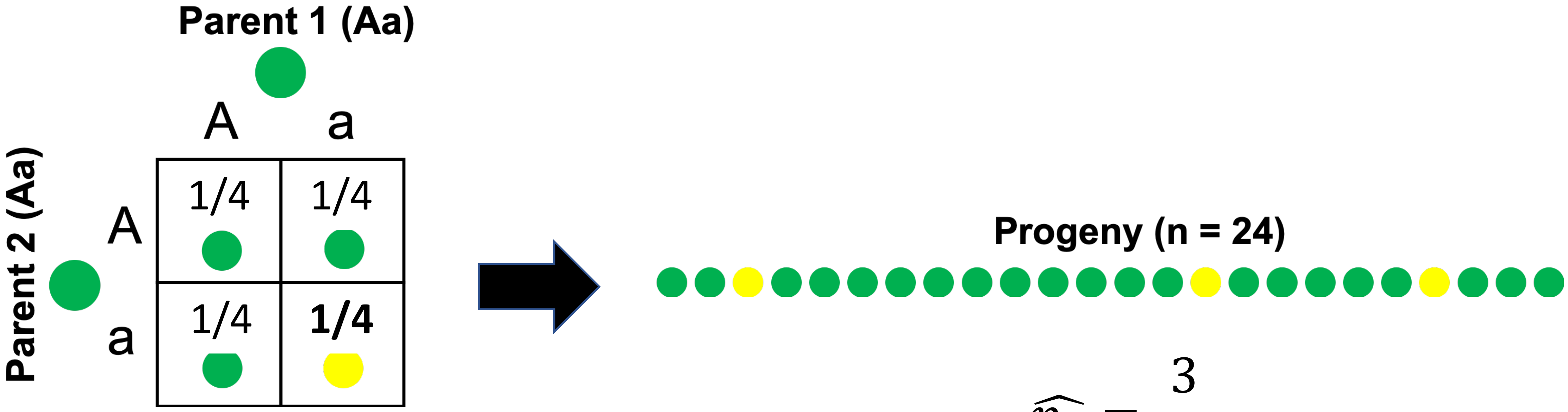


$$\widehat{p_y} = \frac{3}{24}$$

Null hypothesis

$$H_0\!: p_y = \frac{1}{4}$$

How much evidence is this against the null hypothesis?



	A	a
A	1/4 🟢	1/4 🟢
a	1/4 🟢	**1/4** 🟡

🟢🟢🟡🟢🟢🟢🟢🟢🟢🟢🟢🟢🟢🟢🟡🟢🟢🟢🟢🟢🟡🟢🟢🟢

$$\widehat{p_y} = \frac{3}{24}$$

Null hypothesis

$$H_0 : p_y = \frac{1}{4}$$

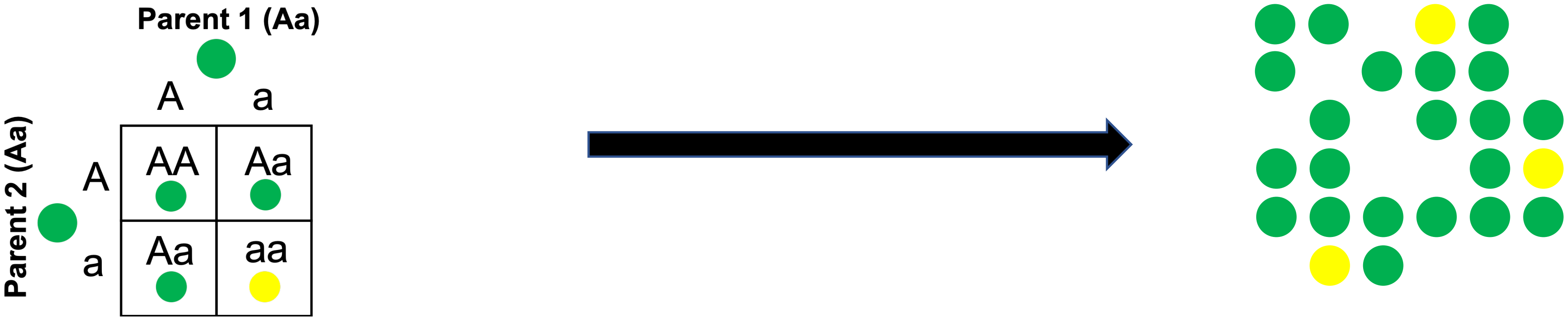
How much evidence is this against the null hypothesis?

What is the distribution of $\widehat{p_y}$ if the null hypothesis is true?

Sampling distribution!

Population

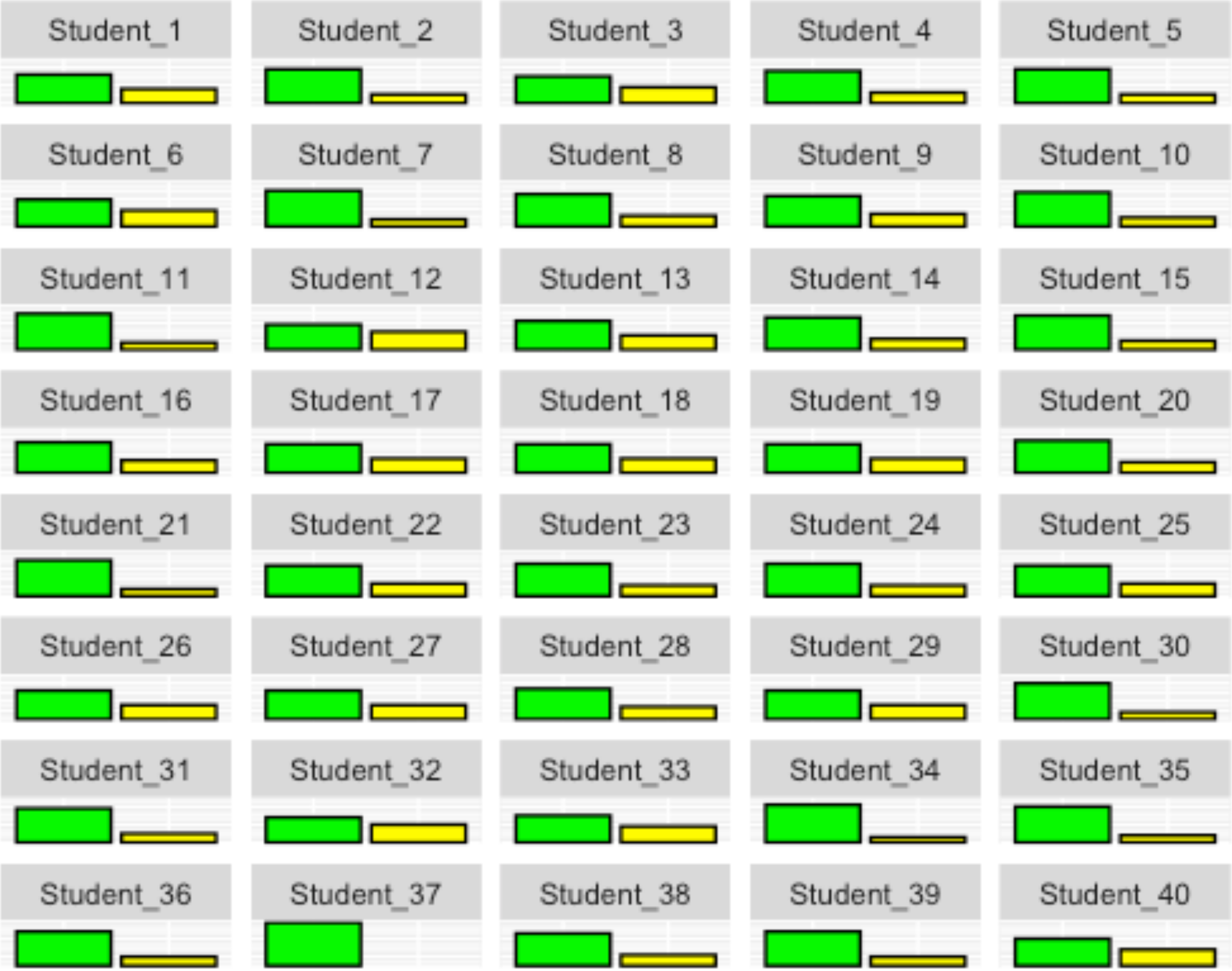
Sample

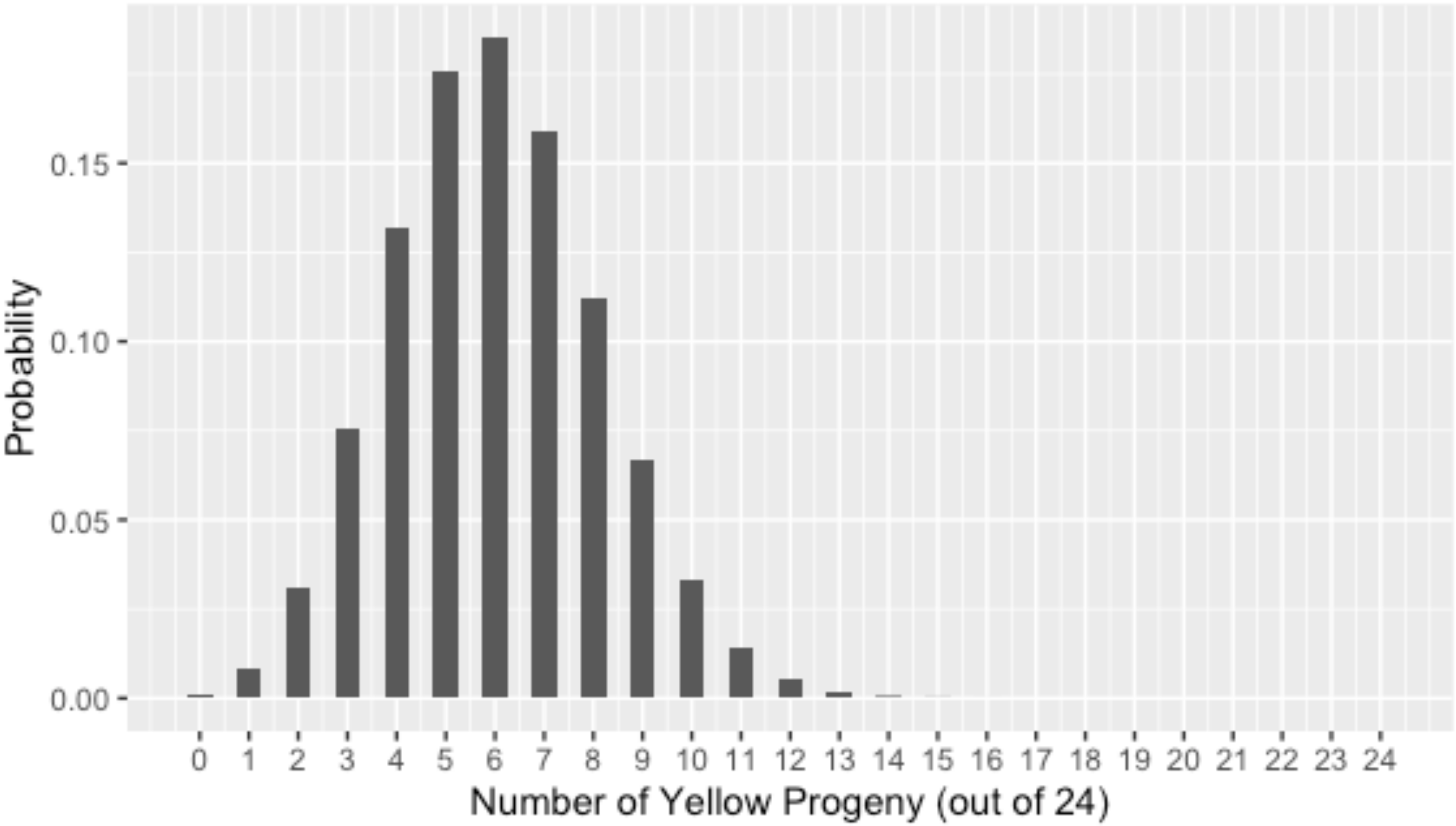


Parent 1 (Aa)

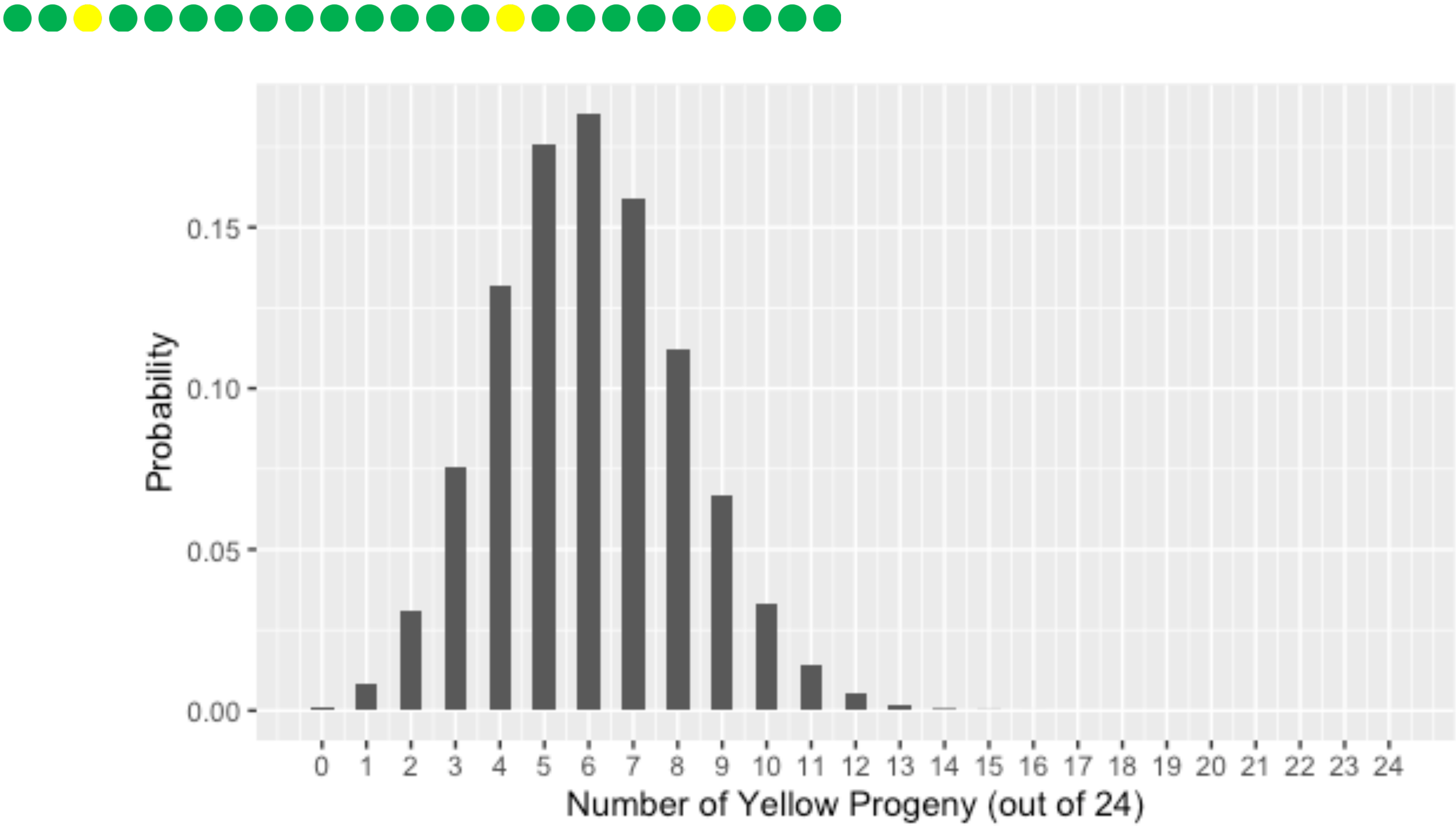
Parent 2 (Aa)

	A	a
A	AA ●	Aa ●
a	Aa ●	aa ●

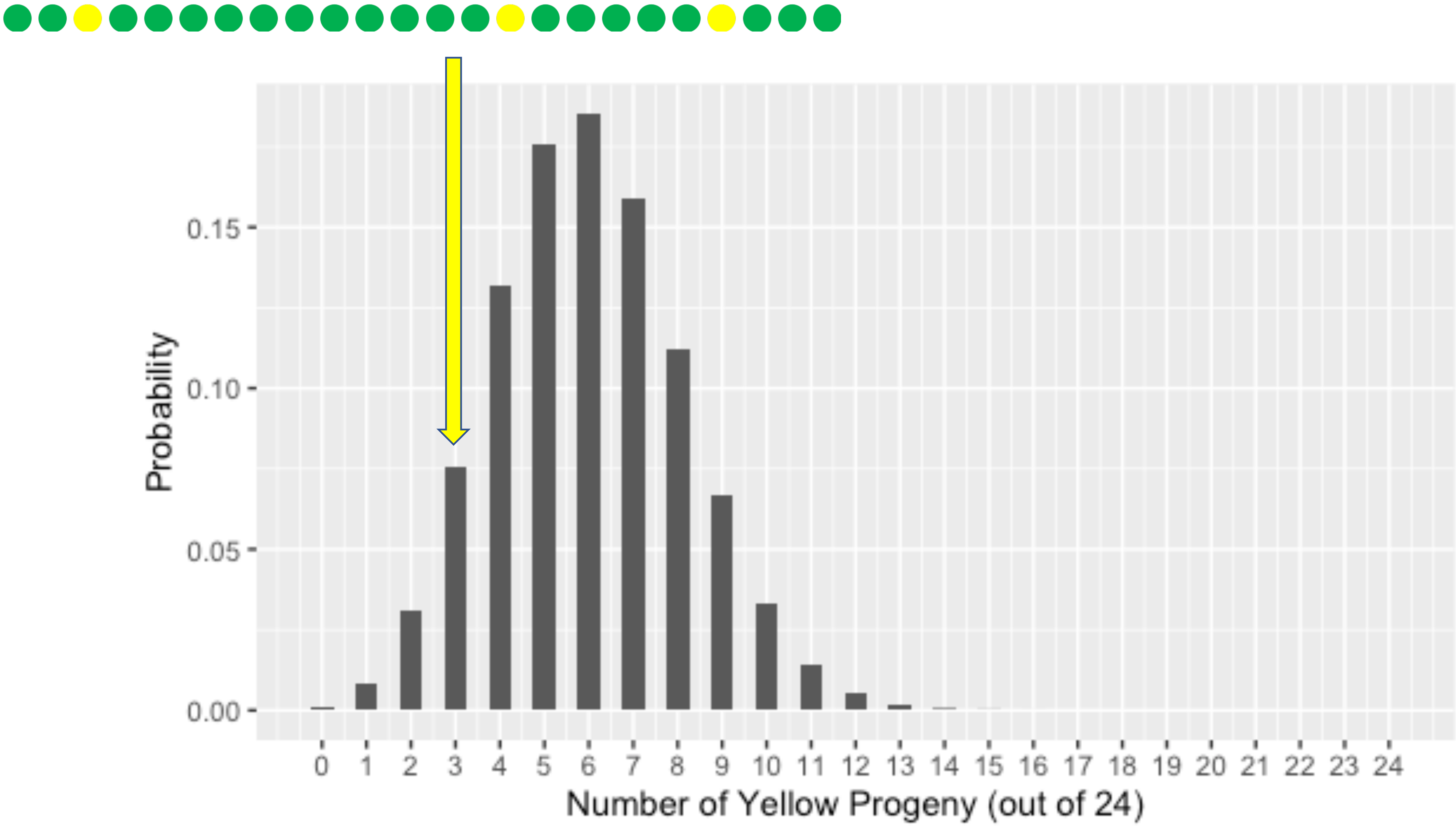

Student_1
Student_2
Student_3
Student_4
Student_5
Student_6
Student_7
Student_8
Student_9
Student_10
Student_11
Student_12
Student_13
Student_14
Student_15
Student_16
Student_17
Student_18
Student_19
Student_20
Student_21
Student_22
Student_23
Student_24
Student_25
Student_26
Student_27
Student_28
Student_29
Student_30
Student_31
Student_32
Student_33
Student_34
Student_35
Student_36
Student_37
Student_38
Student_39
Student_40


Probability
0.00
0.05
0.10
0.15
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
Number of Yellow Progeny (out of 24)

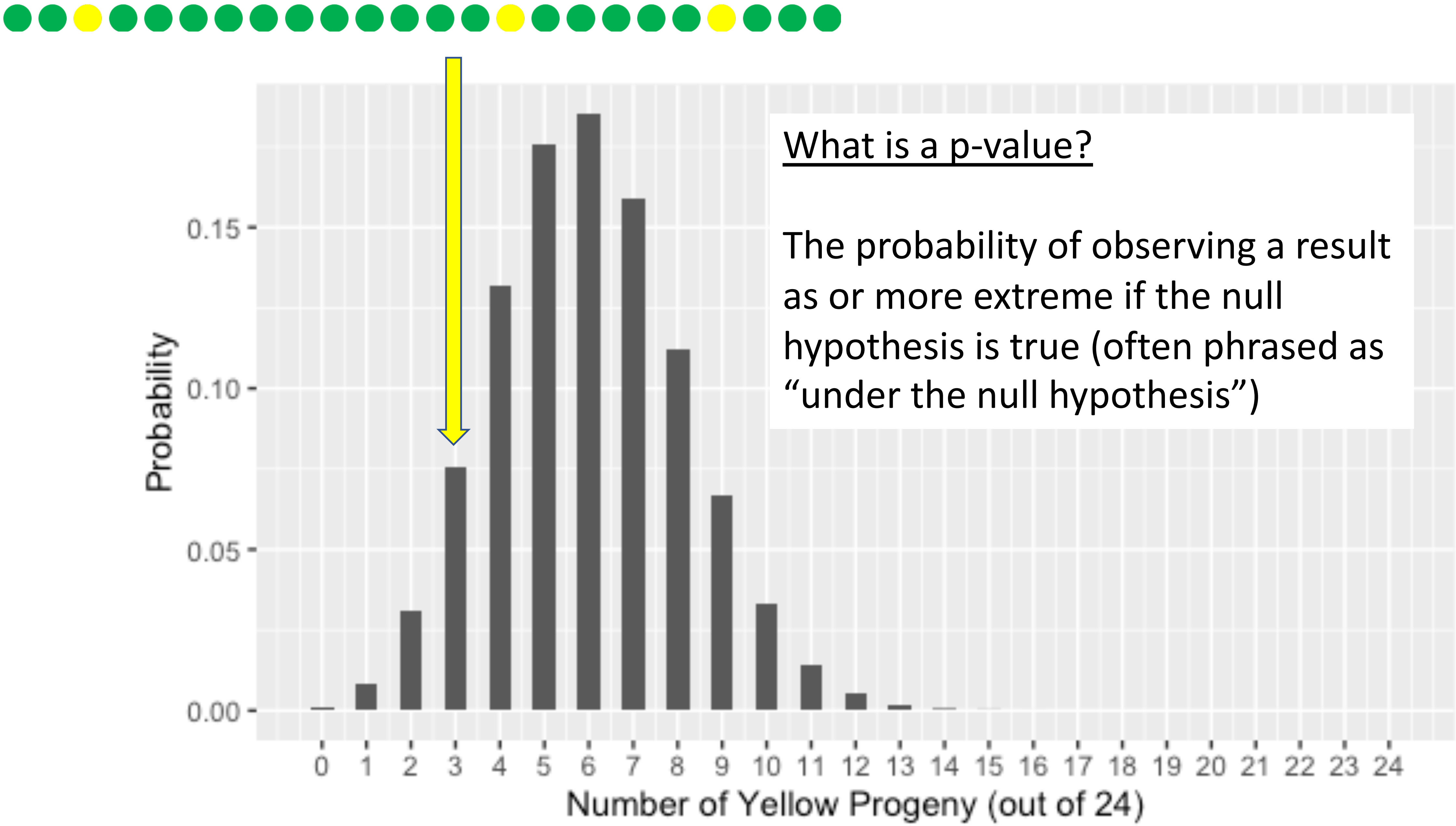
Progeny (n = 24)


Probability
0.00
0.05
0.10
0.15
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
Number of Yellow Progeny (out of 24)

Progeny (n = 24)


Probability
0.00
0.05
0.10
0.15
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
Number of Yellow Progeny (out of 24)

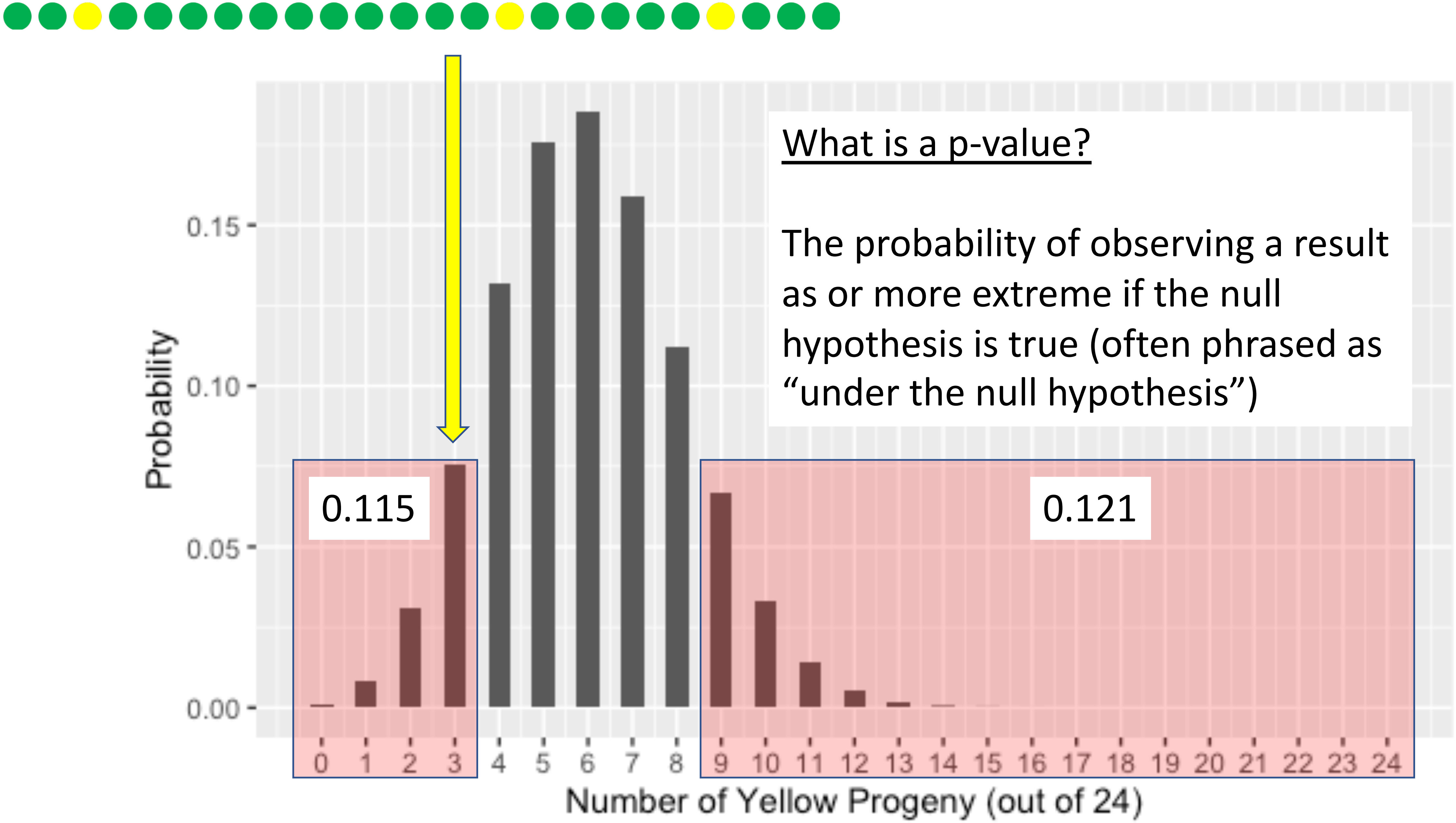
Progeny (n = 24)



What is a p-value?

The probability of observing a result as or more extreme if the null hypothesis is true (often phrased as “under the null hypothesis”)

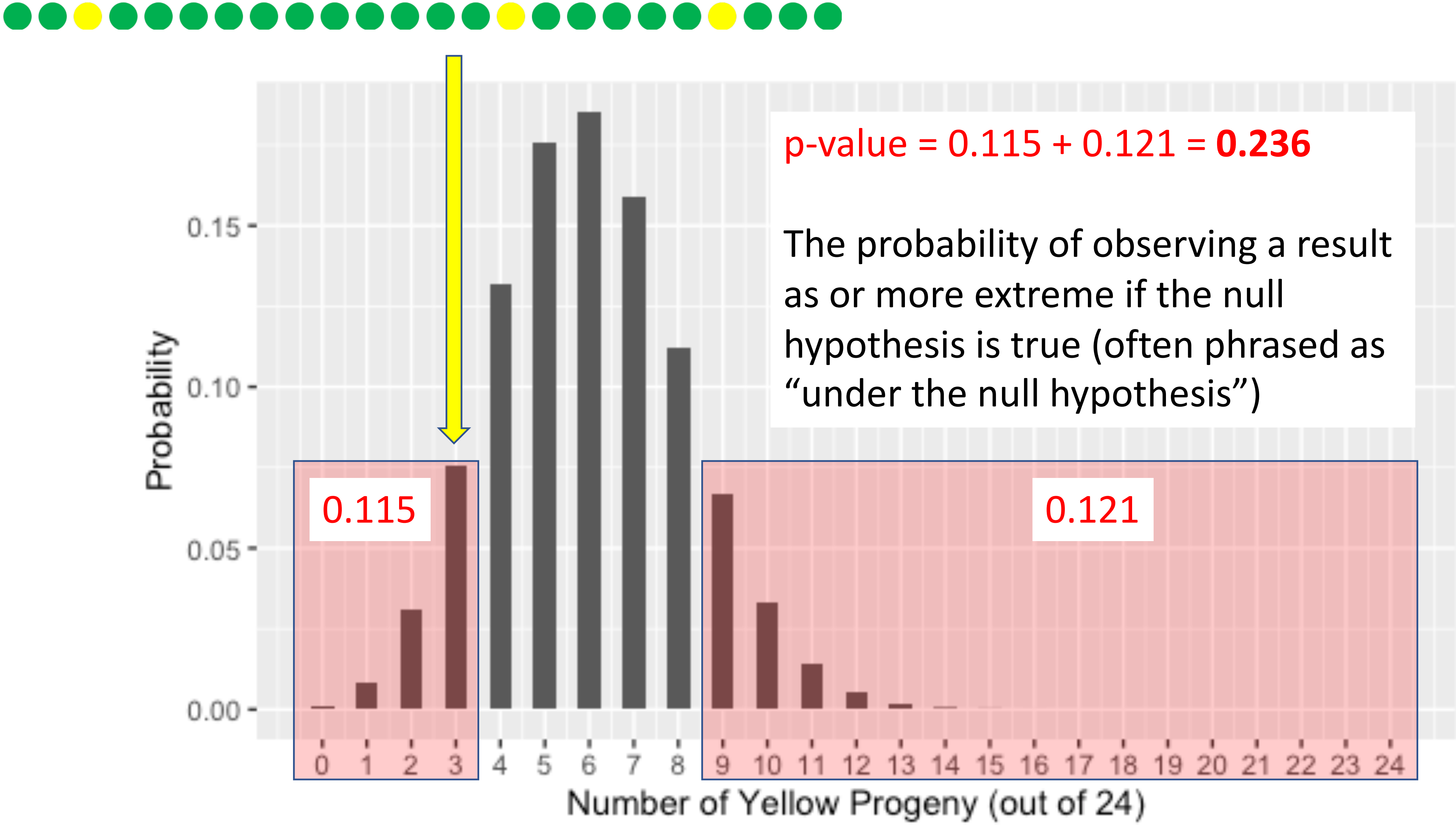
Progeny (n = 24)



What is a p-value?

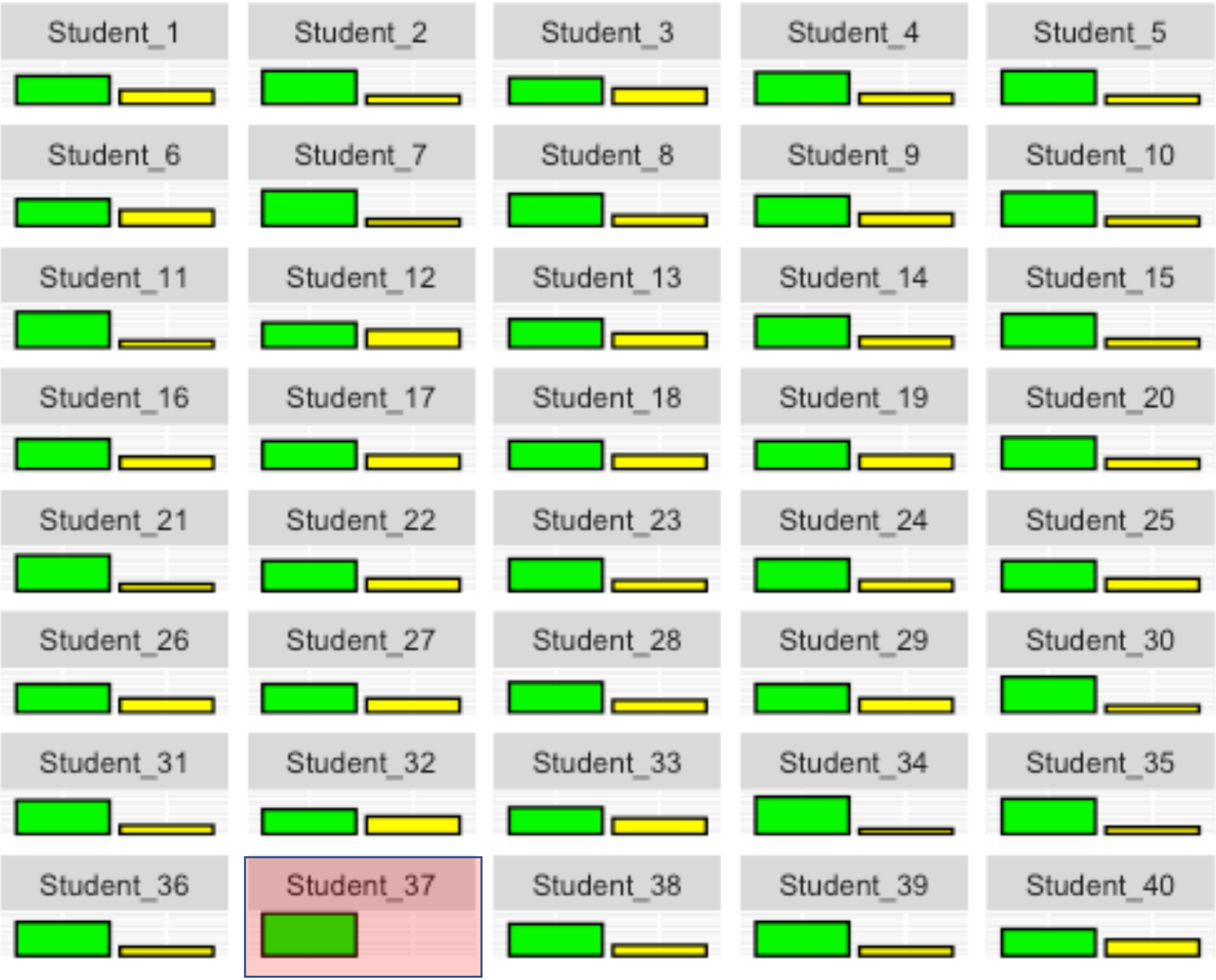
The probability of observing a result as or more extreme if the null hypothesis is true (often phrased as "under the null hypothesis")

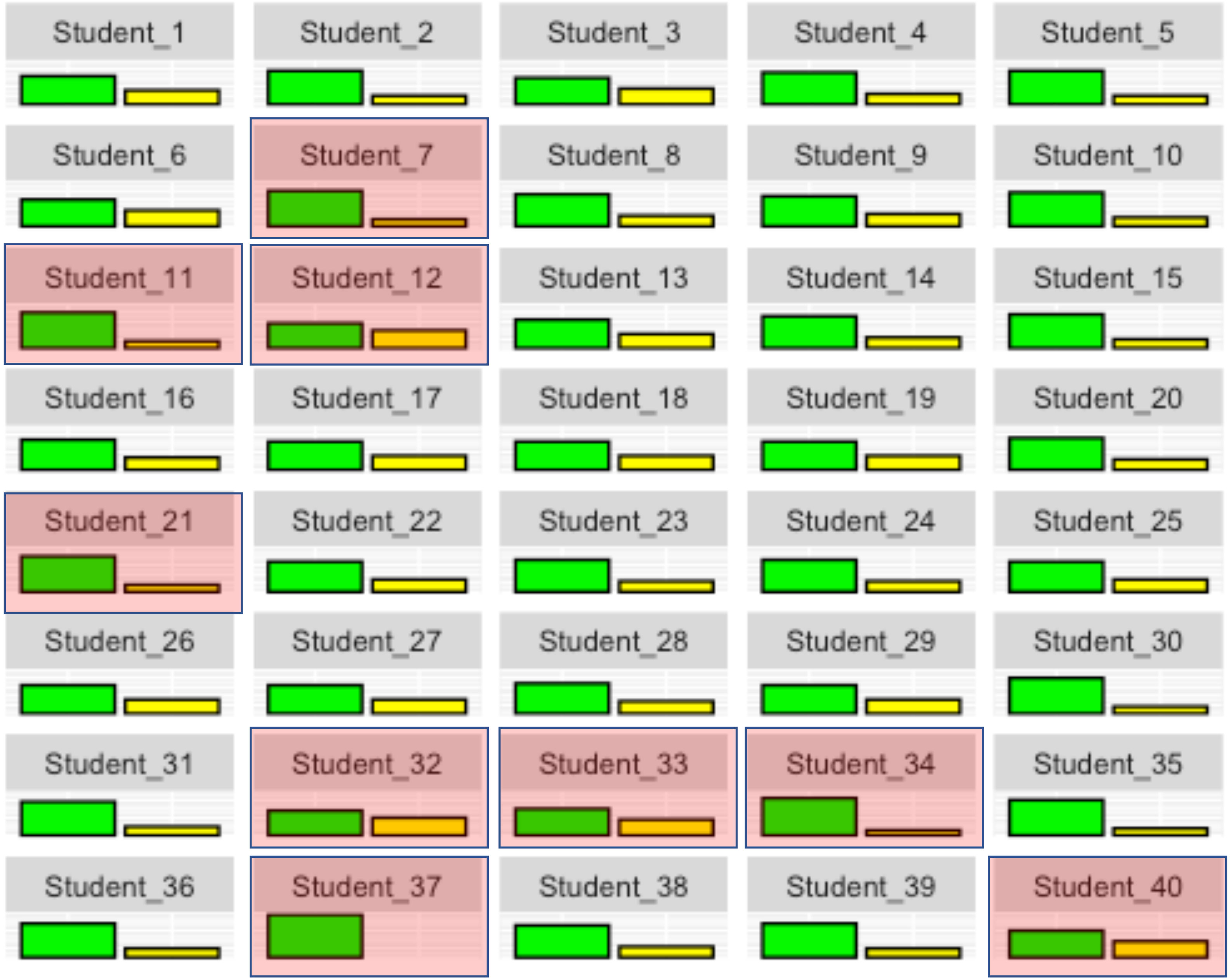
Progeny (n = 24)

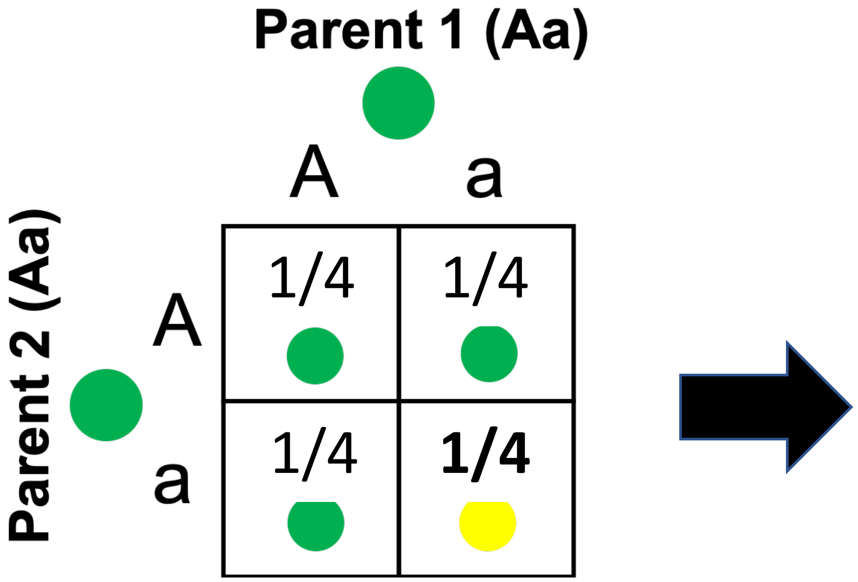


p-value = 0.115 + 0.121 = **0.236**

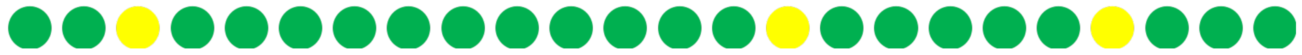
The probability of observing a result as or more extreme if the null hypothesis is true (often phrased as “under the null hypothesis”)


Student_1
Student_2
Student_3
Student_4
Student_5
Student_6
Student_7
Student_8
Student_9
Student_10
Student_11
Student_12
Student_13
Student_14
Student_15
Student_16
Student_17
Student_18
Student_19
Student_20
Student_21
Student_22
Student_23
Student_24
Student_25
Student_26
Student_27
Student_28
Student_29
Student_30
Student_31
Student_32
Student_33
Student_34
Student_35
Student_36
Student_37
Student_38
Student_39
Student_40


Student_1
Student_2
Student_3
Student_4
Student_5
Student_6
Student_7
Student_8
Student_9
Student_10
Student_11
Student_12
Student_13
Student_14
Student_15
Student_16
Student_17
Student_18
Student_19
Student_20
Student_21
Student_22
Student_23
Student_24
Student_25
Student_26
Student_27
Student_28
Student_29
Student_30
Student_31
Student_32
Student_33
Student_34
Student_35
Student_36
Student_37
Student_38
Student_39
Student_40



Progeny (n = 24)



$$\widehat{p_y} = \frac{3}{24}$$

Null hypothesis

$$H_0 : p_y = \frac{1}{4}$$

How much evidence is this against the null hypothesis?

A result this extreme happens 23.6% of the time…

Key concepts

- Statistical inference is the process of drawing conclusions about an underlying population based on a sample or subset of the data

- Hypothesis testing is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis

Key concepts

- Statistical inference is the process of drawing conclusions about an underlying population based on a sample or subset of the data

- Hypothesis testing is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis

Key concepts

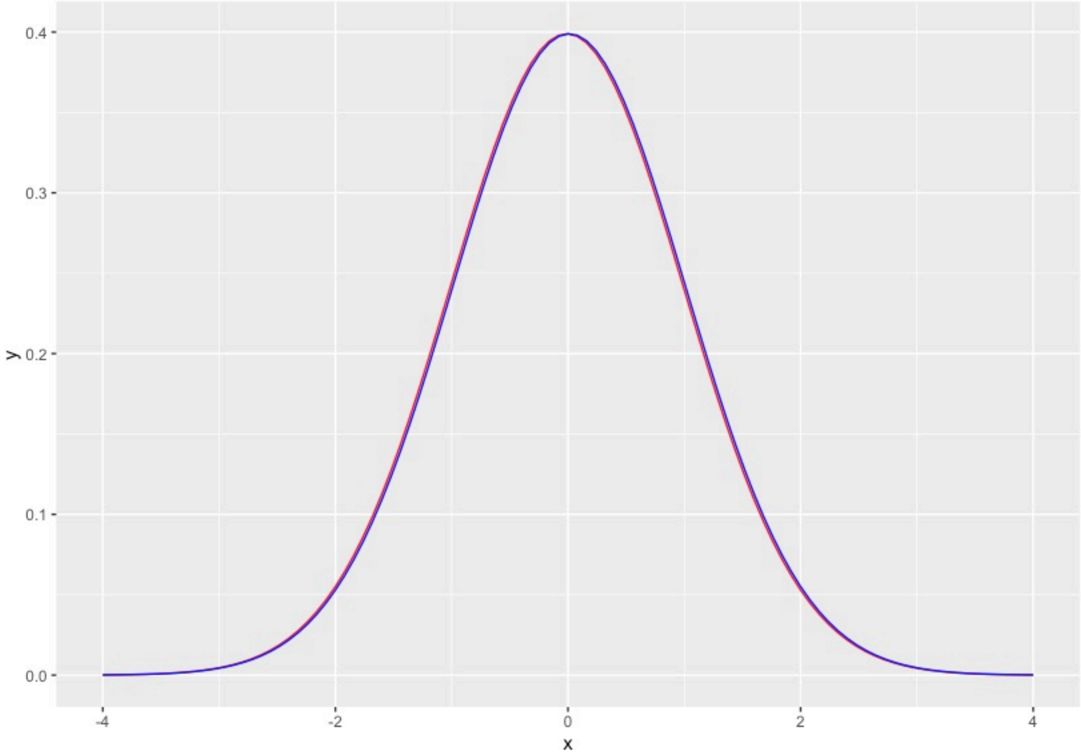
- Statistical inference is the process of drawing conclusions about an underlying population based on a sample or subset of the data

- Hypothesis testing is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis

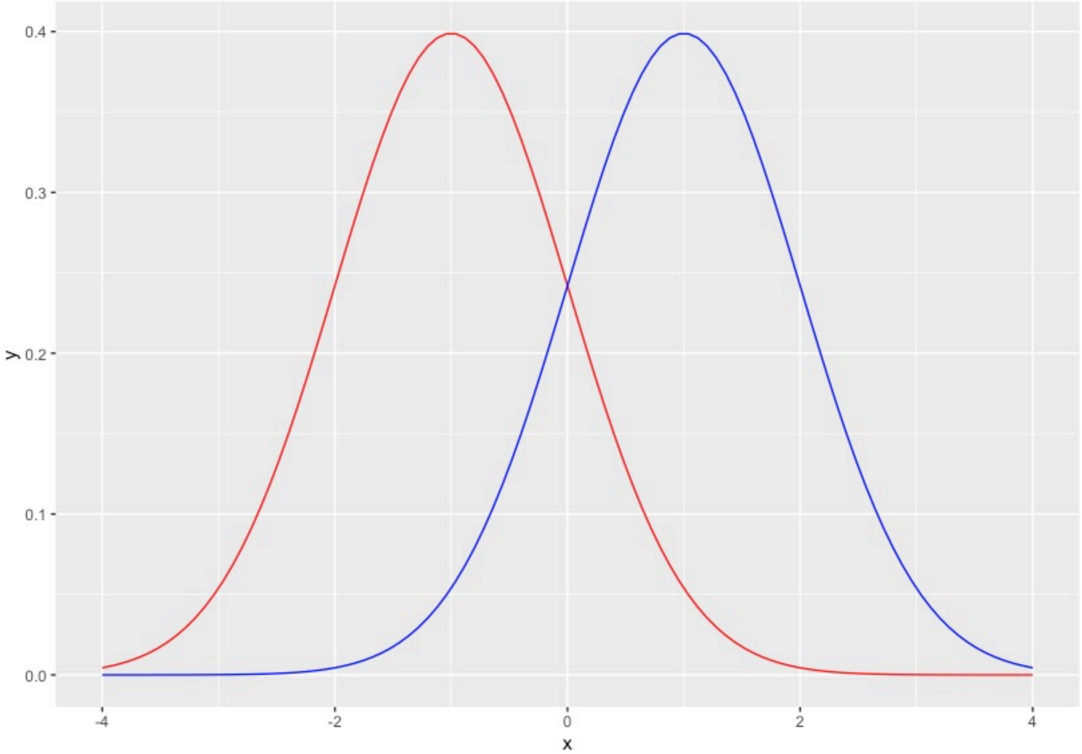
- For hypothesis testing, define any useful (test) statistic and figure out its sampling distribution under the null hypothesis. That's it…

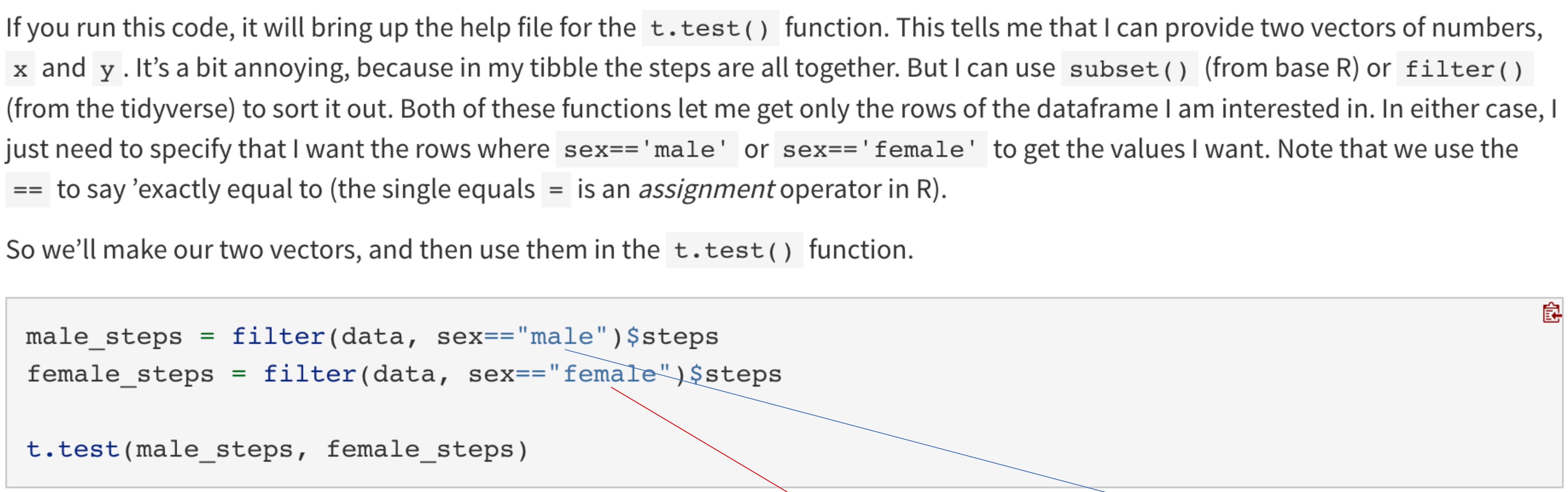
Working example: two sample test

H_0: Distributions A and B have the same mean



H_A: Distributions A and B have different means



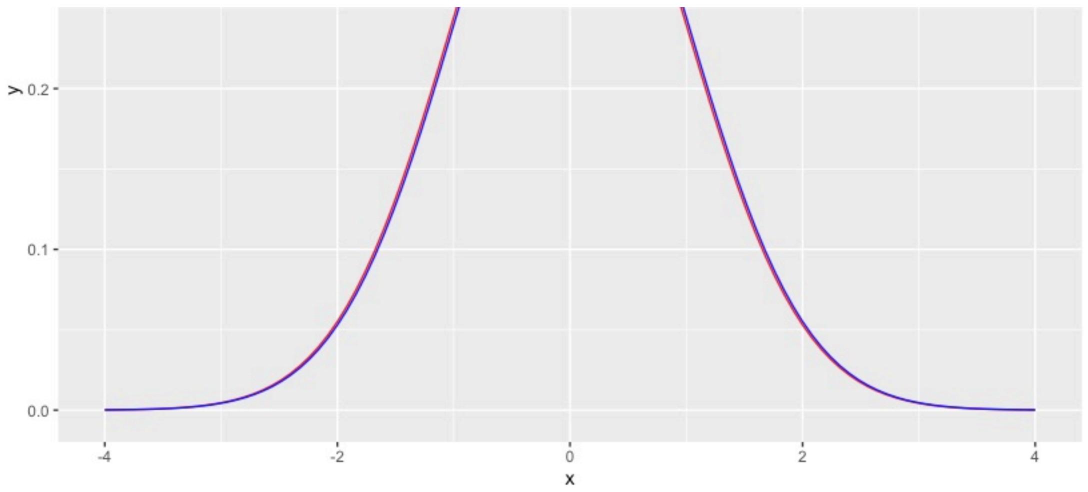


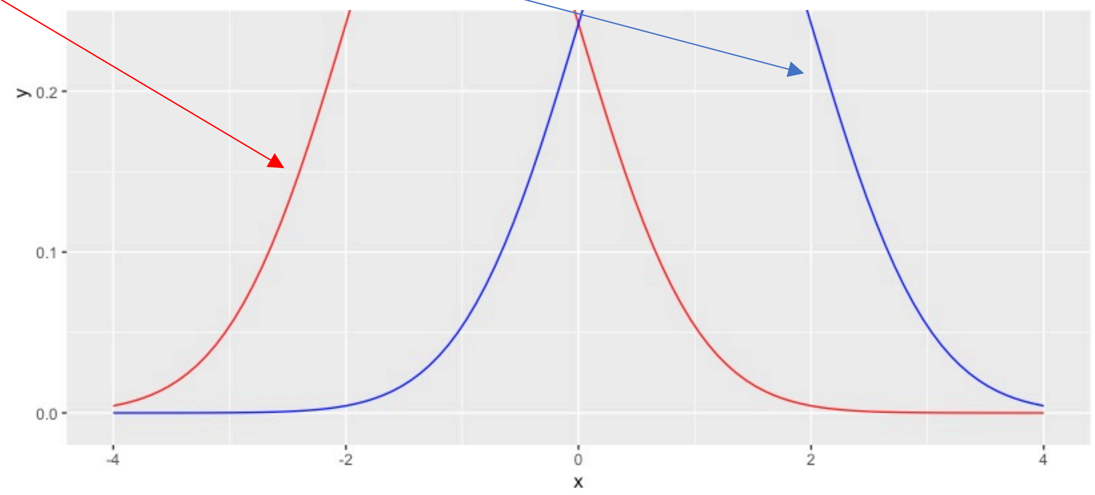
If you run this code, it will bring up the help file for the `t.test()` function. This tells me that I can provide two vectors of numbers, `x` and `y`. It’s a bit annoying, because in my tibble the steps are all together. But I can use `subset()` (from base R) or `filter()` (from the tidyverse) to sort it out. Both of these functions let me get only the rows of the dataframe I am interested in. In either case, I just need to specify that I want the rows where `sex=='male'` or `sex=='female'` to get the values I want. Note that we use the `==` to say ’exactly equal to (the single equals `=` is an *assignment* operator in R).

So we’ll make our two vectors, and then use them in the `t.test()` function.

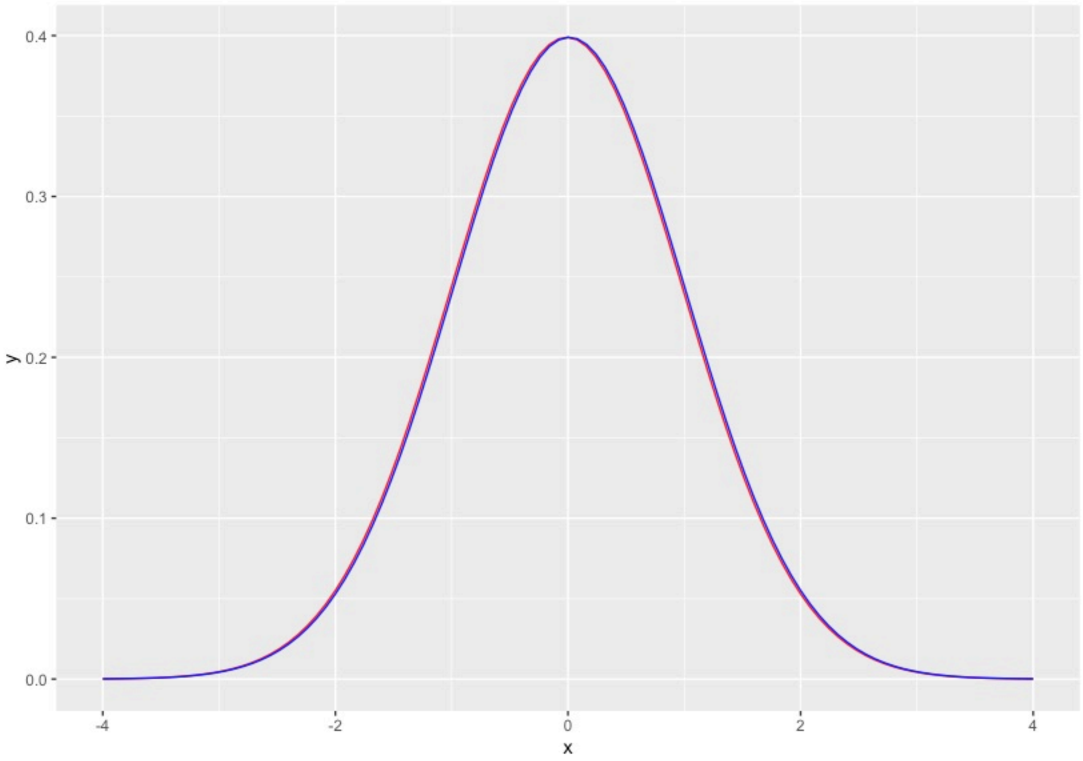
```
male_steps = filter(data, sex=="male")$steps
female_steps = filter(data, sex=="female")$steps

t.test(male_steps, female_steps)
```

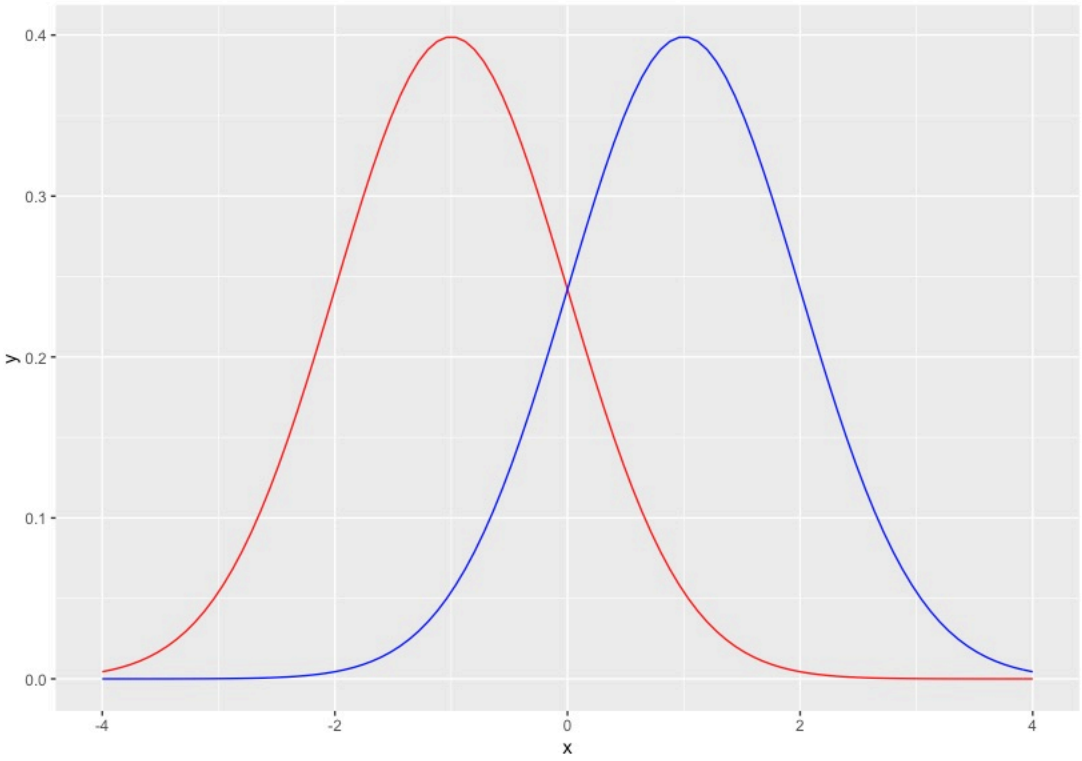




H_0: Distributions A and B have the same mean



H_A: Distributions A and B have different means



- Sample 50 numbers from distribution A (call these set_A)
- Sample 50 numbers from distribution B (call these set_B)
- Perform *some kind of* two-sample test

Worked exercise: Two-sample t-test

The plan:

- Read in the data from “GroupData12Aug.csv”
- Do a little data wrangling
- Do a little data visualisation
- Perform a simple t-test

- *Interpret this in terms of the sampling distribution*
  - *Sampling distribution of what?!?*

Worked exercise: Permutation t-test

The plan:

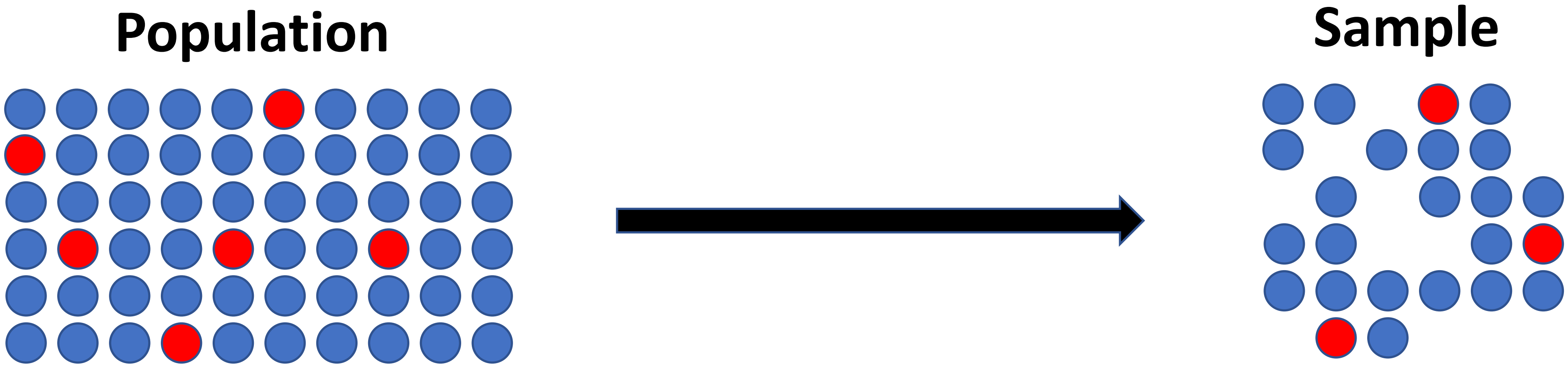
- Do not assume the t-distribution!
- Instead, simulate null distribution using permutation

- Combine Set_A and Set_B into one set of length 100
- Randomly reassign the labels to create new sets
- Compute the test statistic
- Repeat n_reps times
- Compare real value to this distribution

OLD →	NEW
1	3
2	2
3	5
4	4
5	1

Extension

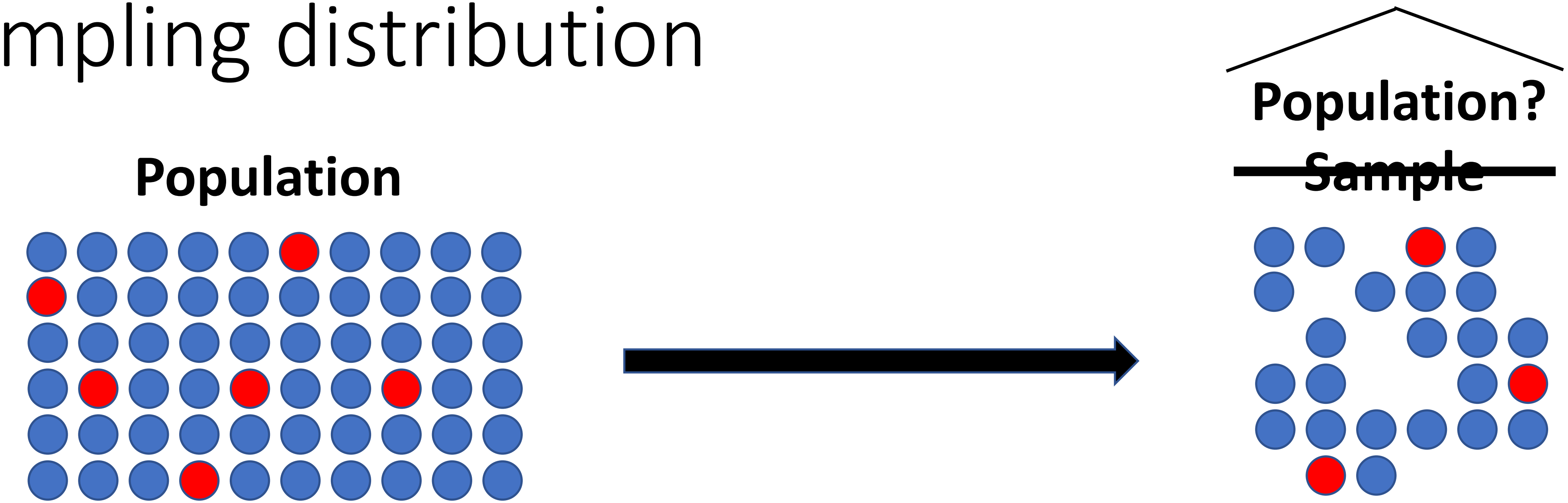
Sampling distribution



Population

Sample

Sampling distribution



Population

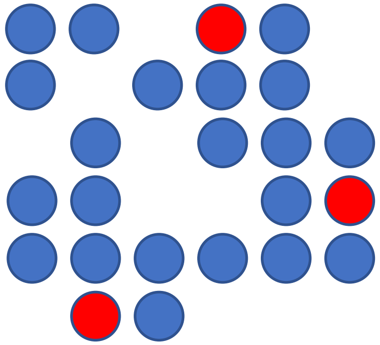
Population?

~~**Sample**~~

Bootstrap sampling distribution

Your sample

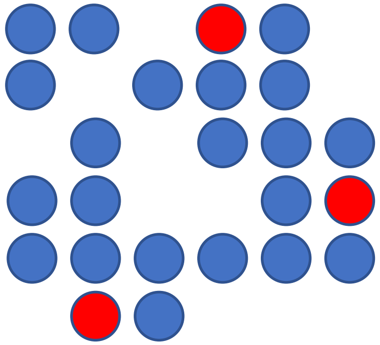
***n* = 24**



Bootstrap sampling distribution

Your sample

***n* = 24**



Sample with replacement

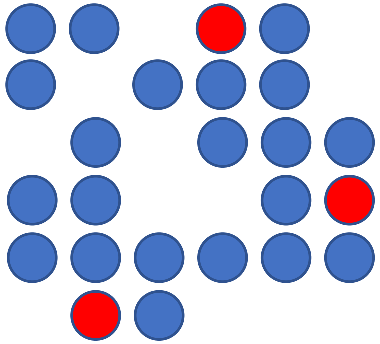
Bootstrap sample 1

***n* = 24**

Bootstrap sampling distribution

Your sample

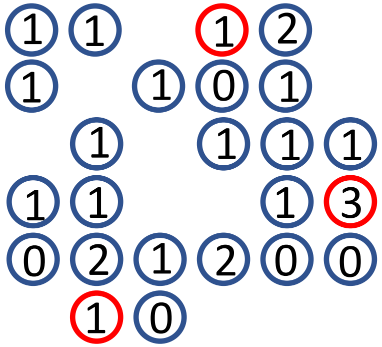
***n* = 24**



Sample with replacement

Bootstrap sample 1

***n* = 24**



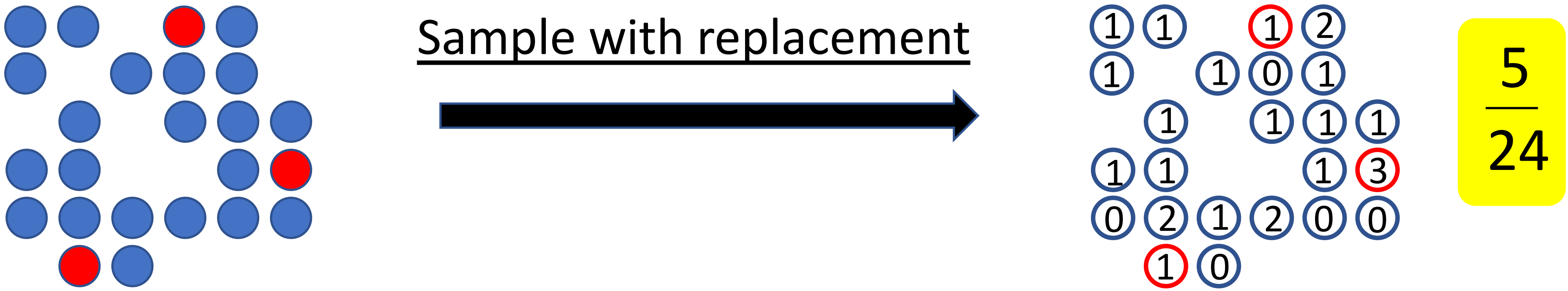
Bootstrap sampling distribution

Your sample

***n* = 24**

Bootstrap sample 1

***n* = 24**



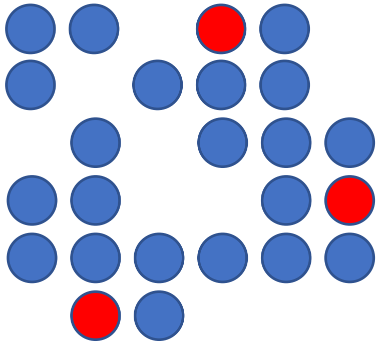
Sample with replacement

$$\frac{5}{24}$$

Bootstrap sampling distribution

Your sample

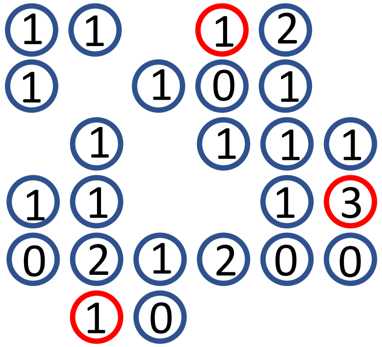
***n* = 24**



Sample with replacement

Bootstrap sample 1

***n* = 24**



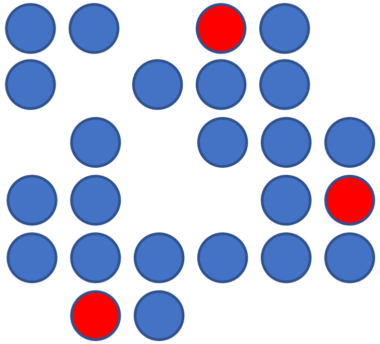
$\frac{5}{24}$

Many bootstrap samples

Exercise: Bootstrap sampling distribution in R

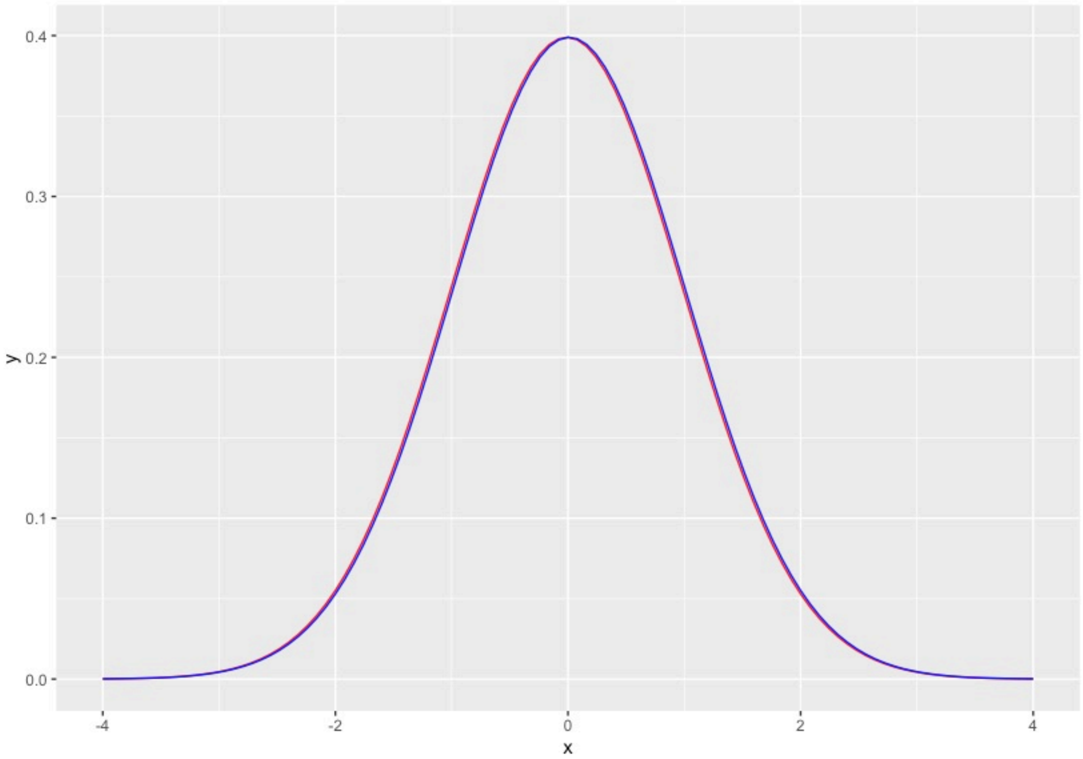
Your sample

***n* = 24**

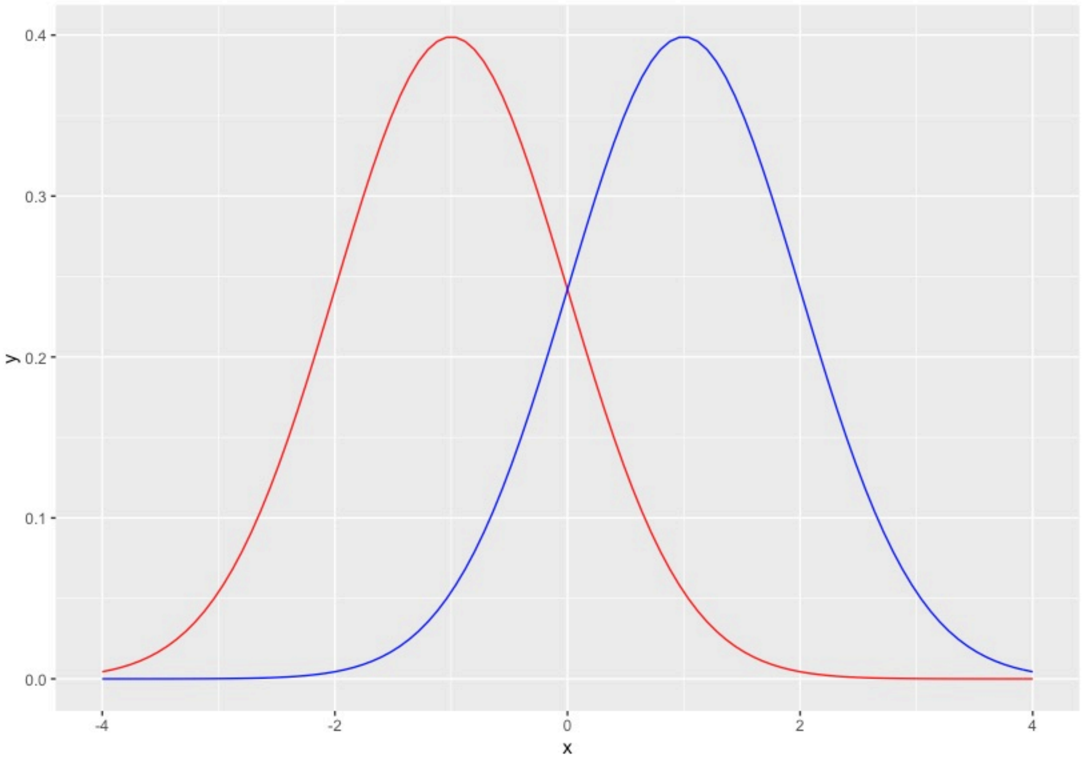


1. Write code to sample with replacement a sample of size 24 from your sample (3 red; 21 blue).
2. Compute the statistic #red/24
3. Wrap these steps in a for loop (or use apply) to repeat the process n_reps = 10000 times, recording the statistic in a vector
4. Plot the histogram of the statistics computed in the previous step *(bonus: and compare the sampling distribution from before)*

H_0: Distributions A and B have the same mean



H_A: Distributions A and B have different means



- Sample 50 numbers from distribution A (call these set_A)
- Sample 50 numbers from distribution B (call these set_B)
- Perform *some kind of* two-sample test

Recall: Two-sample t-test

The plan:

- Read in the data from “GroupData12Aug.csv”
- Do a little data wrangling
- Do a little data visualisation
- Perform a simple t-test

- *Interpret this in terms of the sampling distribution*
  - *Sampling distribution of what?!?*

Recall: Permutation t-test

The plan:

- Do not assume the t-distribution!
- Instead, simulate null distribution using permutation

- Combine Set_A and Set_B into one set of length 100
- Randomly reassign the labels to create new sets
- Compute the test statistic
- Repeat n_reps times
- Compare real value to this distribution

OLD	→	NEW
1		3
2		2
3		5
4		4
5		1

Exercise: Bootstrap t-test

The plan:

- This time, simulate the null distribution using the bootstrap

- *How would you go about doing this?*

More Computational Statistics
and
Data Visualisation