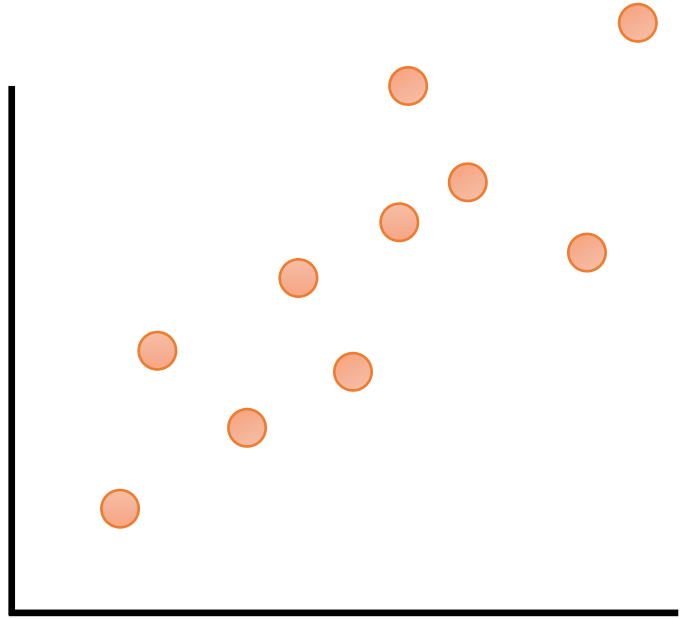
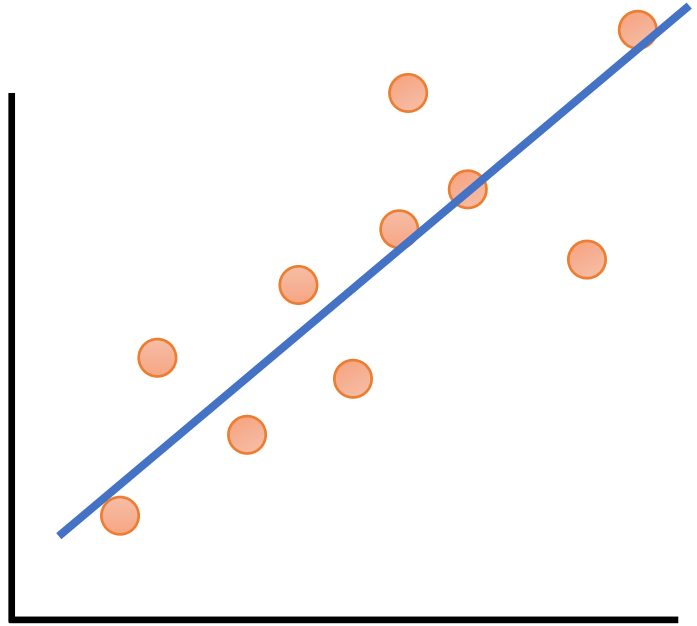
Linear Models

Part II

Last time



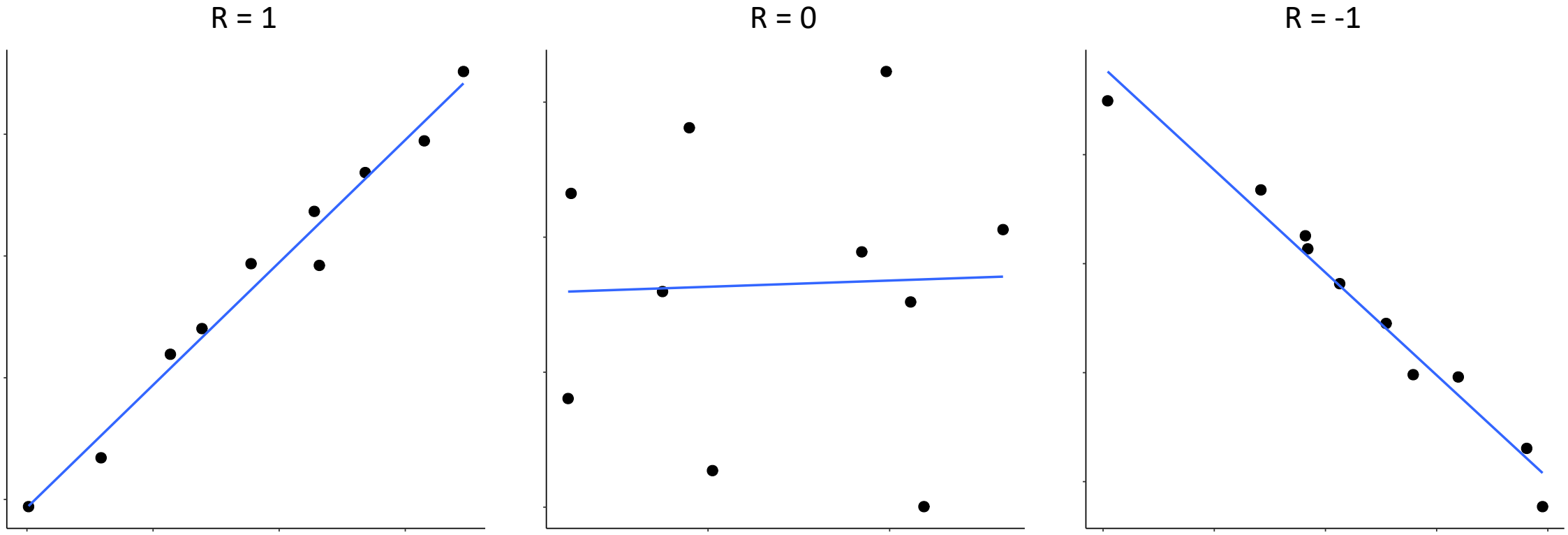
Last time



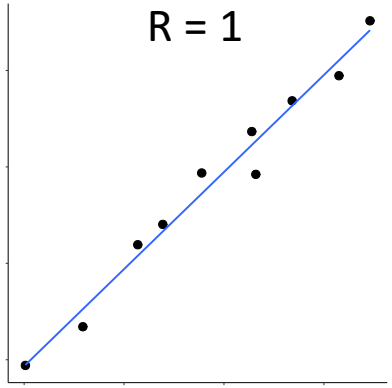
```
Call:
lm(formula = y ~ x, data = df)

Residuals:
      1       2       3       4       5       6       7
 5.4033 -4.9961 -0.1866 -0.8958  1.6027 -0.6310 -0.2965

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.93686    6.12210    0.48 0.651679
x            0.94070    0.09324   10.09 0.000164 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.407 on 5 degrees of freedom
Multiple R-squared:  0.9532,    Adjusted R-squared:  0.9438
F-statistic: 101.8 on 1 and 5 DF,  p-value: 0.0001638
```

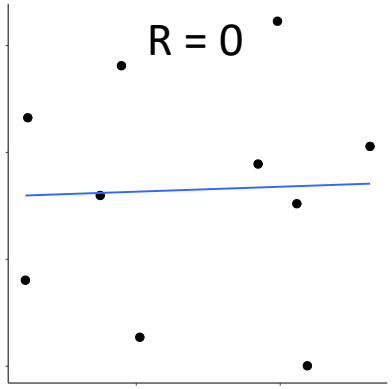
```
Call:
lm(formula = y ~ x, data = df)

Residuals:
      1       2       3       4       5       6       7
 5.4033 -4.9961 -0.1866 -0.8958  1.6027 -0.6310 -0.2965

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.93686    6.12210    0.48 0.651679
x            0.94070    0.09324   10.09 0.000164 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.407 on 5 degrees of freedom
Multiple R-squared:  0.9532,    Adjusted R-squared:  0.9438
F-statistic: 101.8 on 1 and 5 DF,  p-value: 0.0001638
```

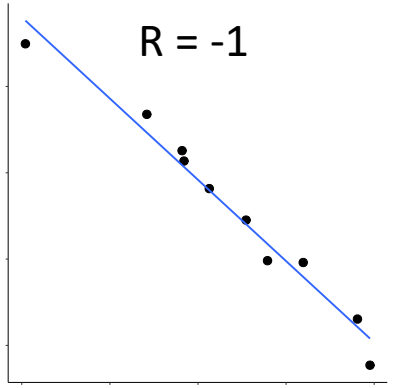
Correlation-Coefficient “R”



R and R^2





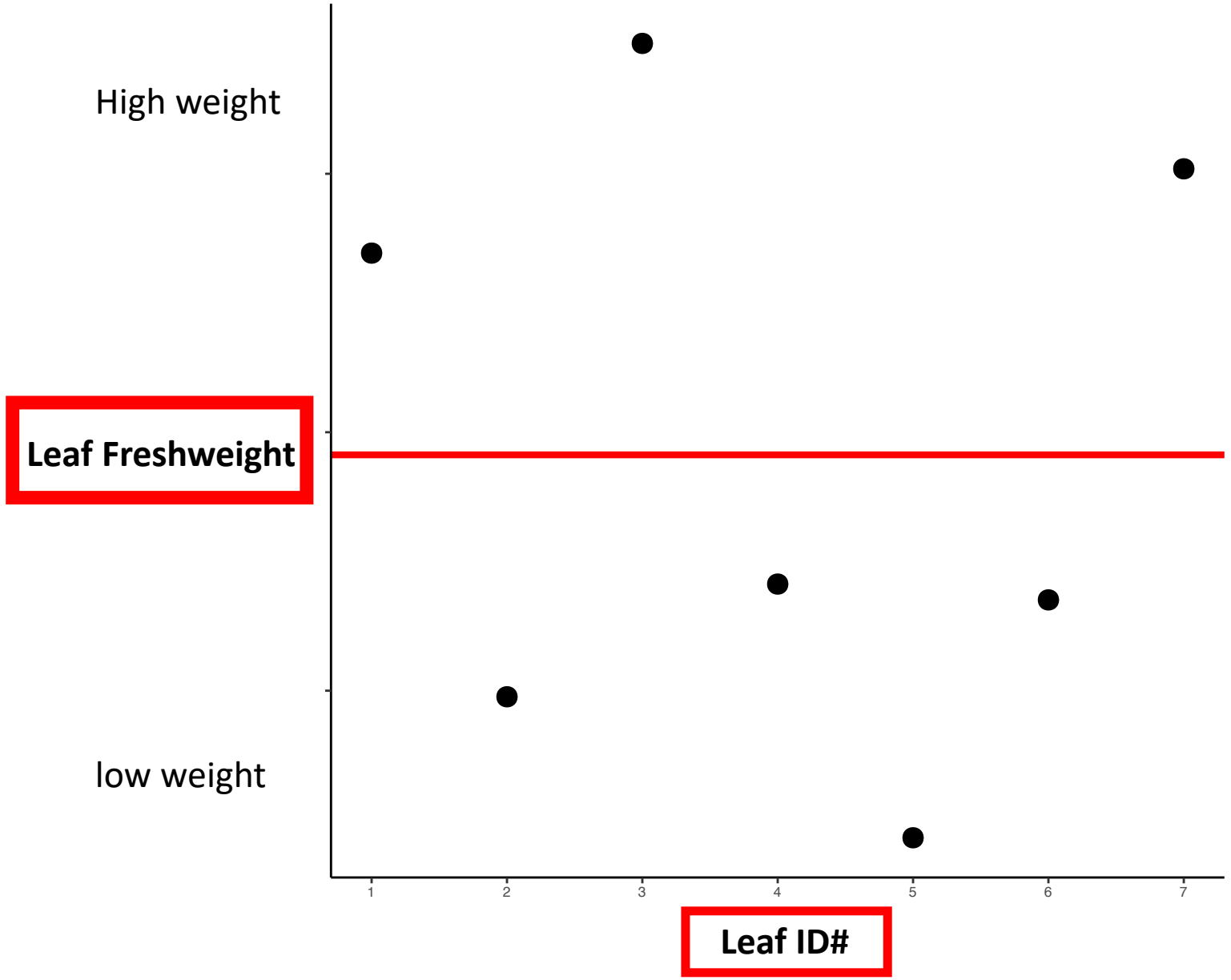


You probably know about correlation (regular "R")

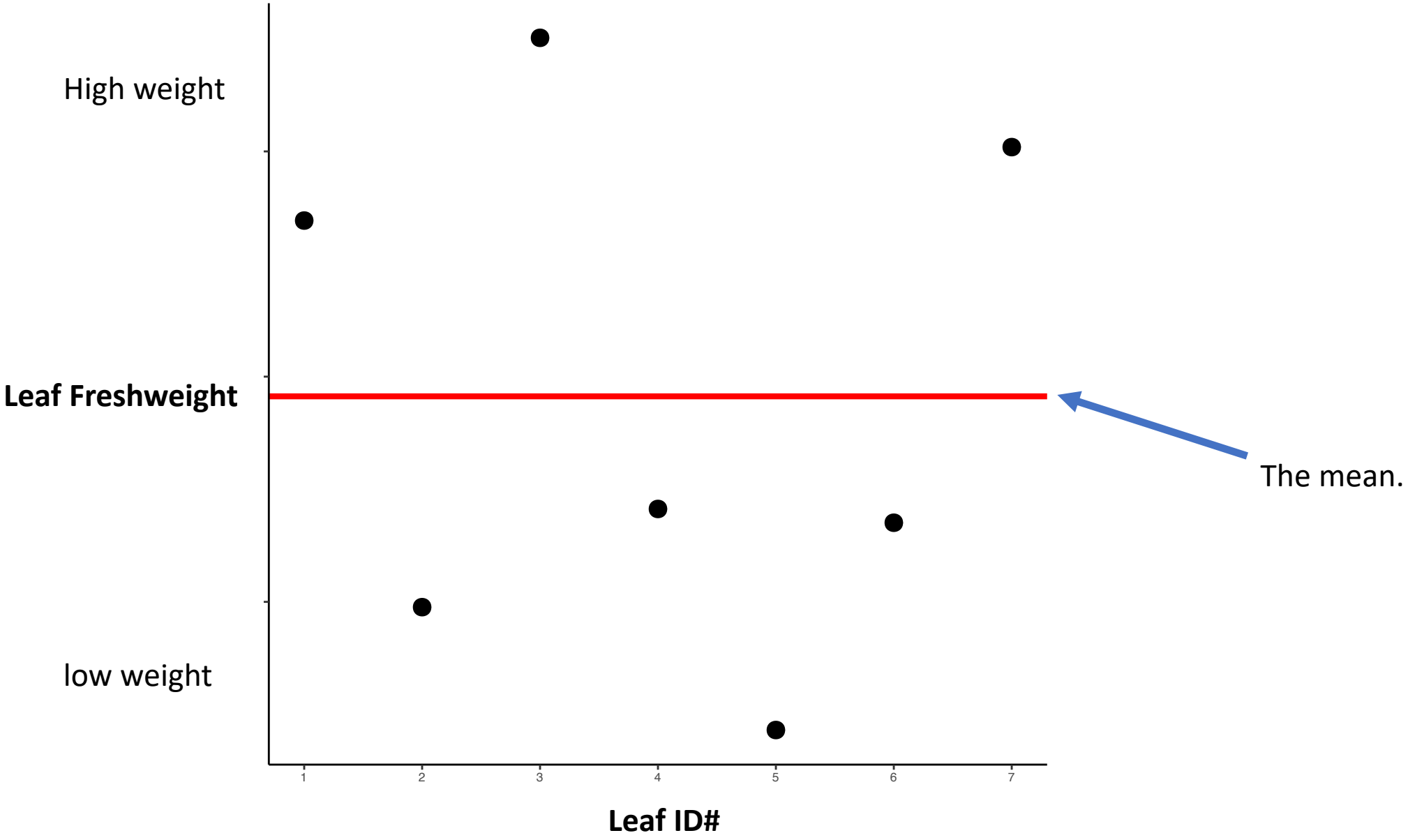
- Correlation values close to 1 or -1 are good and tell you two quantitative variables are strongly related.
- Correlation values close to 0 are lame

Why should we care about R^2? (and what even is that?)

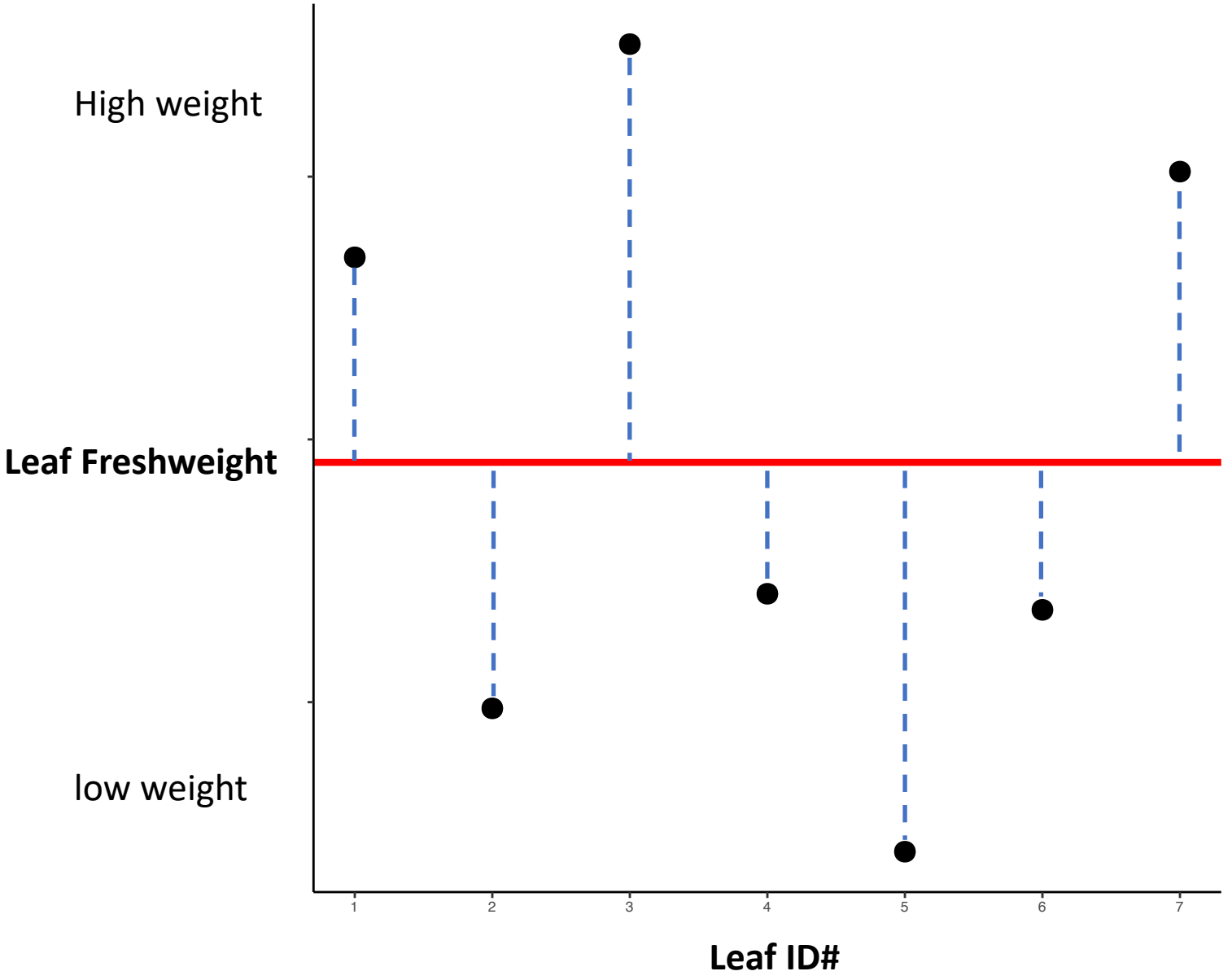
- R^2 is very similar to its hipper cousin, R, but..
- Interpretation is easier:
  - It's not obvious that (regular) R = 0.7 is twice as good as R = 0.5
  - However, R^2=0.7 is what it looks like, 1.4 times as good as R^2=0.5


High weight
Leaf Freshweight
low weight
1
2
3
4
5
6
7
Leaf ID#

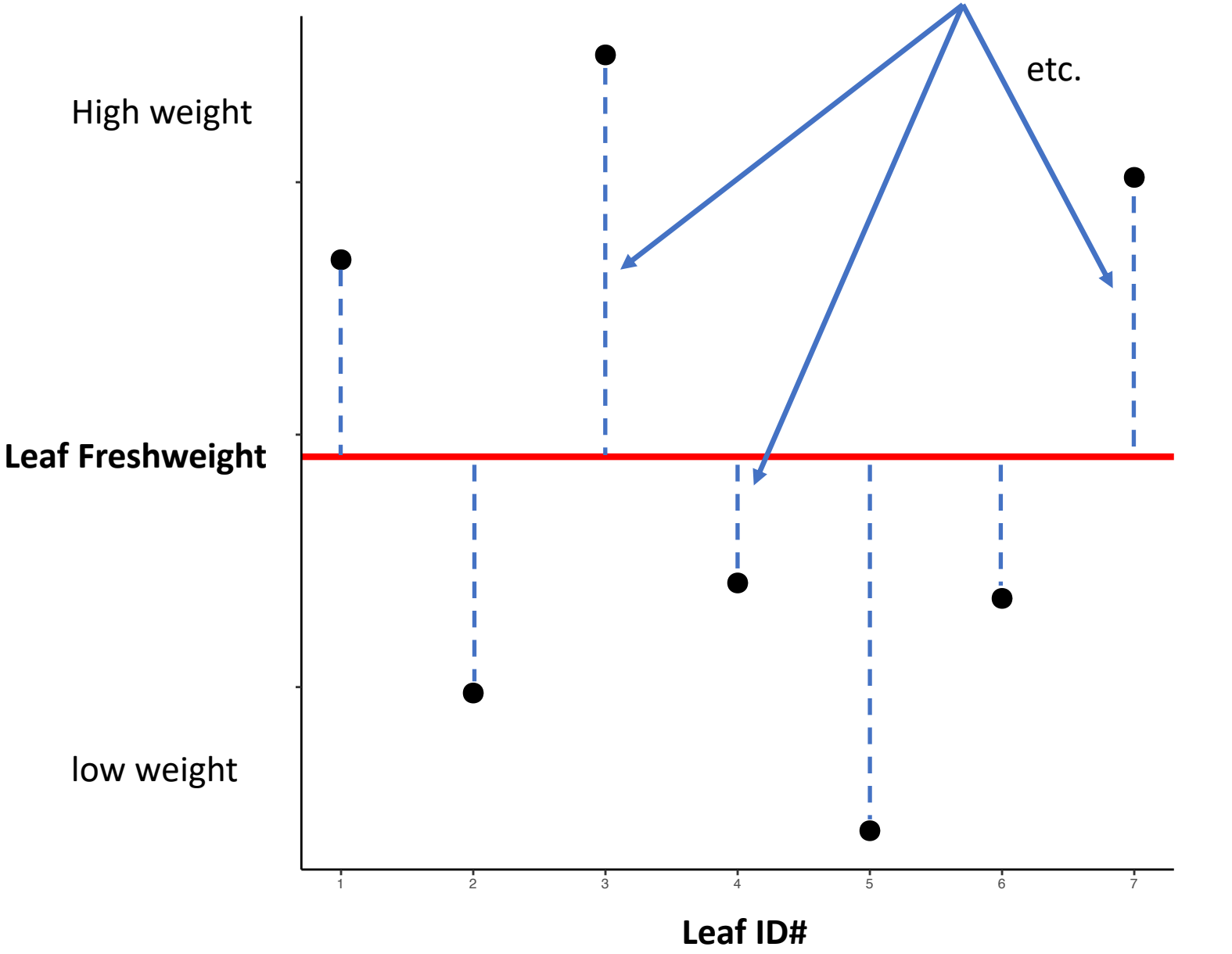
The Mean (or Average) of leaf weights



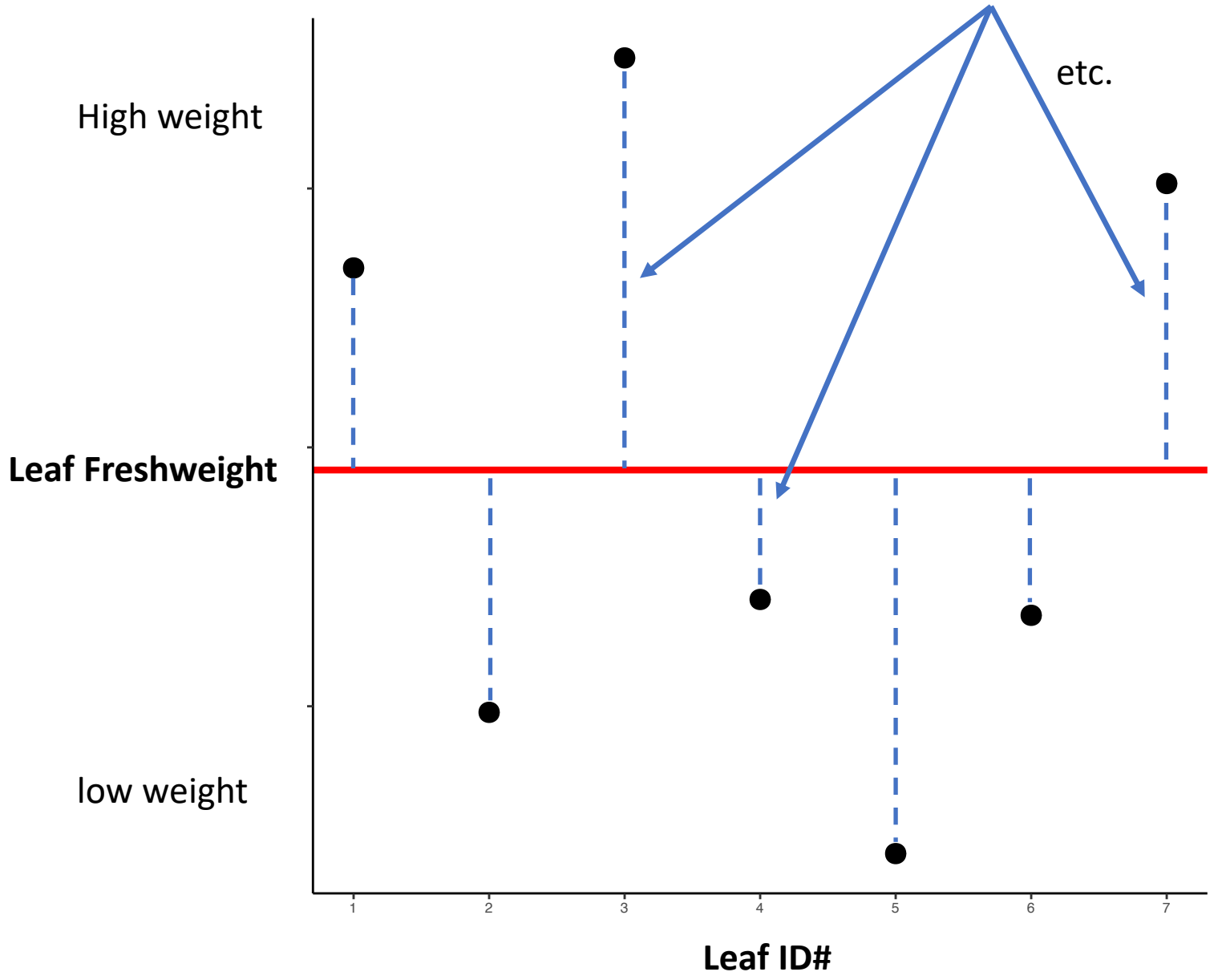
The Variation of the data = Sum(weight – mean)^2



The Variation of the data = Sum(weight – mean)^2

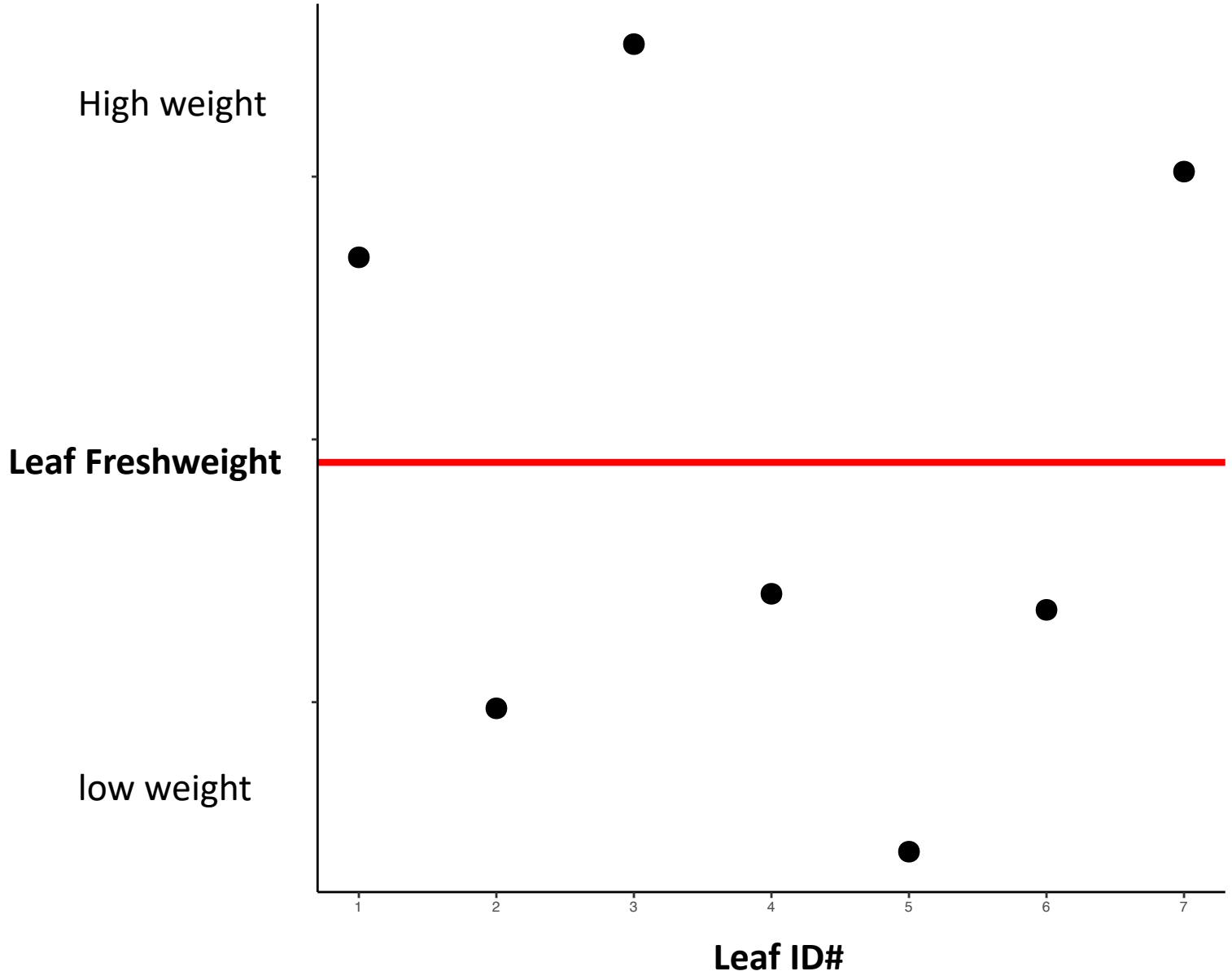


The Variation of the data = Sum(weight – mean)^2

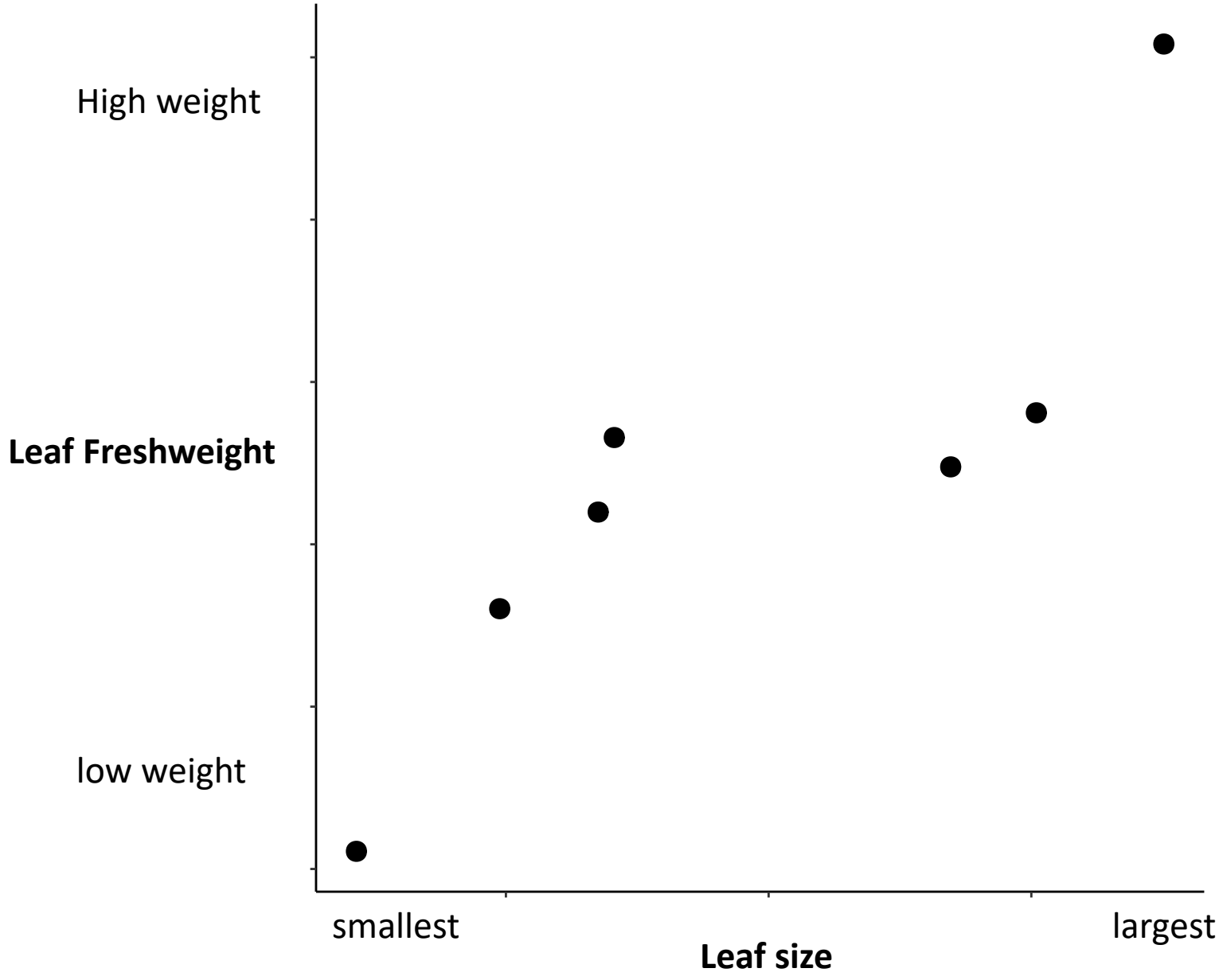


The difference between each data point is squared so that points below the. Mean don't cancel out points above the mean.

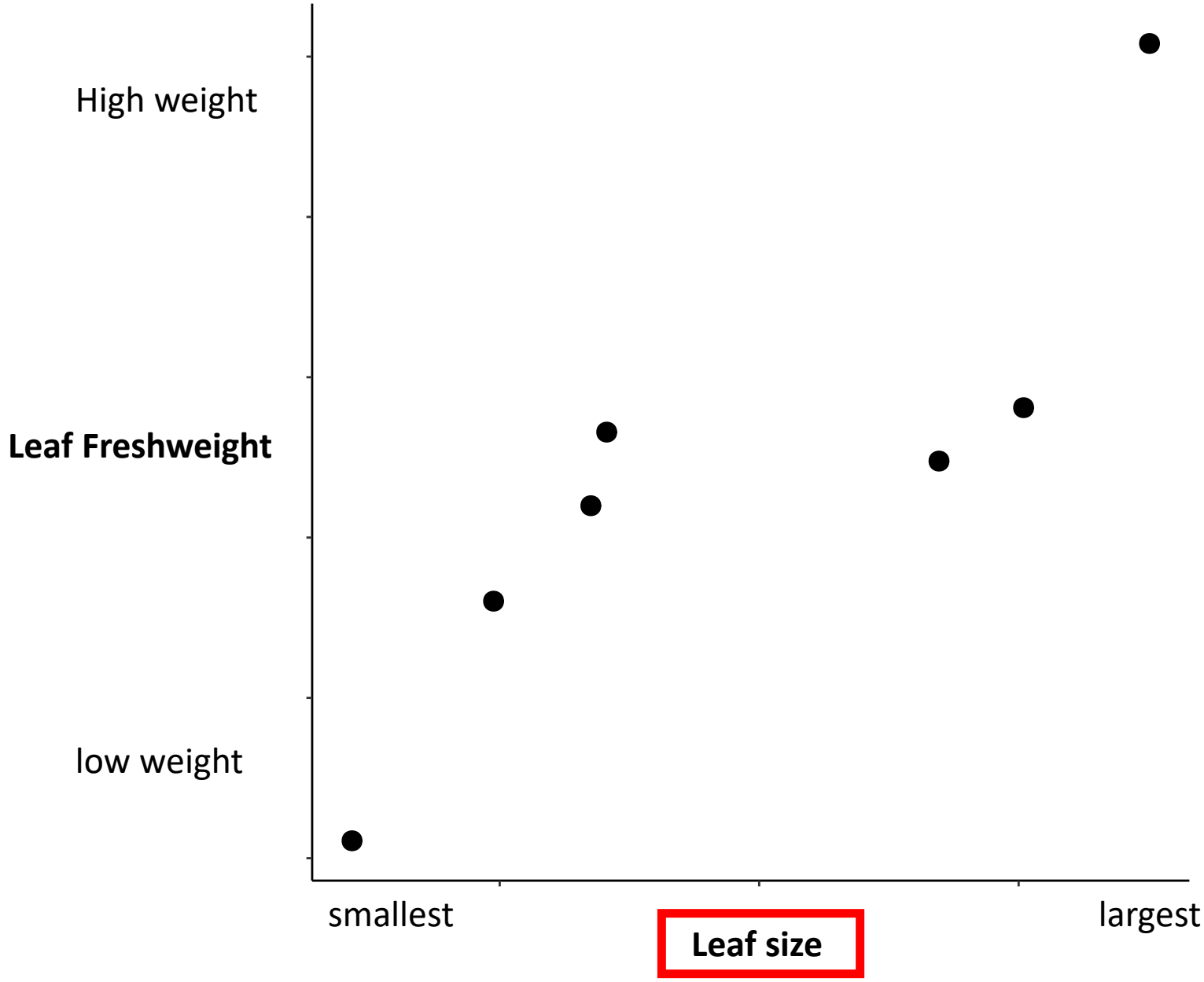
Now, what if, instead of ordering our leafs by their ID#, we ordered them by their size?



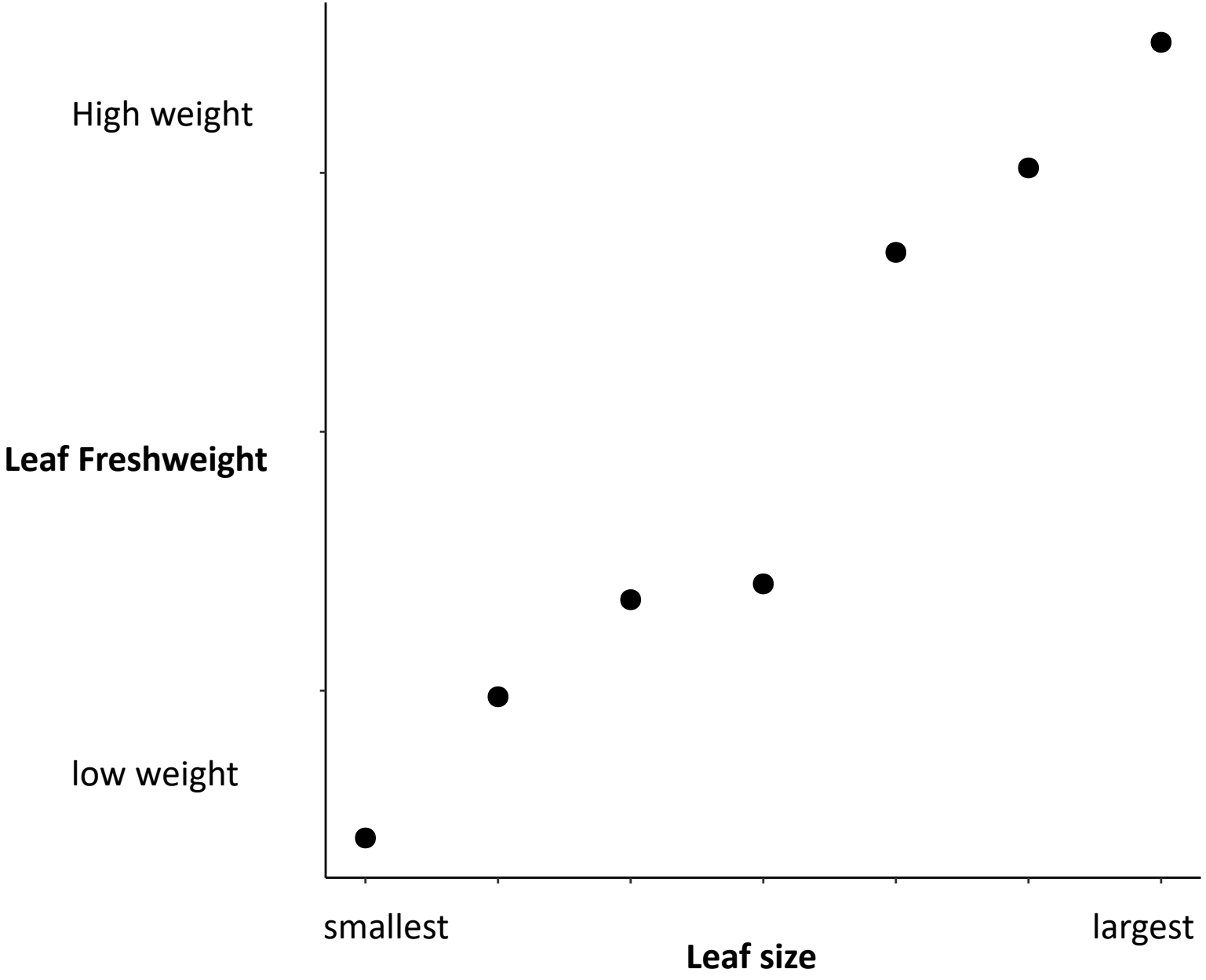
Now, what if, instead of ordering our leafs by their ID#, we ordered them by their size?



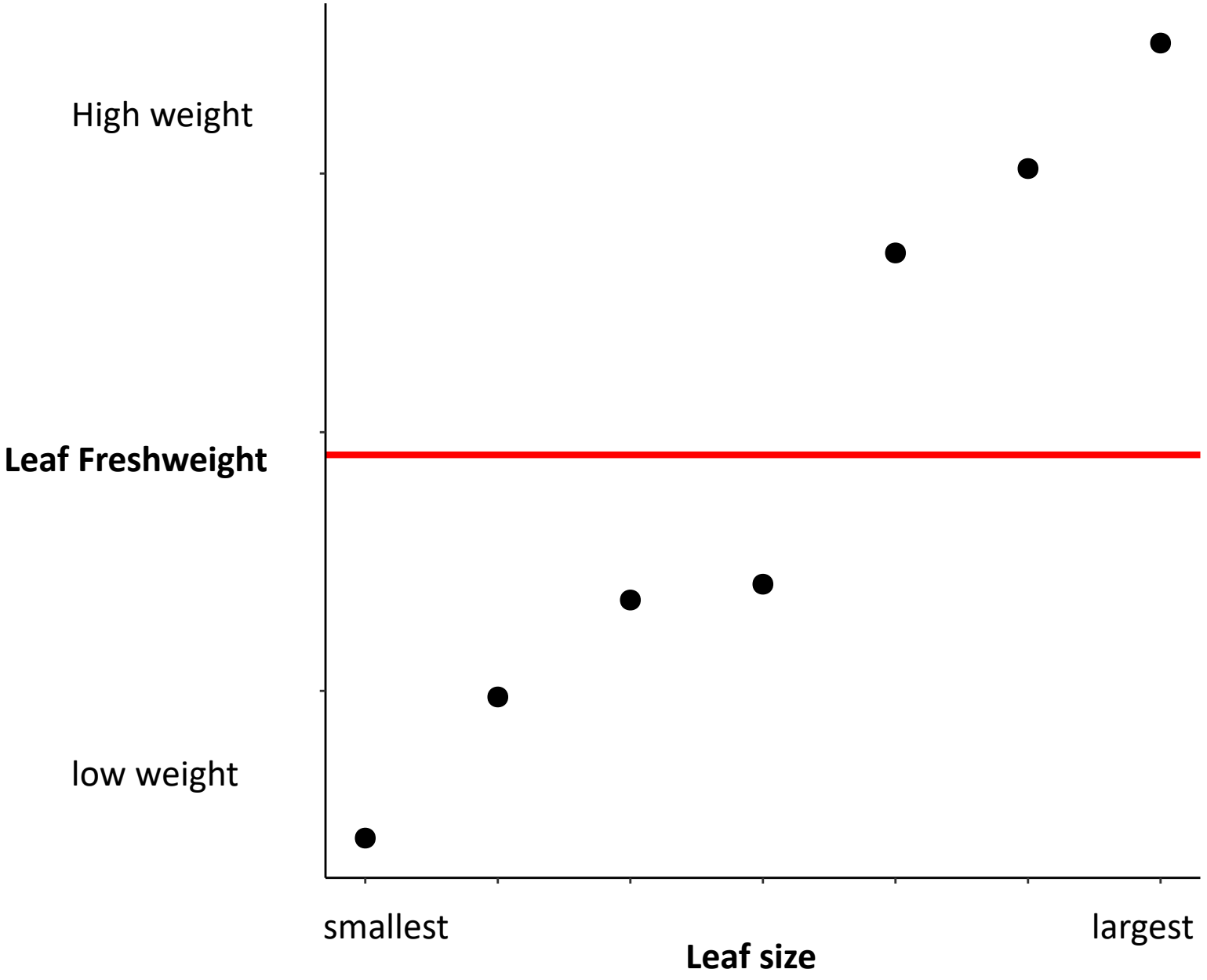
Now, what if, instead of ordering our leafs by their ID#, we ordered them by their size?



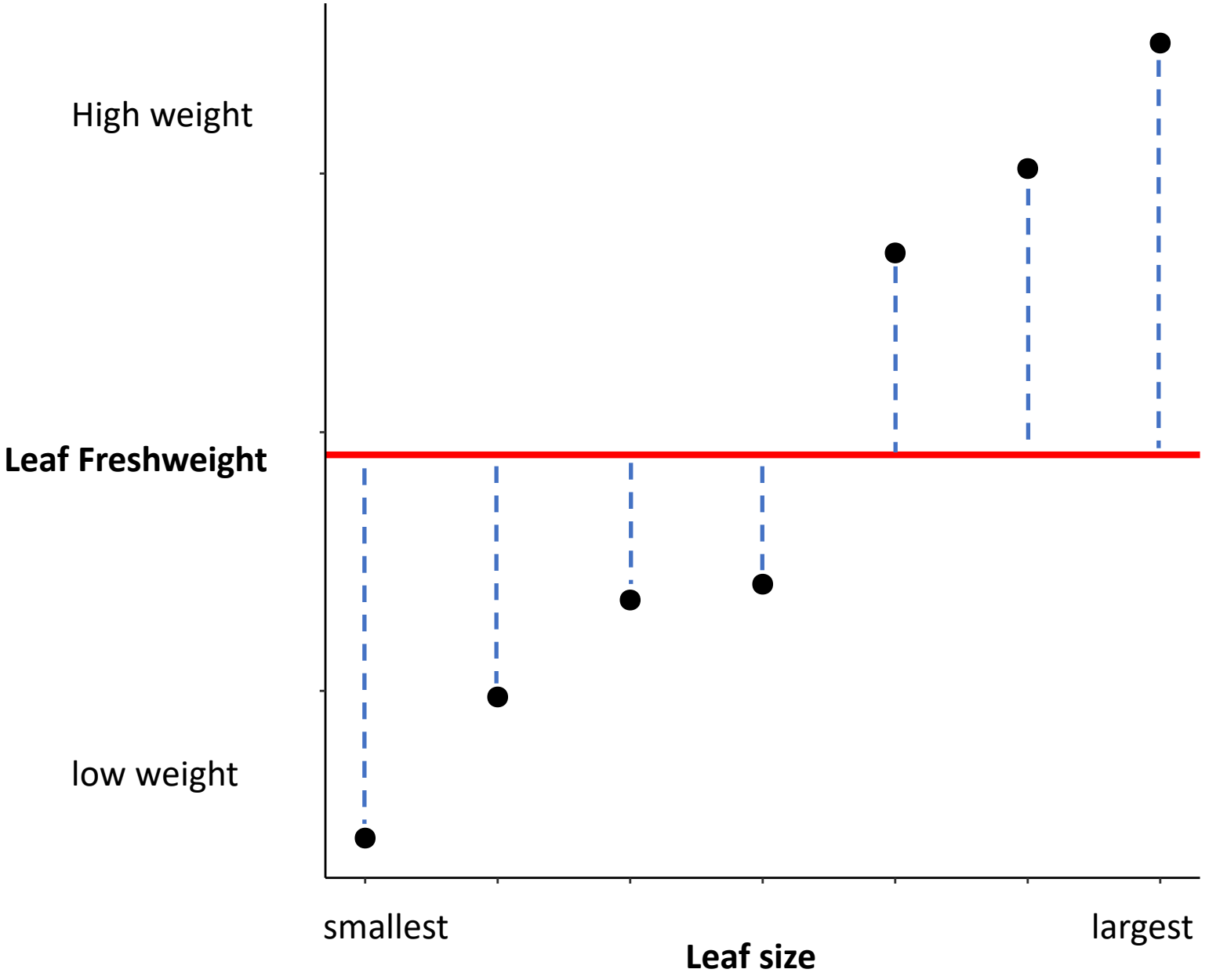
Now, what if, instead of ordering our leafs by their ID#, we ordered them by their size?



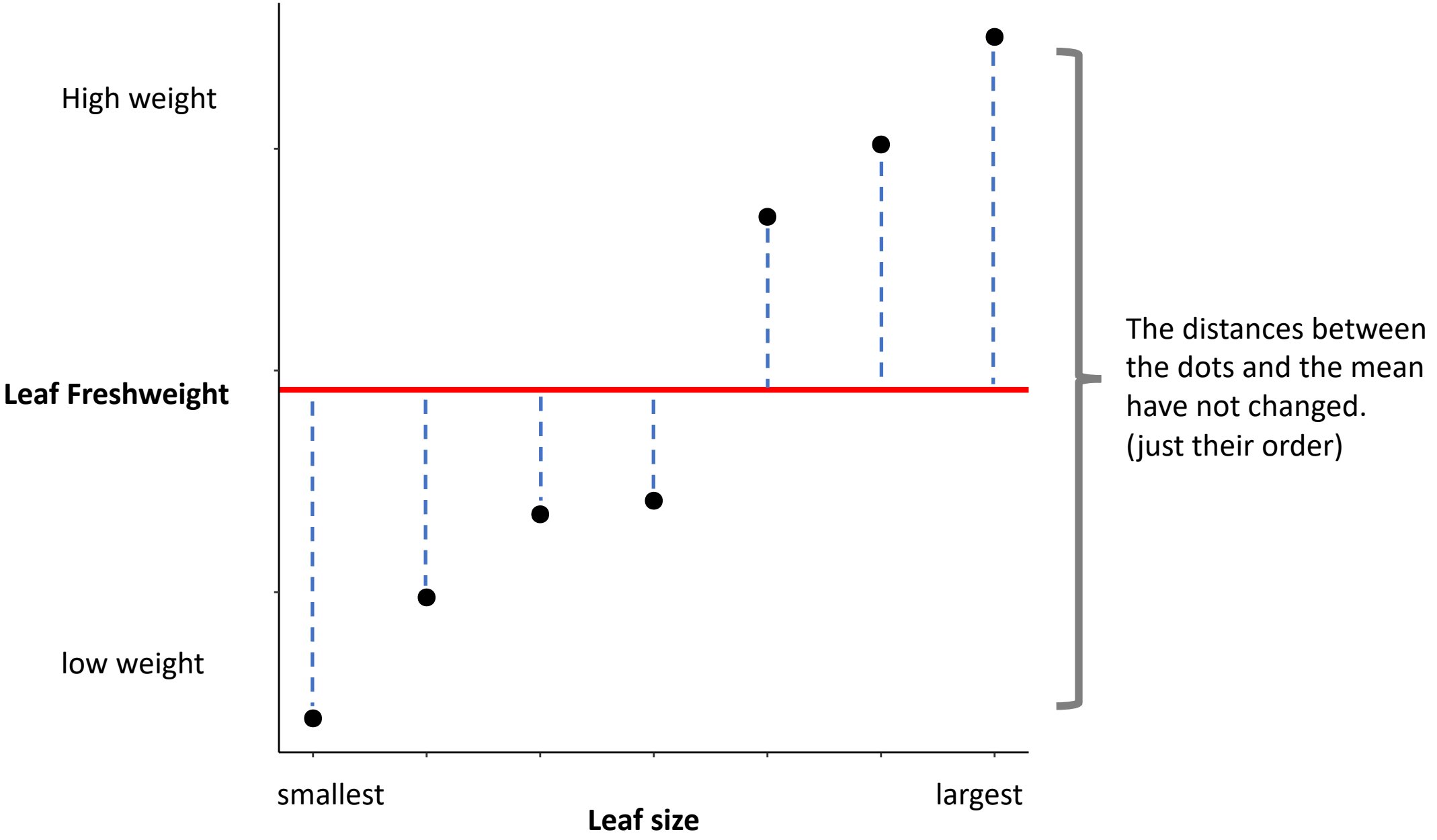
All we have done is reorder the data on the X-axis.
The mean and variation are the exact same as before.



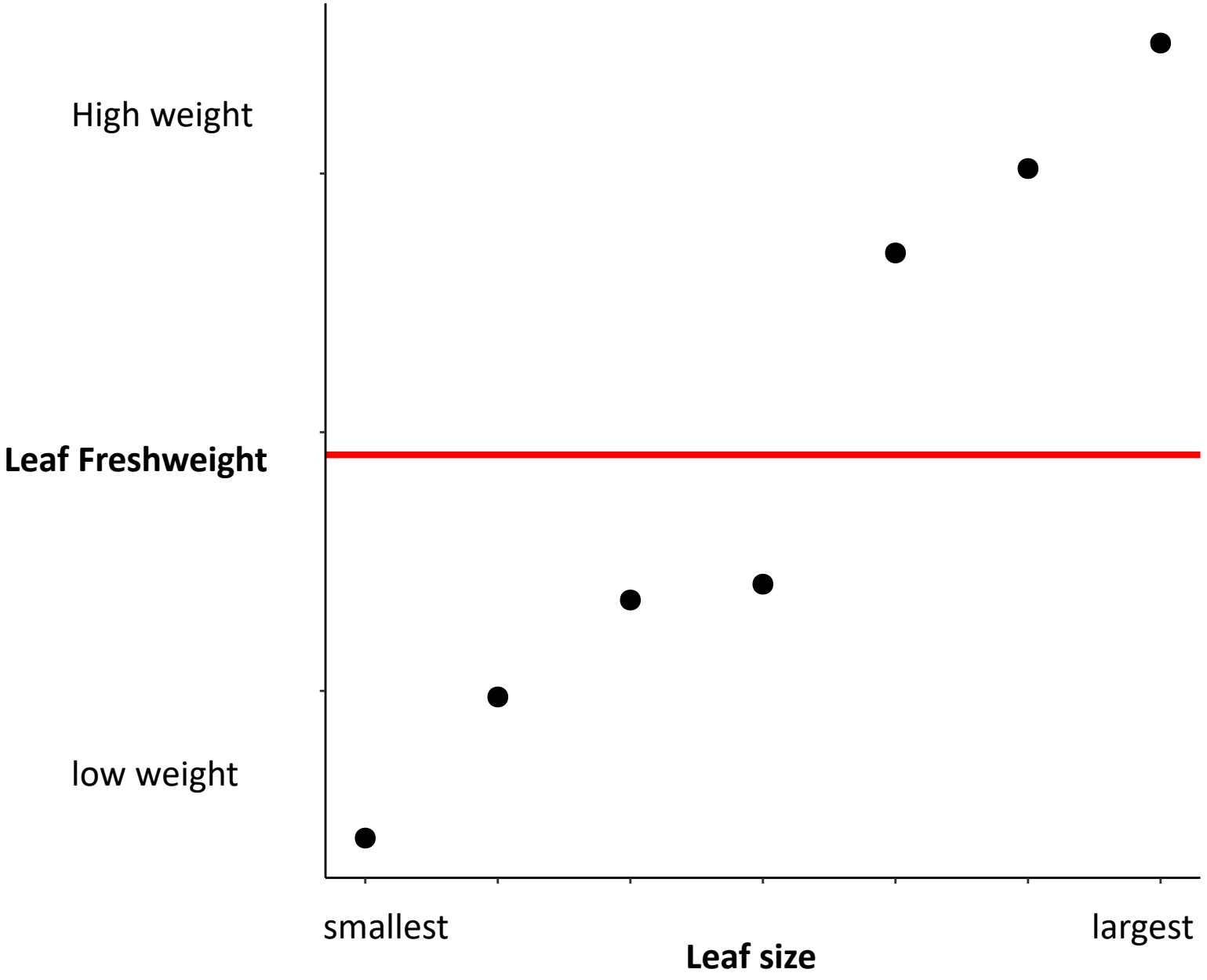
All we have done is reorder the data on the X-axis.
The mean and variation are the exact same as before.



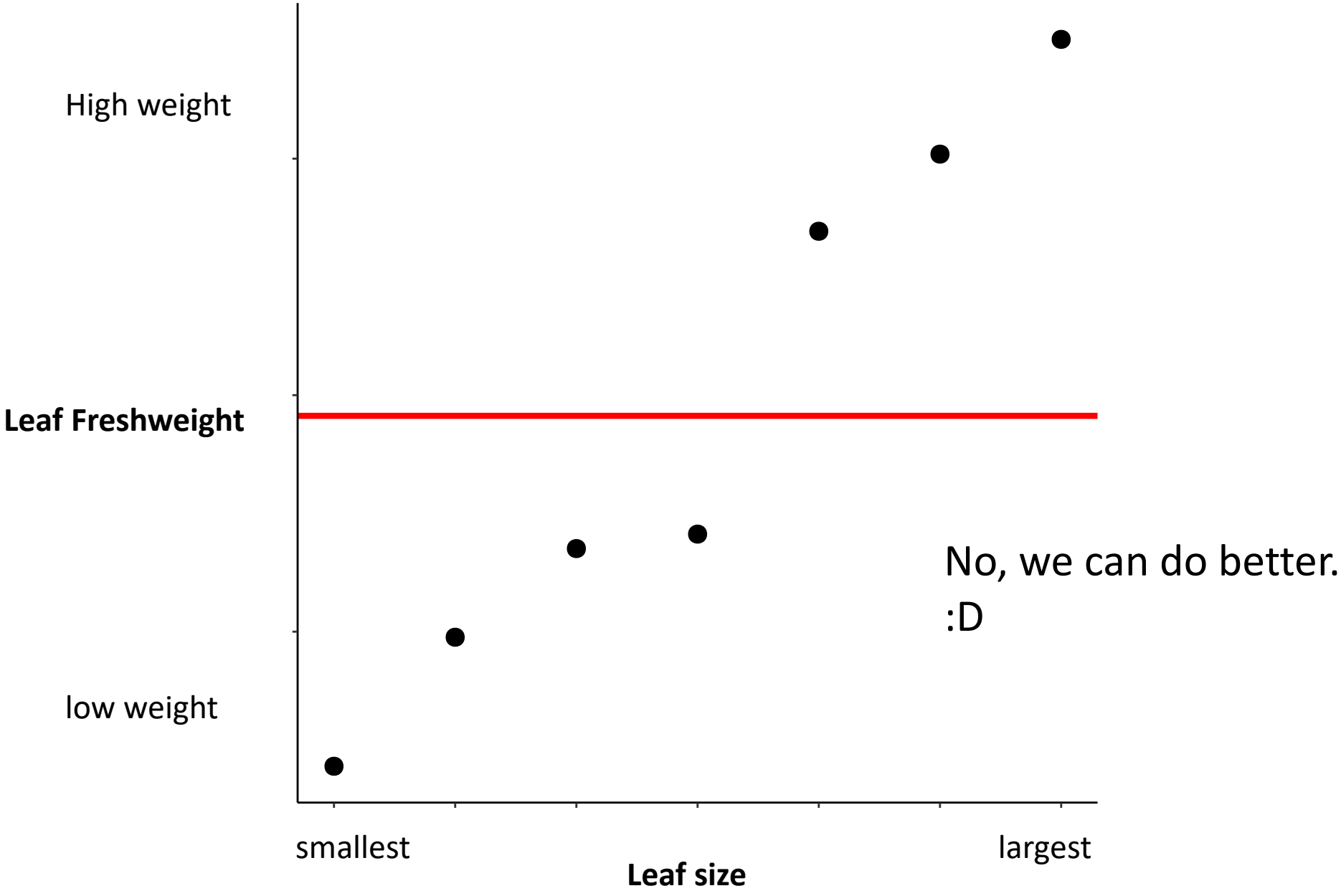
All we have done is reorder the data on the X-axis.
The mean and variation are the exact same as before.



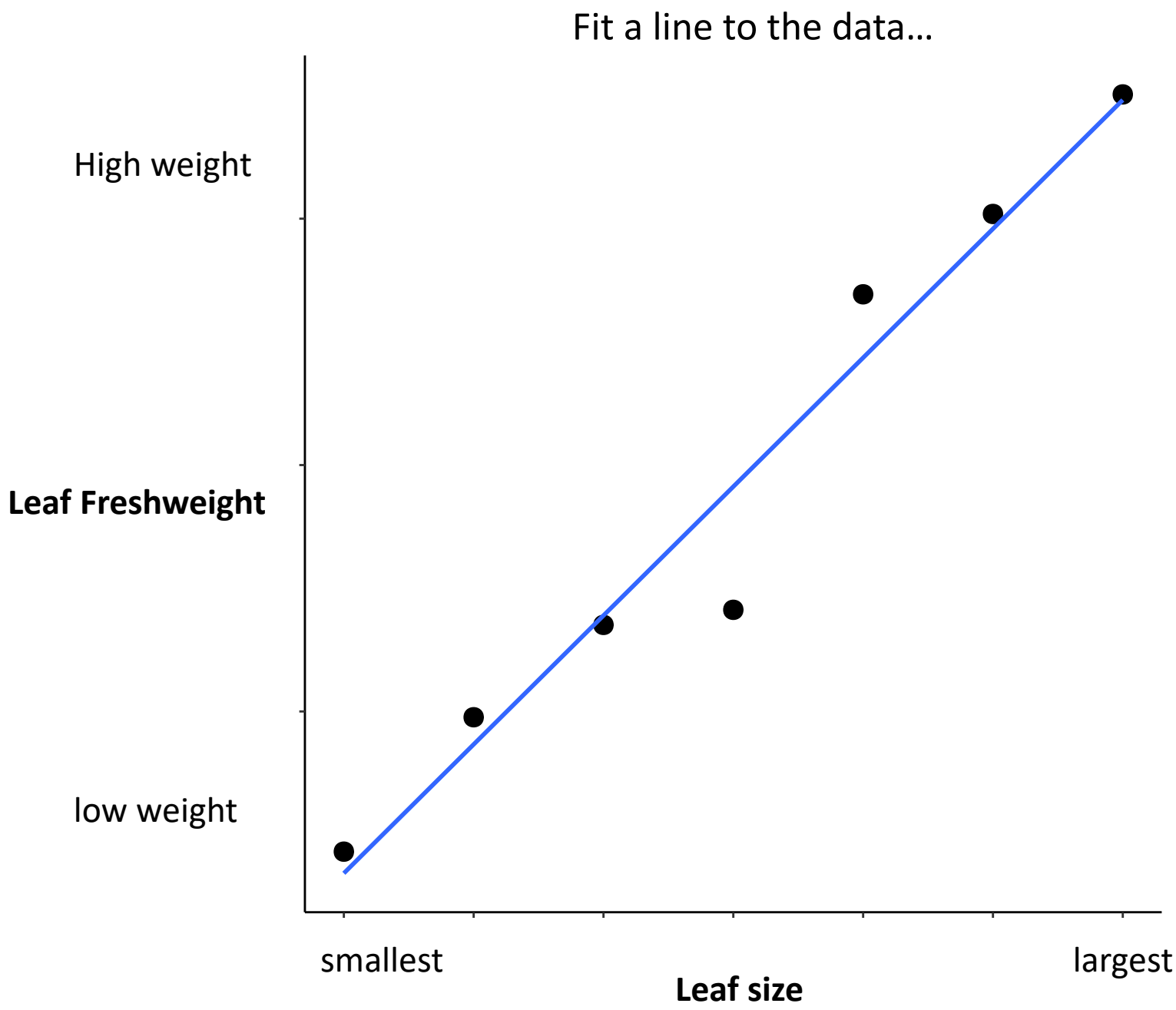
Question: Given that we know the leaf sizes,
Is the mean weight the best way to predict leaf weight?

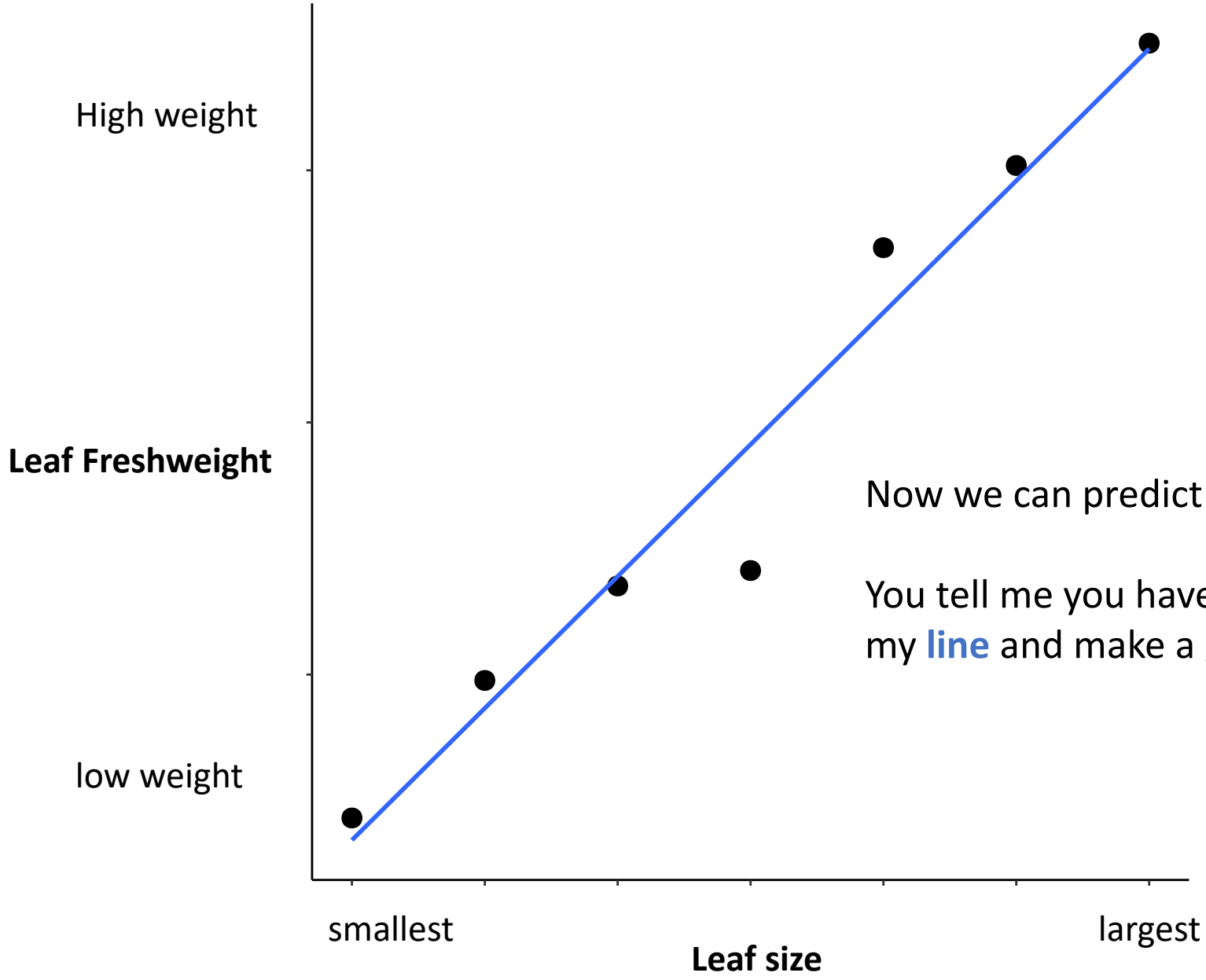


Question: Given that we know the leaf sizes,
Is the mean weight the best way to predict leaf weight?



No, we can do better.
:D

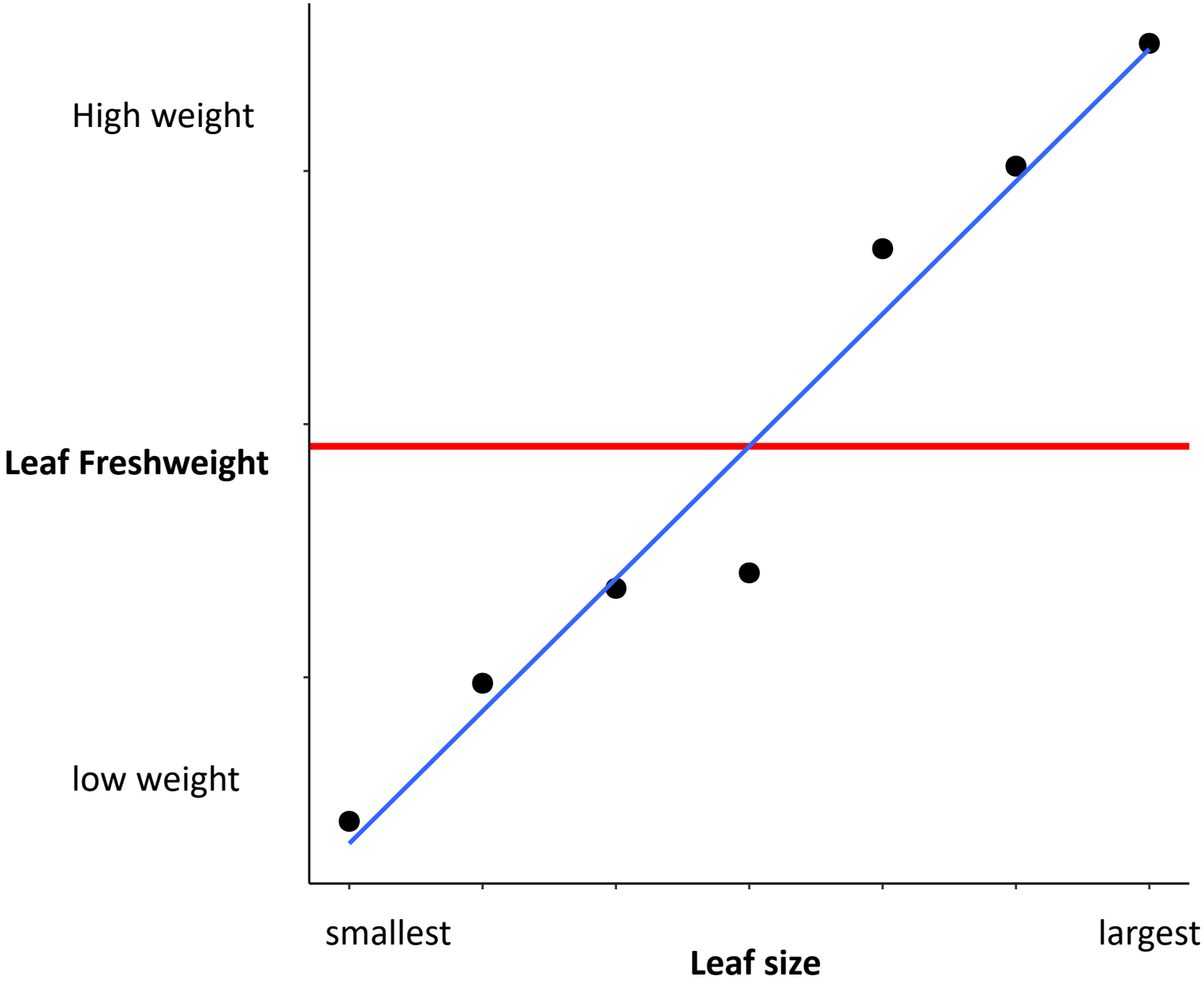

Fit a line to the data…
High weight
Leaf Freshweight
low weight
smallest
largest
Leaf size



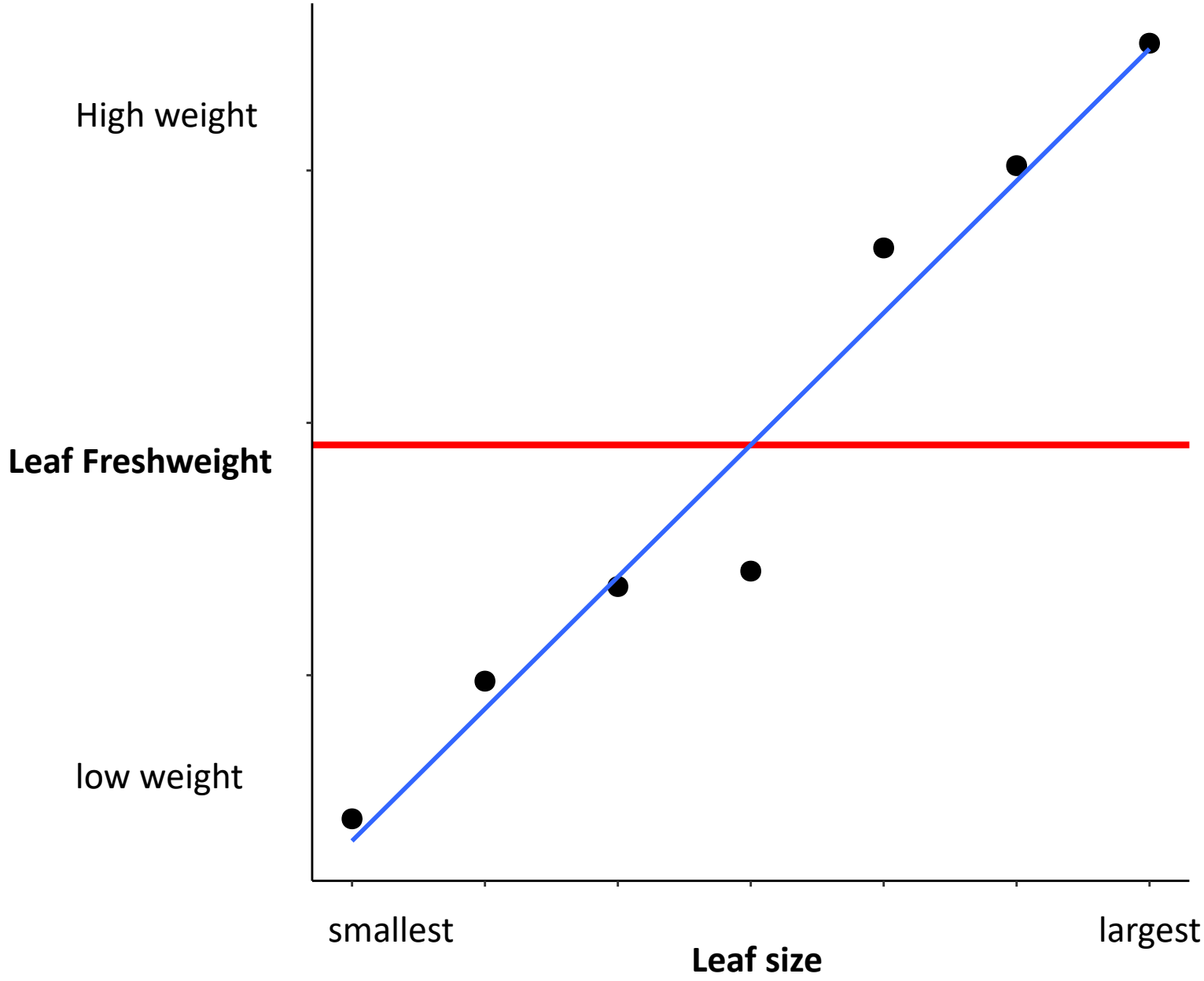
Now we can predict weight with our **line**.

You tell me you have a large leaf, I can look at my **line** and make a good guess about its weight.

Question: Does the **line** fit the data better than the **mean**?
If so, how much better?

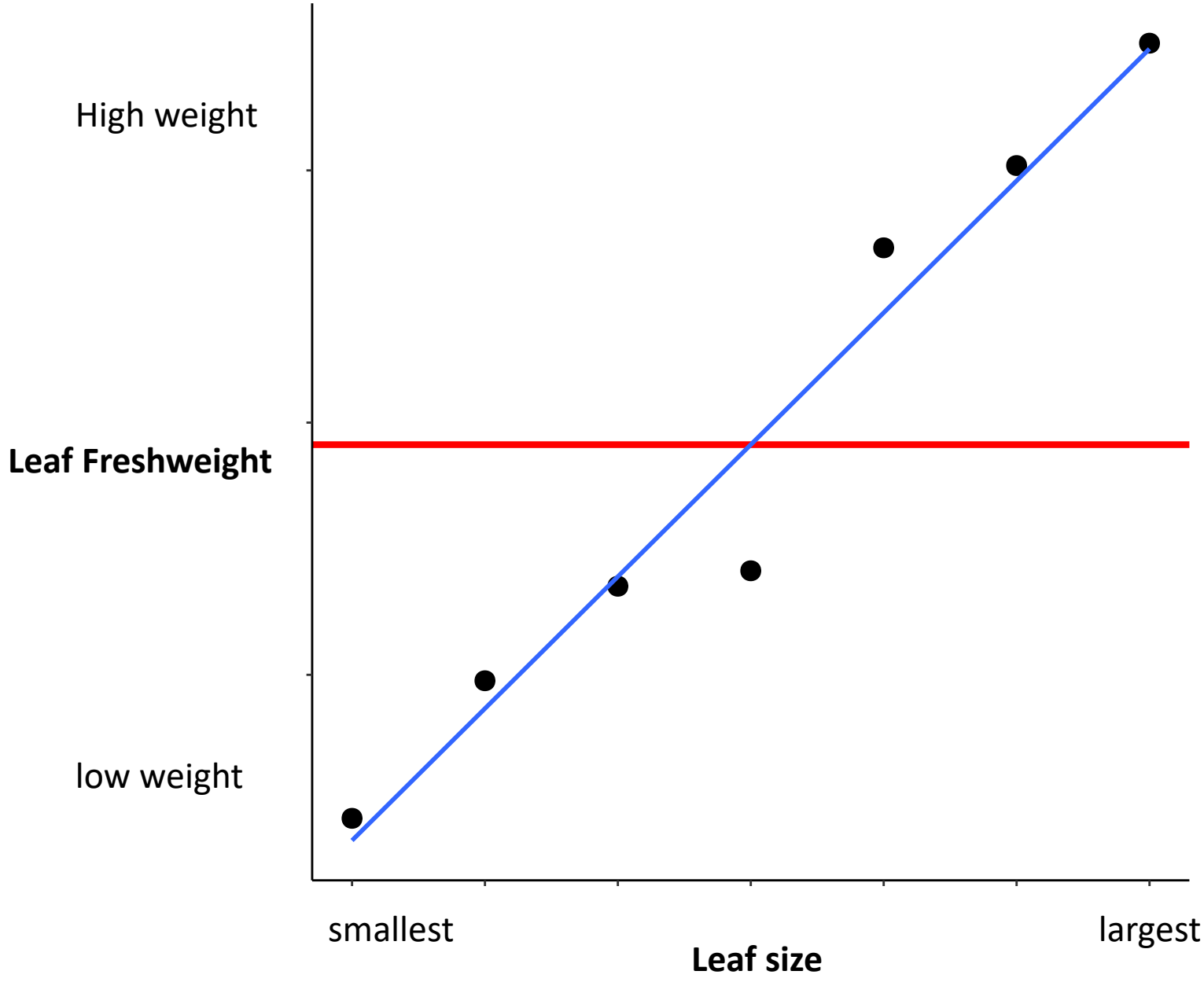


Question: Does the **line** fit the data better than the **mean**?
If so, how much better?



By eye, it looks like the **line** fits the data better than the man.

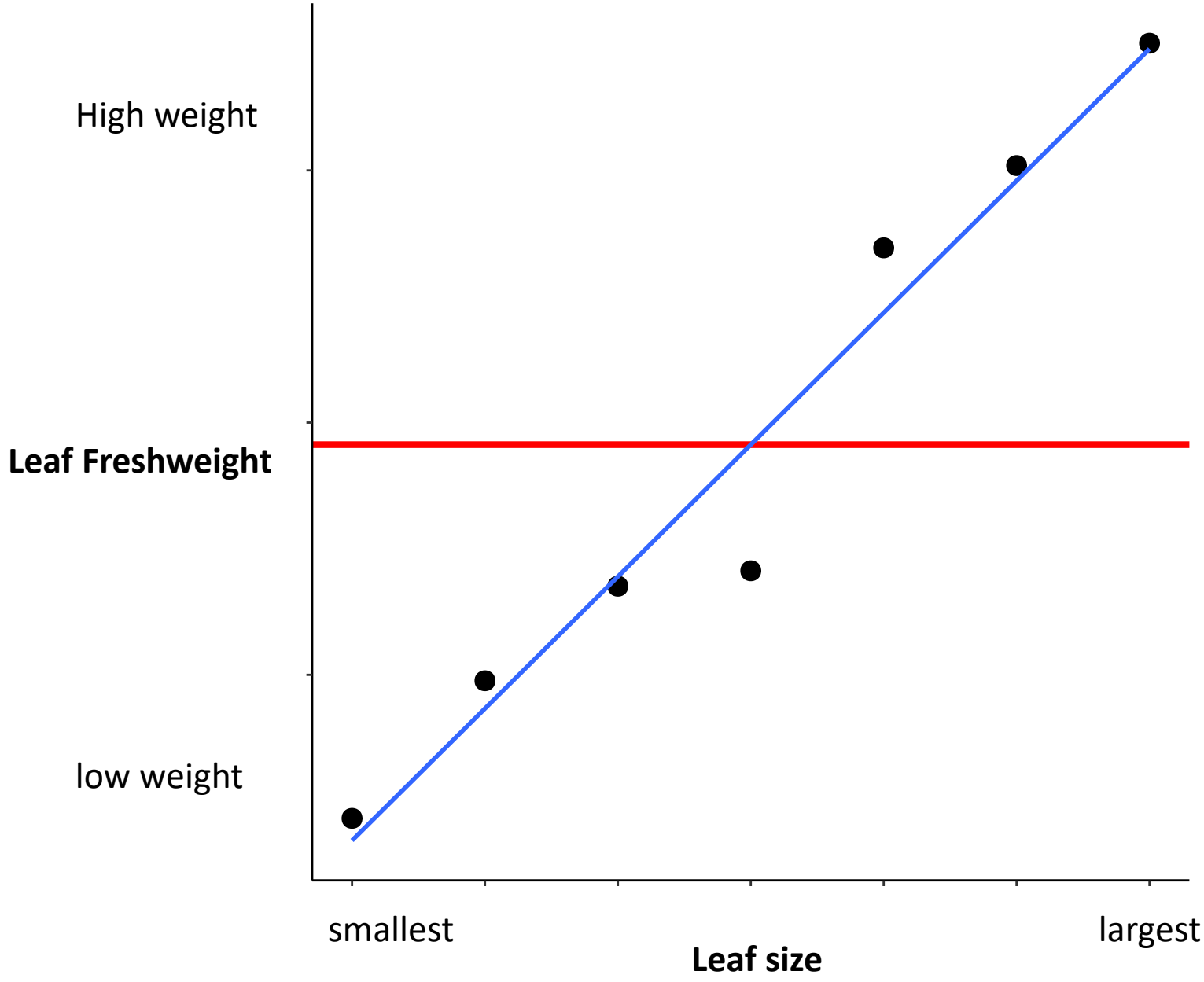
Question: Does the **line** fit the data better than the **mean**?
If so, how much better?



By eye, it looks like the **line** fits the data better than the man.

How do we quantify that difference?

Question: Does the **line** fit the data better than the **mean**?
If so, how much better?

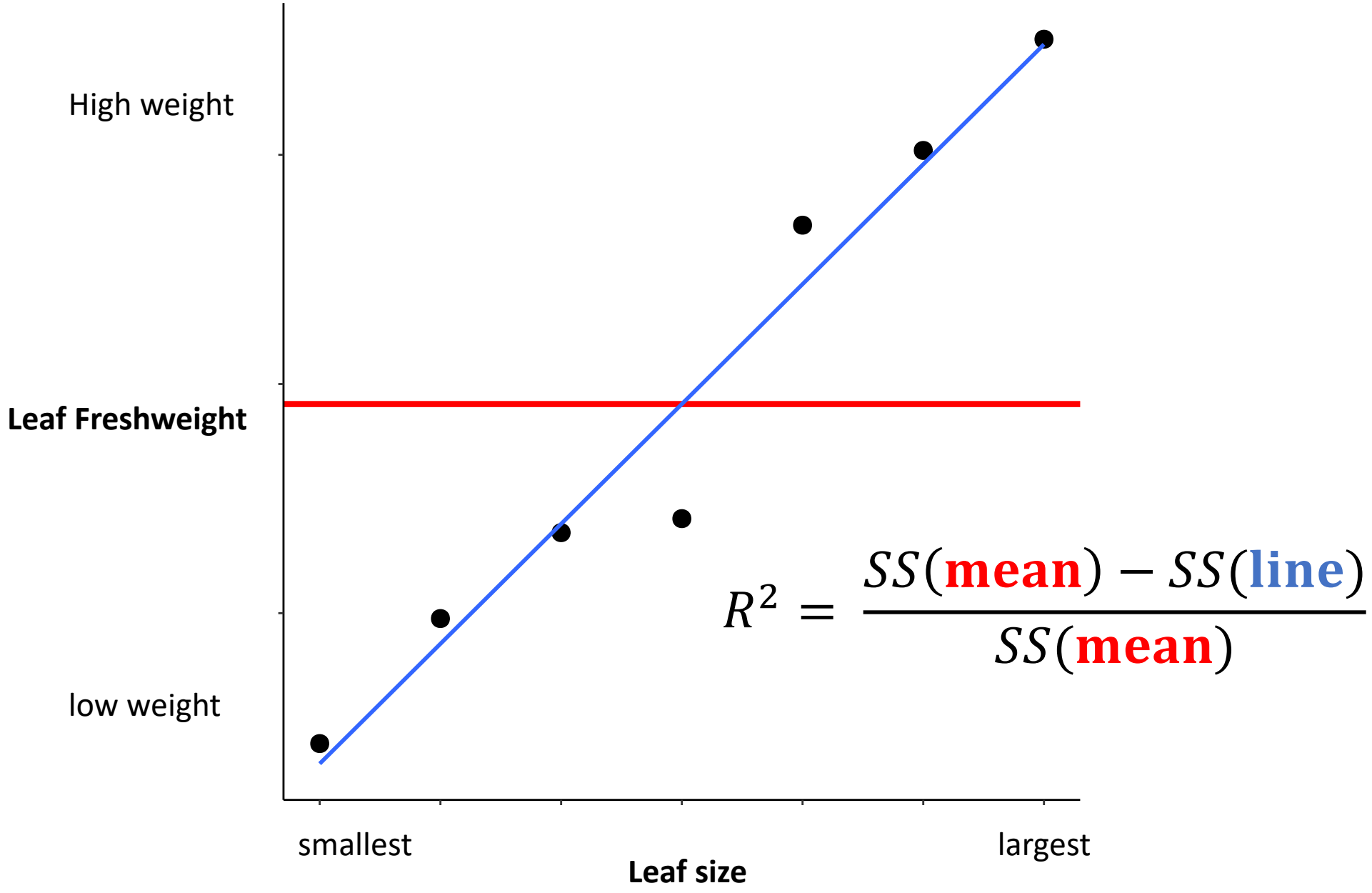


By eye, it looks like the **line** fits the data better than the man.

How do we quantify that difference?

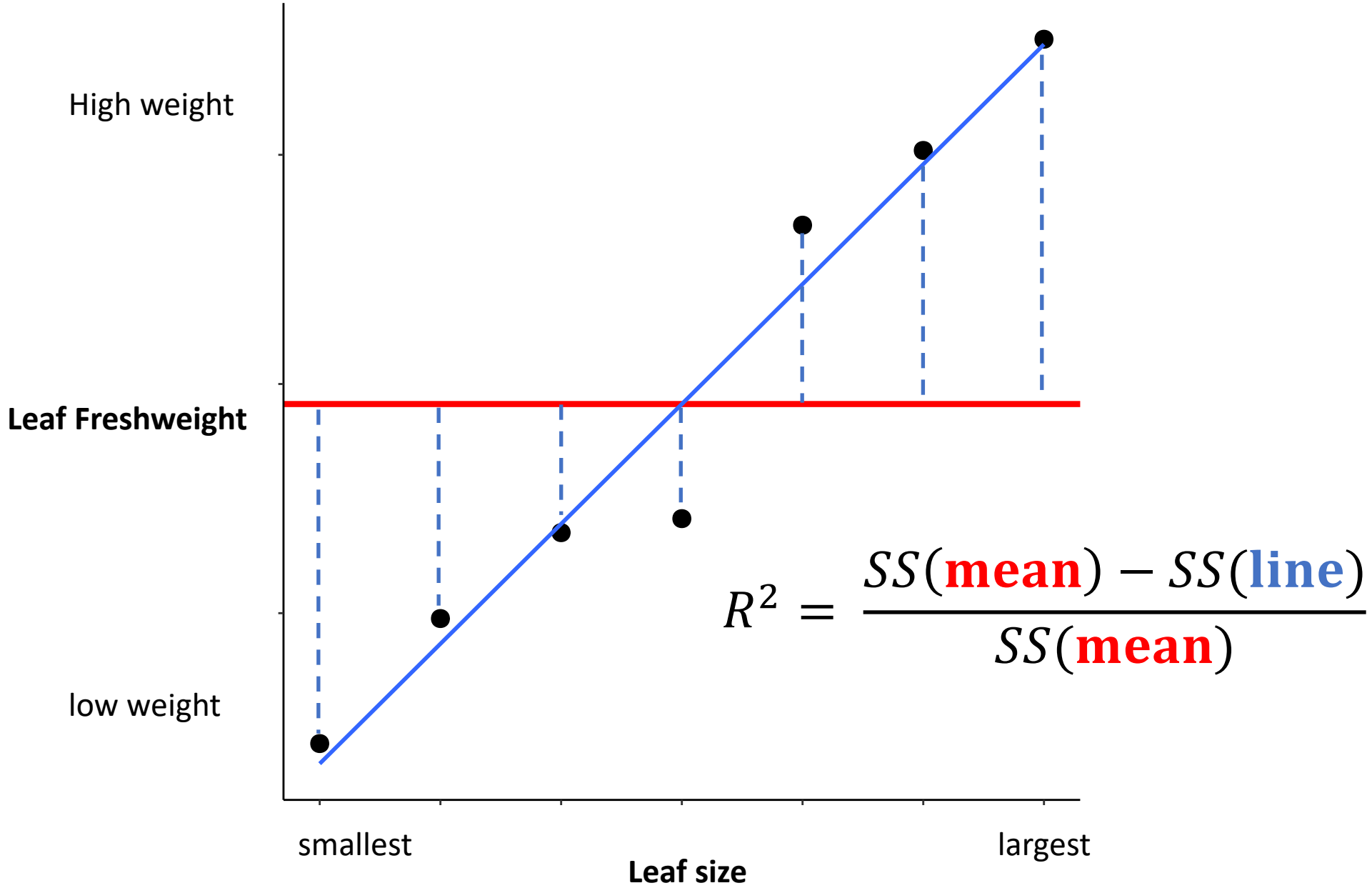
R^2

Quantifying the difference between the **line** and the mean.
i.e. Calculating R^2

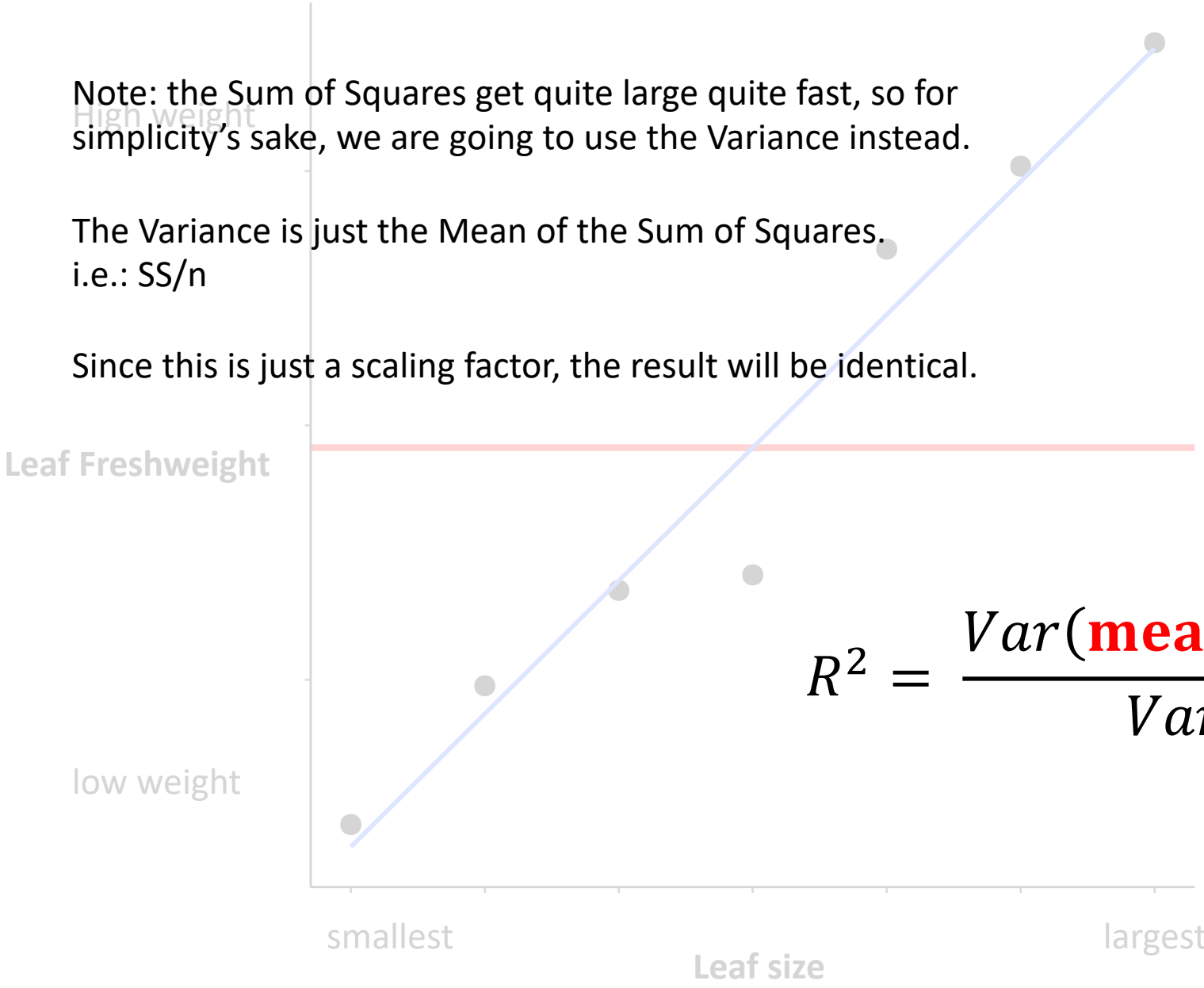


$$R^2 = \frac{SS(\mathbf{mean}) - SS(\mathbf{line})}{SS(\mathbf{mean})}$$

Quantifying the difference between the **line** and the mean.
i.e. Calculating R^2



Quantifying the difference between the line and the mean.
i.e. Calculating R^2



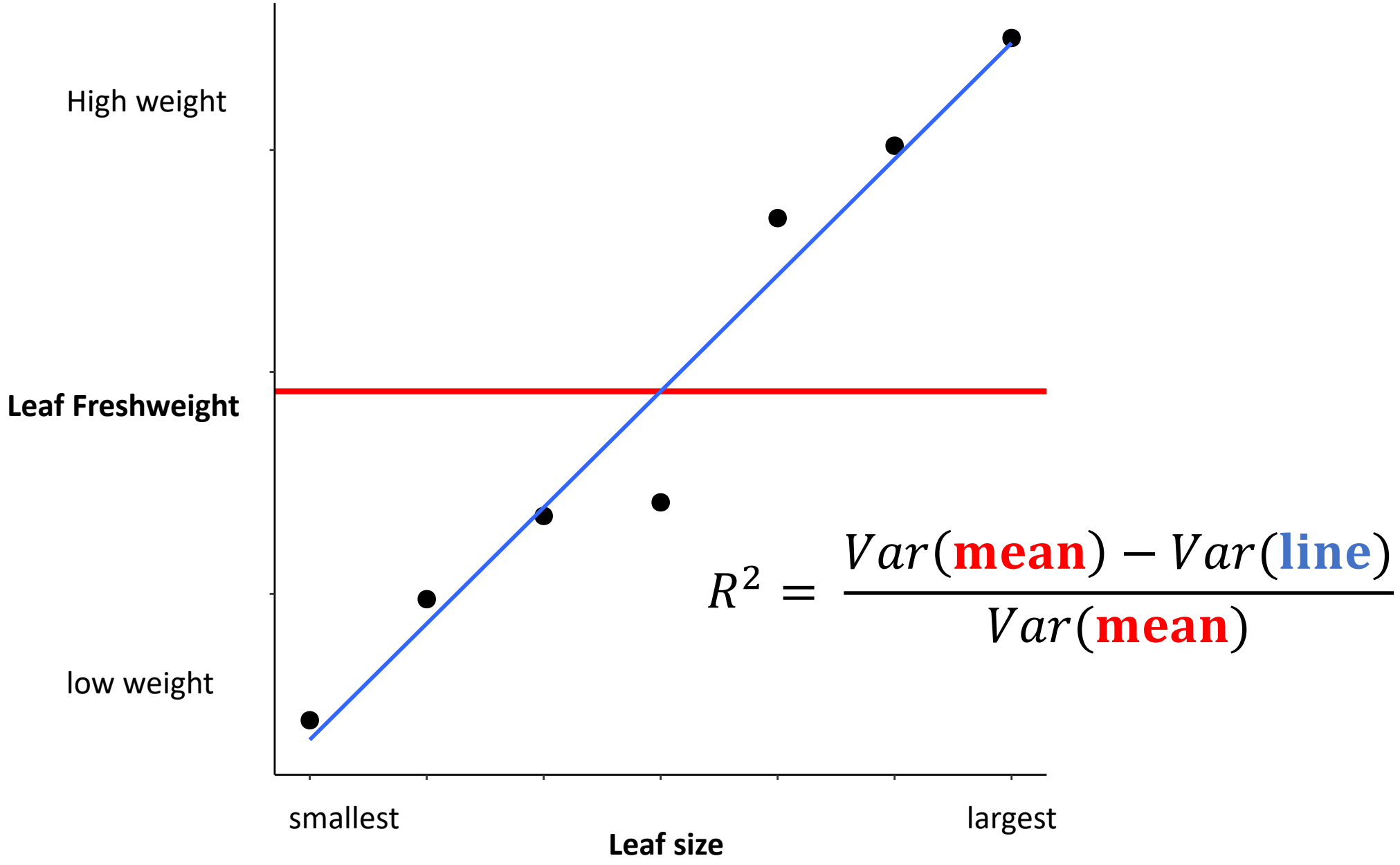
Note: the Sum of Squares get quite large quite fast, so for simplicity's sake, we are going to use the Variance instead.

The Variance is just the Mean of the Sum of Squares.
i.e.: SS/n

Since this is just a scaling factor, the result will be identical.

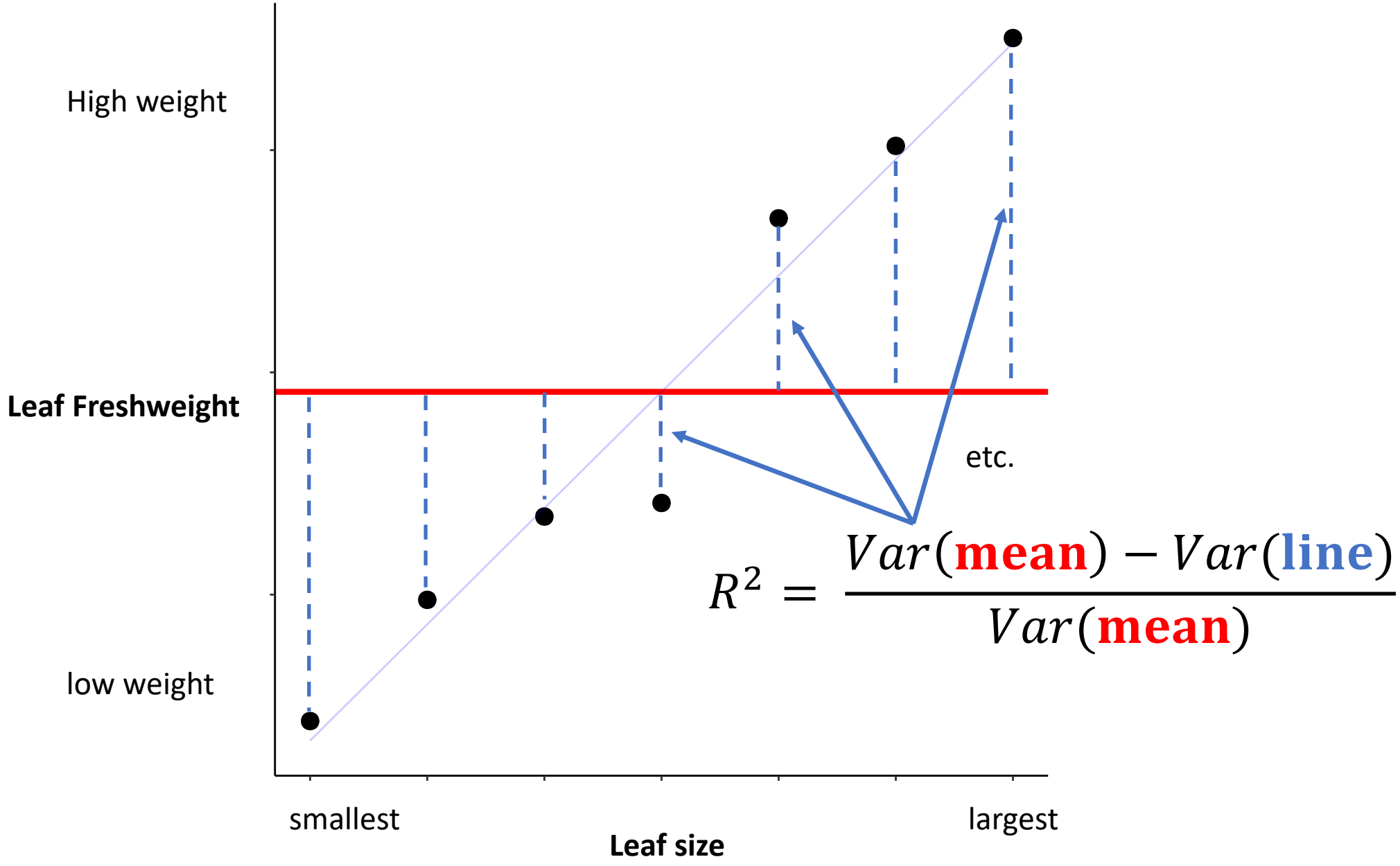
$$R^2 = \frac{Var(\mathbf{mean}) - Var(\mathbf{line})}{Var(\mathbf{mean})}$$

Quantifying the difference between the **line** and the mean.
i.e. Calculating R^2

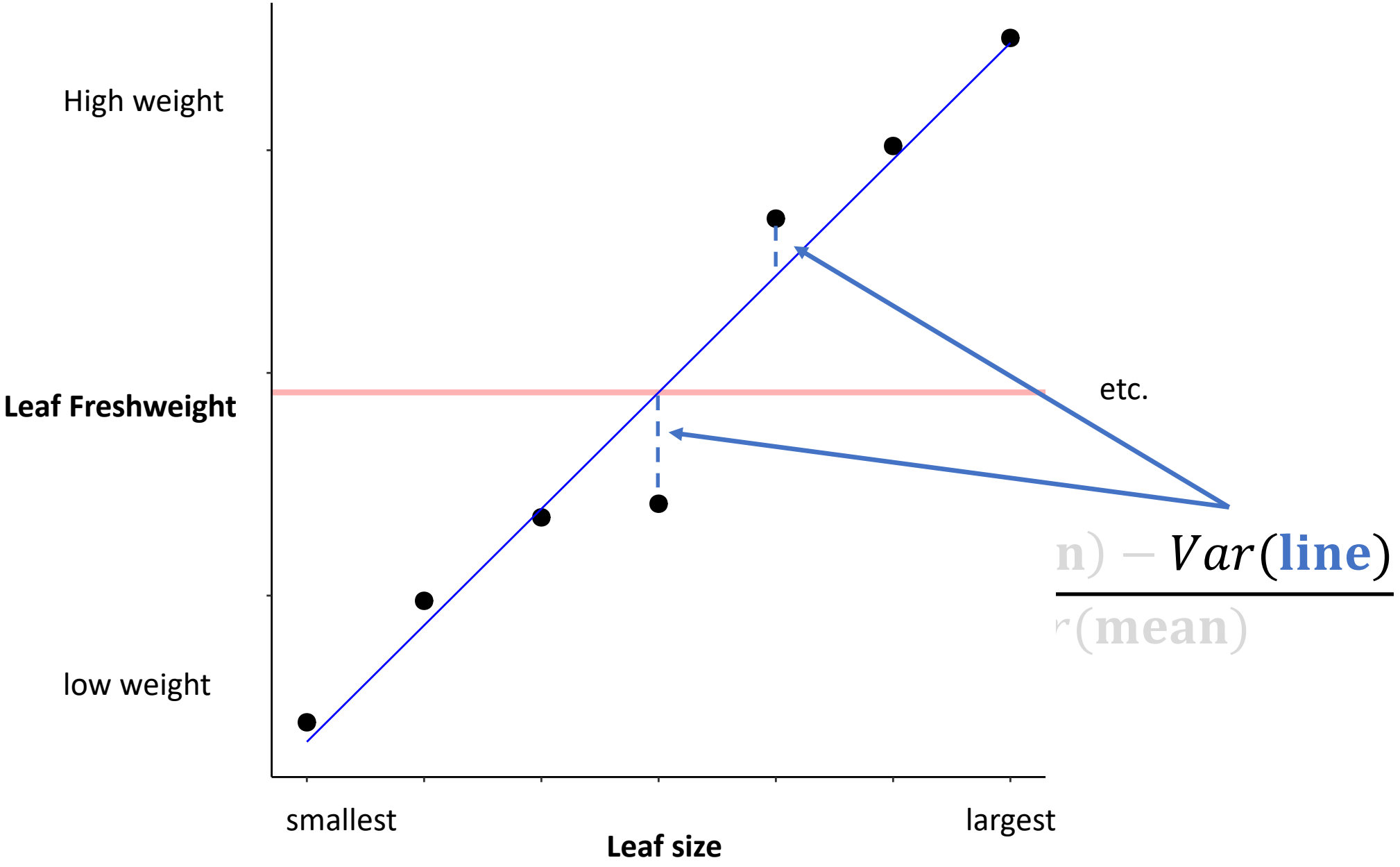


$$R^2 = \frac{Var(\mathbf{mean}) - Var(\mathbf{line})}{Var(\mathbf{mean})}$$

Quantifying the difference between the **line** and the mean.
i.e. Calculating R^2

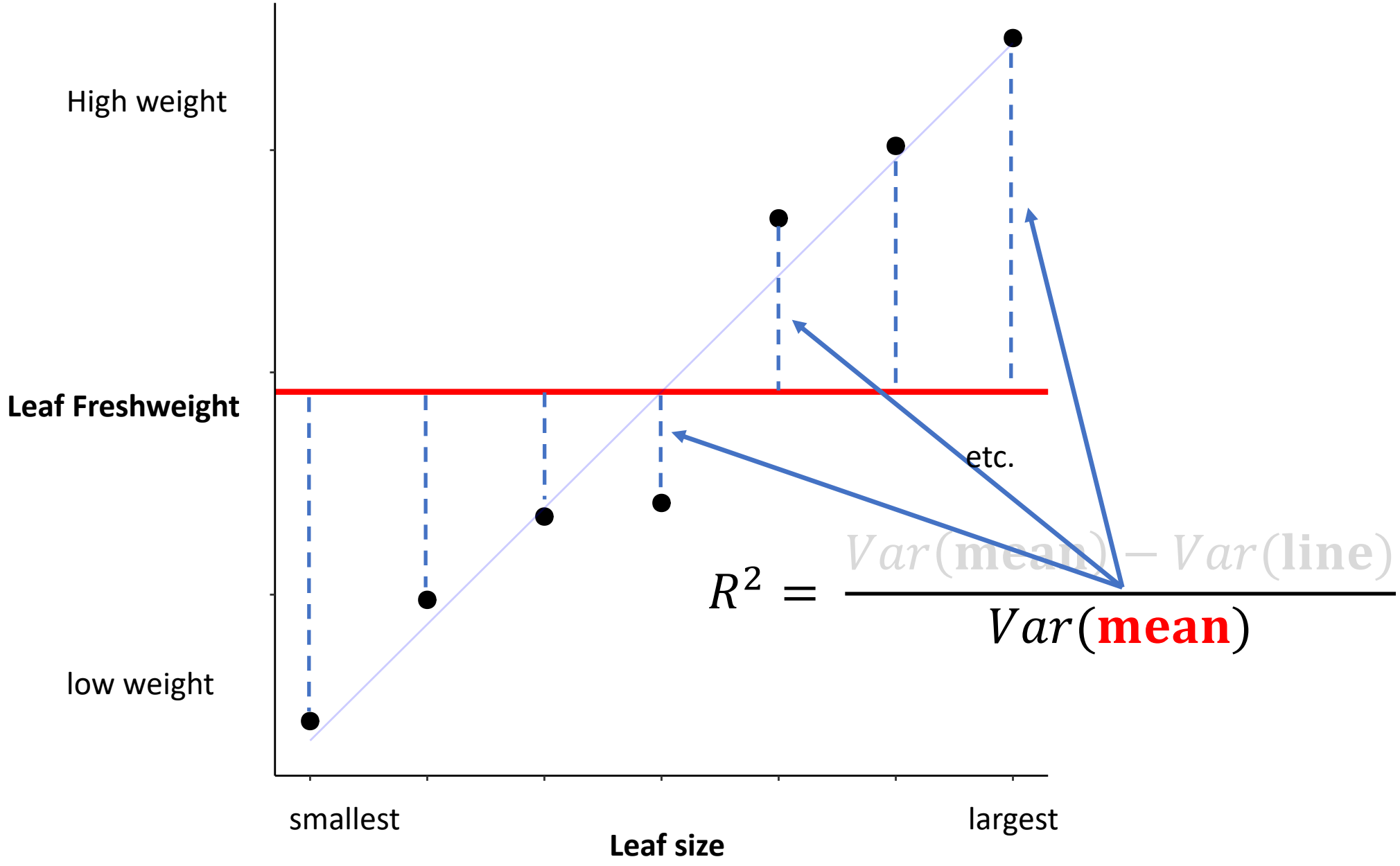


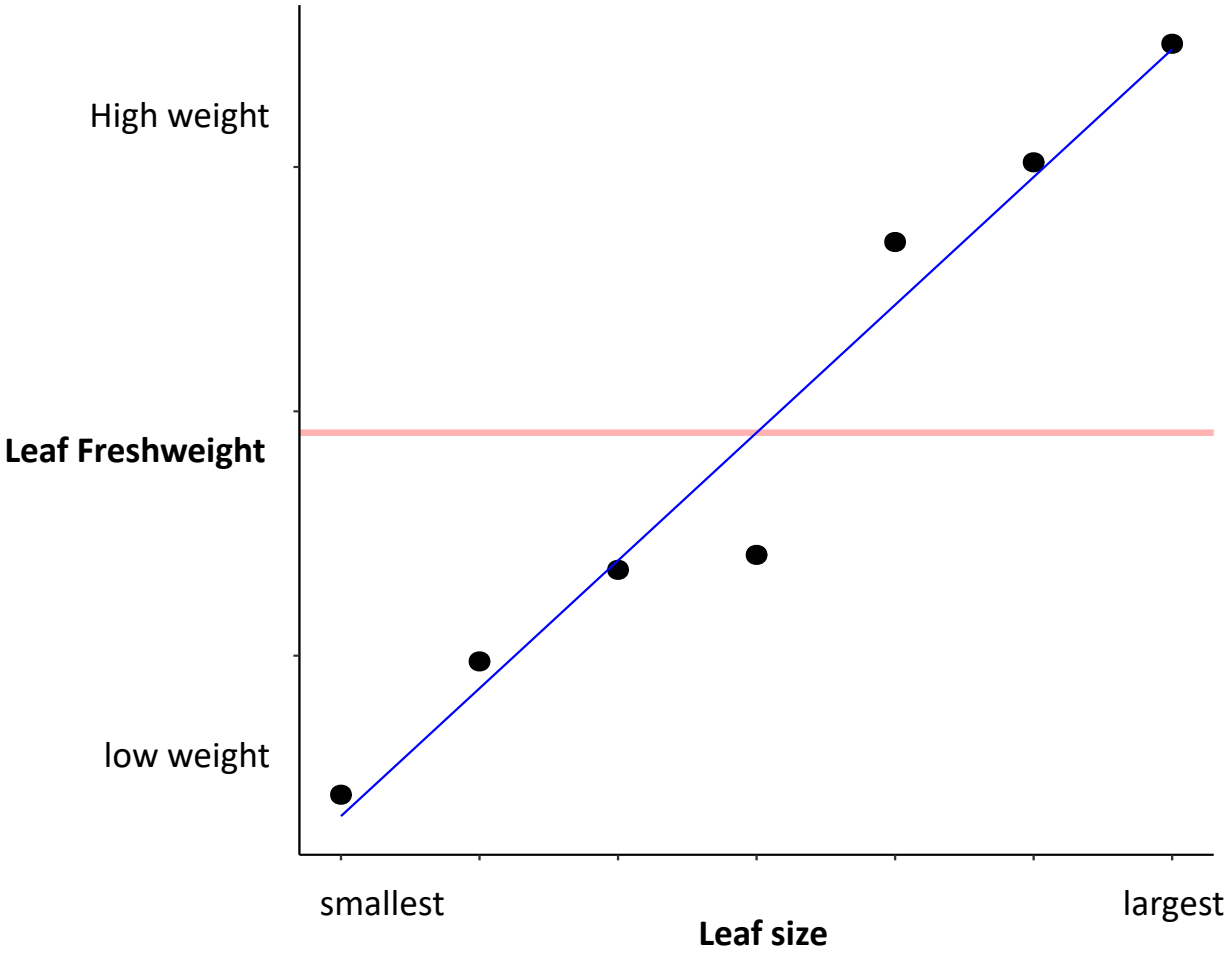
Quantifying the difference between the **line** and the mean.
i.e. Calculating R^2



$$\frac{\text{n}) - Var(\textbf{line})}{\text{r}(\textbf{mean})}$$

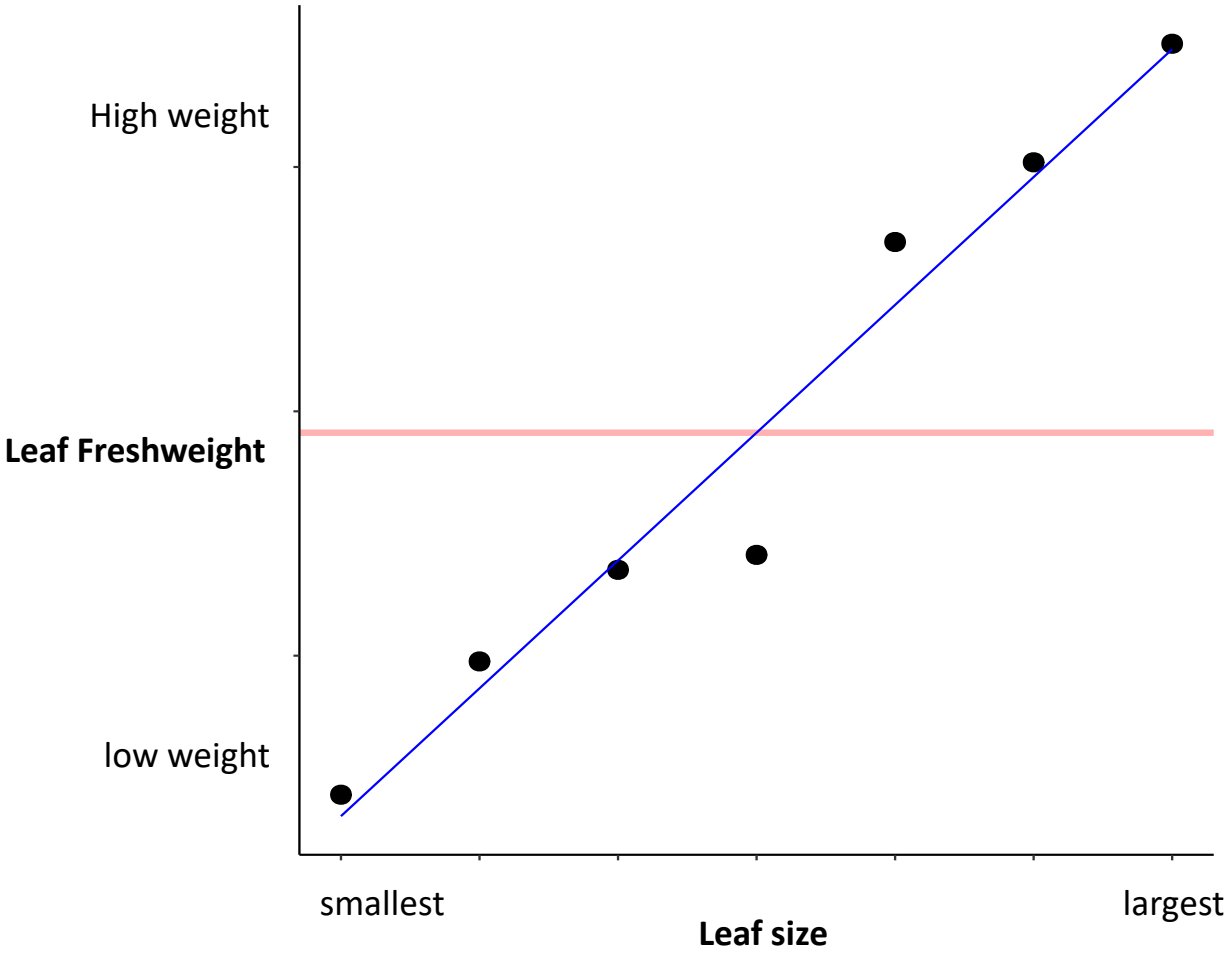
Quantifying the difference between the **line** and the mean.
i.e. Calculating R^2



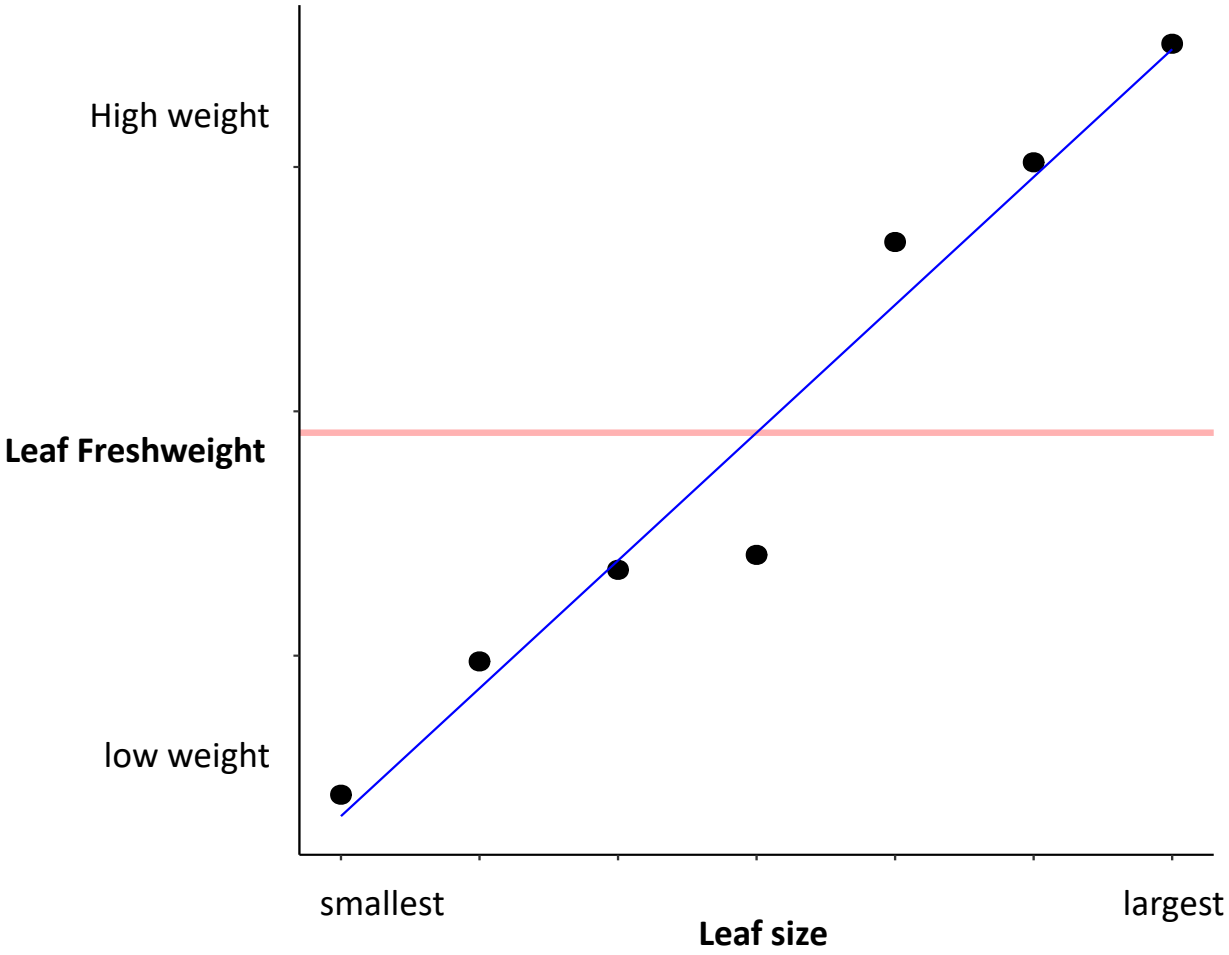


Var(**mean**) = 59.38

Var(**line**) = 7.74



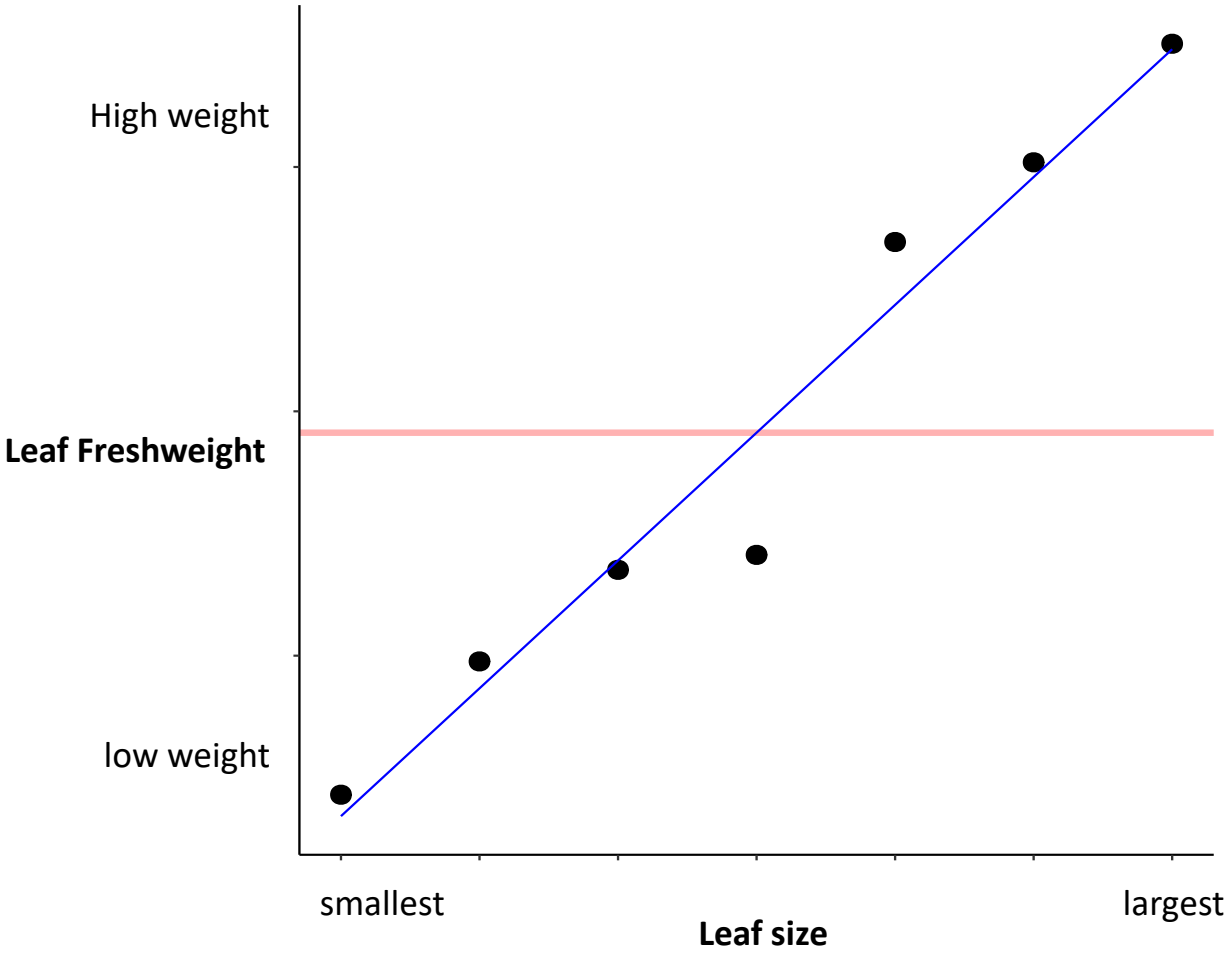
Var(**mean**) = 59.38



Var(**mean**) = 59

Var(**line**) = 8

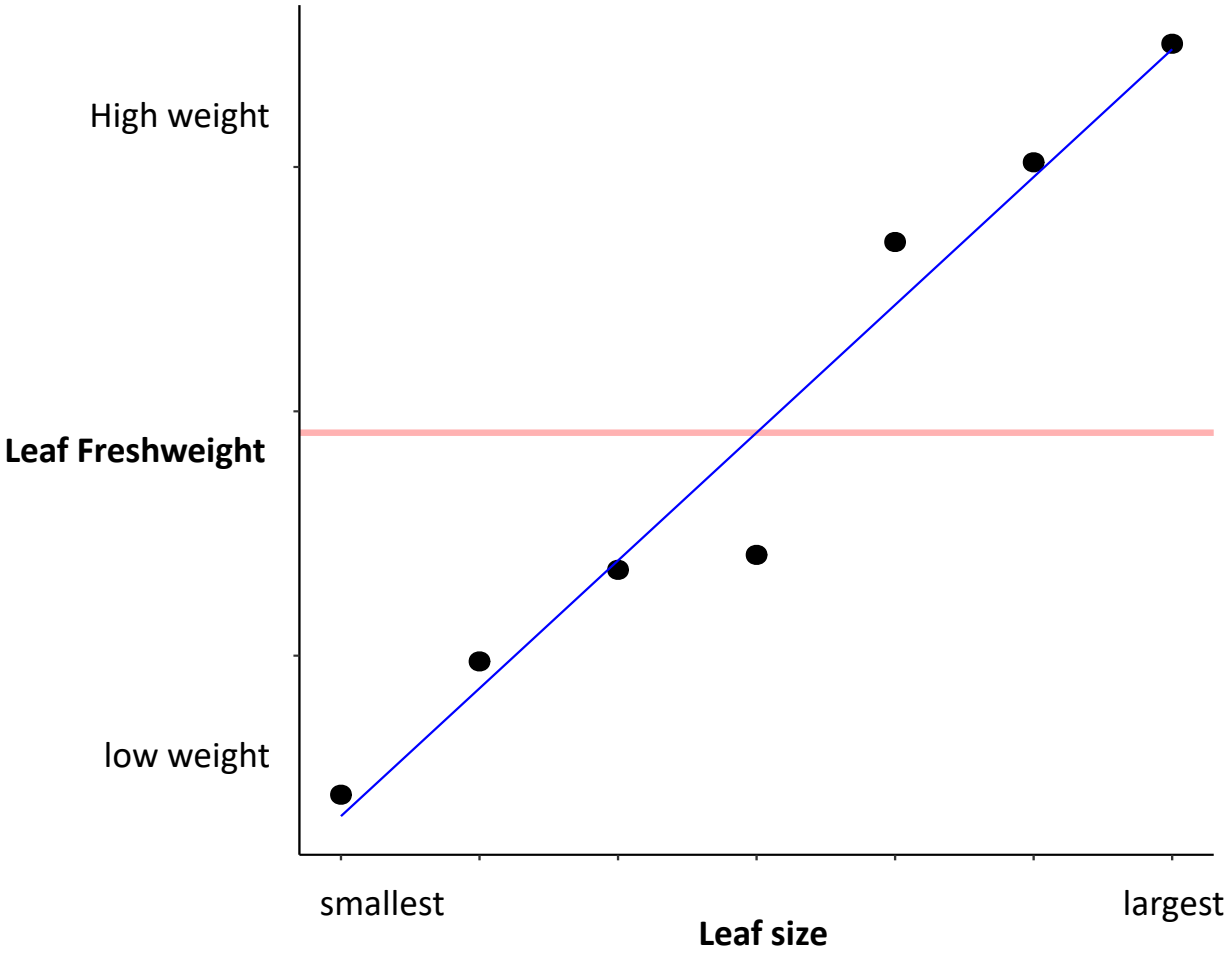
$$R^2 = \frac{Var(\textbf{mean}) - Var(\textbf{line})}{Var(\textbf{mean})}$$



Var(**mean**) = 59

Var(**line**) = 8

$$R^2 = \frac{\mathbf{59} - \mathbf{8}}{\mathbf{59}}$$

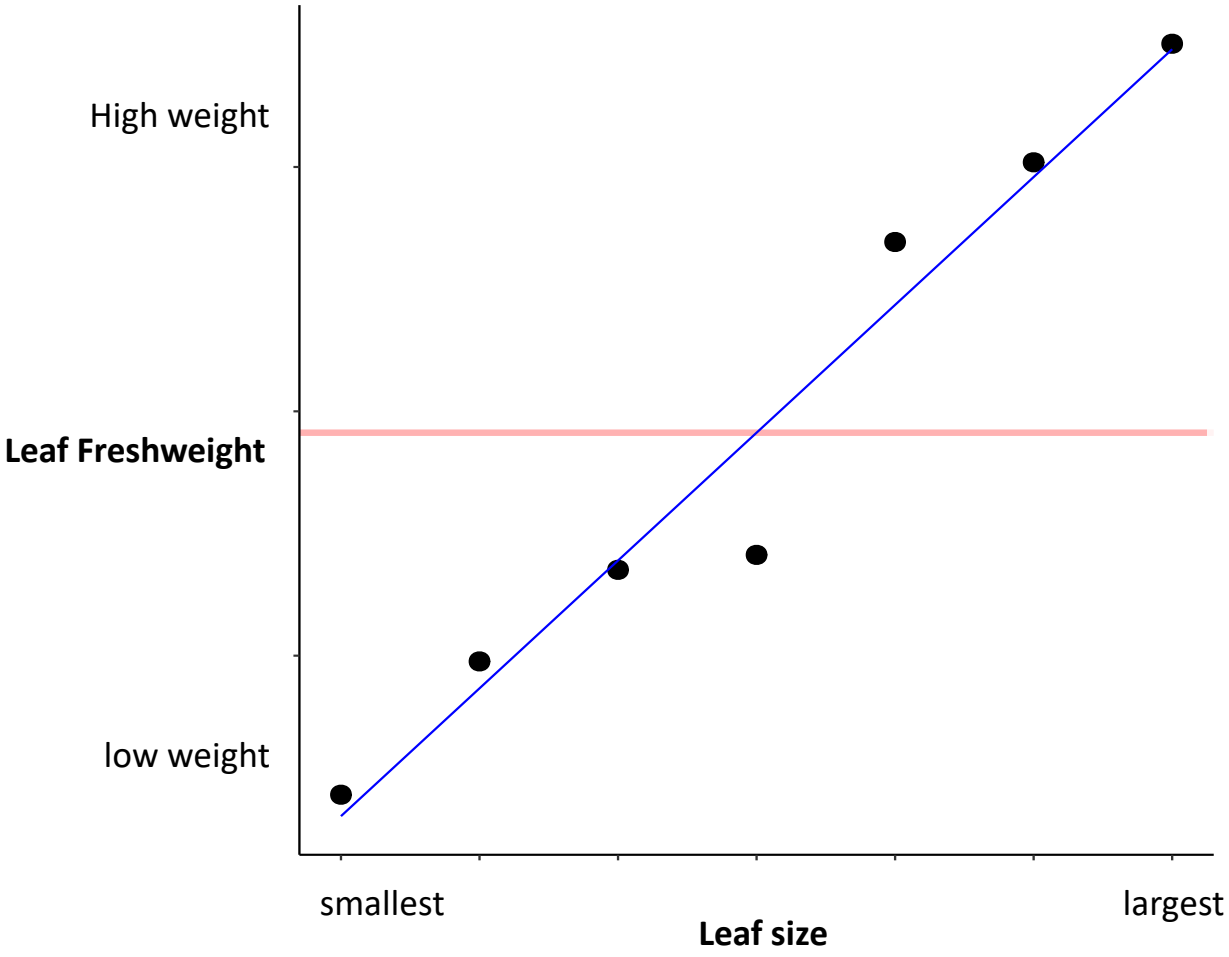


Var(**mean**) = 59

Var(**line**) = 8

$$R^2 = \frac{\mathbf{59} - \mathbf{8}}{\mathbf{59}}$$

$$R^2 = \frac{51}{\mathbf{59}} = 0.86 = 86\%$$



Var(**mean**) = 59

Var(**line**) = 8

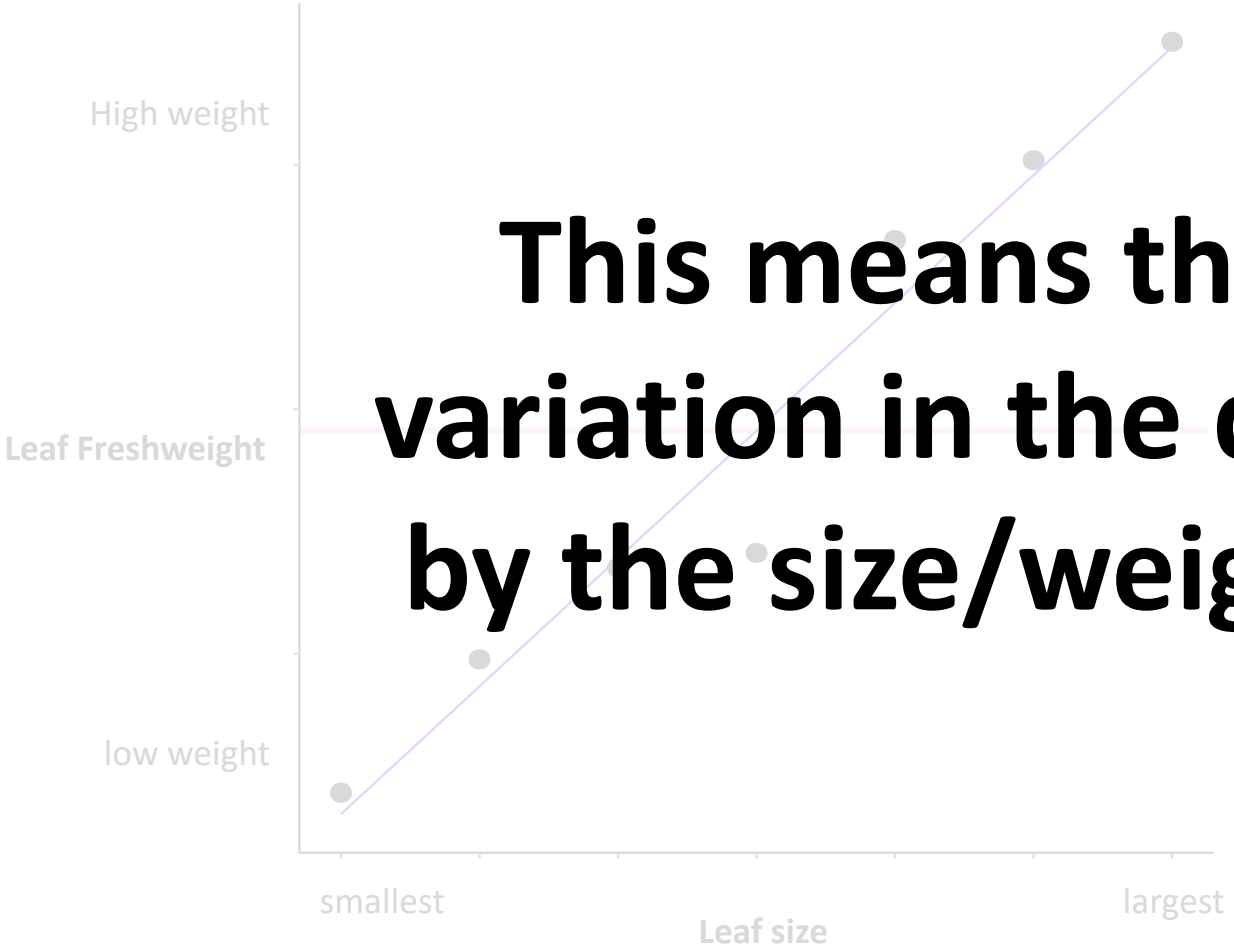
$$R^2 = \frac{59 - 8}{59}$$

$$R^2 = \frac{51}{59} = 0.86 = 86\%$$

There is 86% less variation around the **line** than the **mean**.

...or...

The size/weight relationship accounts for 81% of the variation



This means that most of the variation in the data is explained by the size/weight relationship.

Var(**mean**) = 59

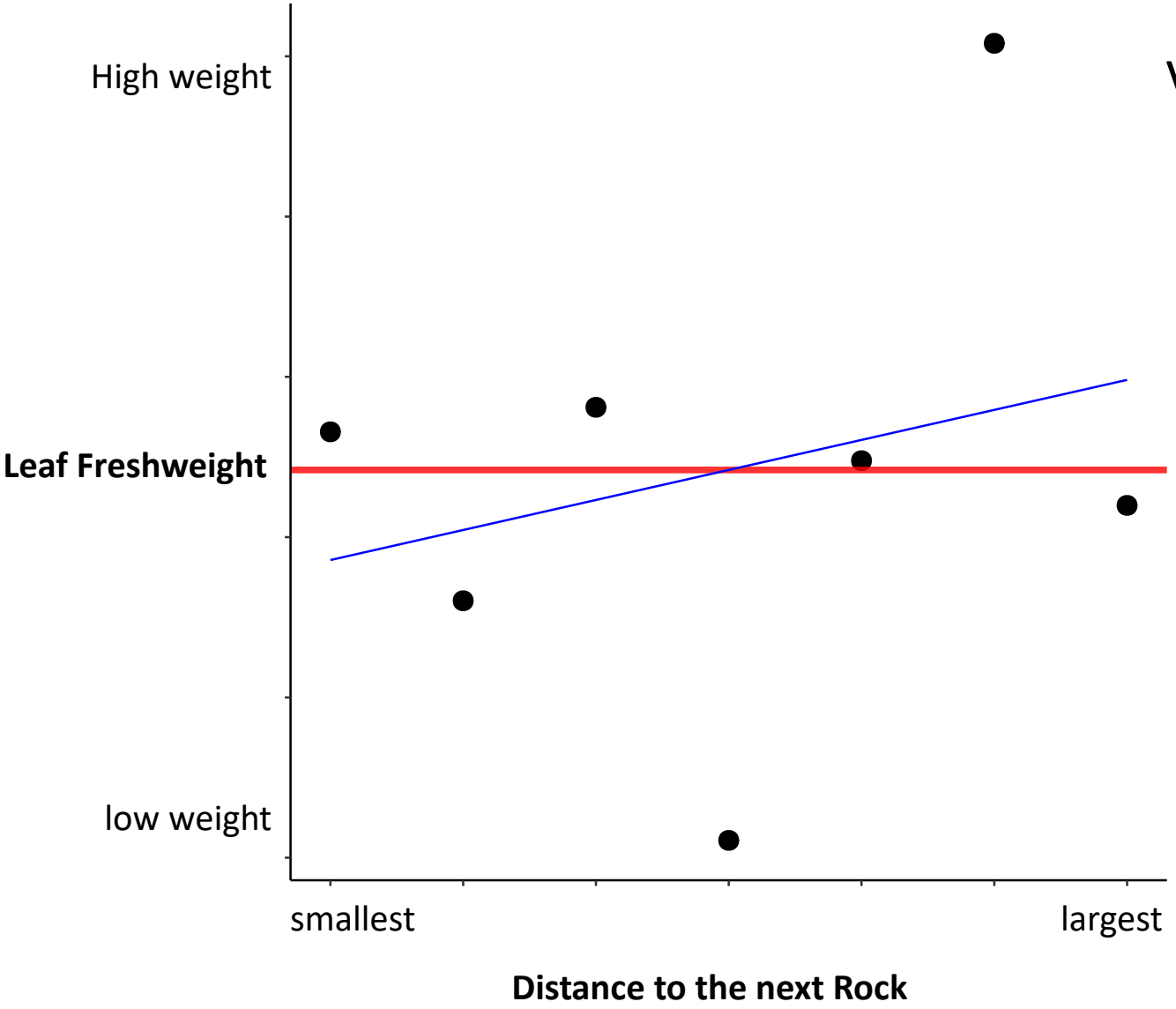
Var(**line**) = 8

$$R^2 = \frac{59 - 8}{59}$$

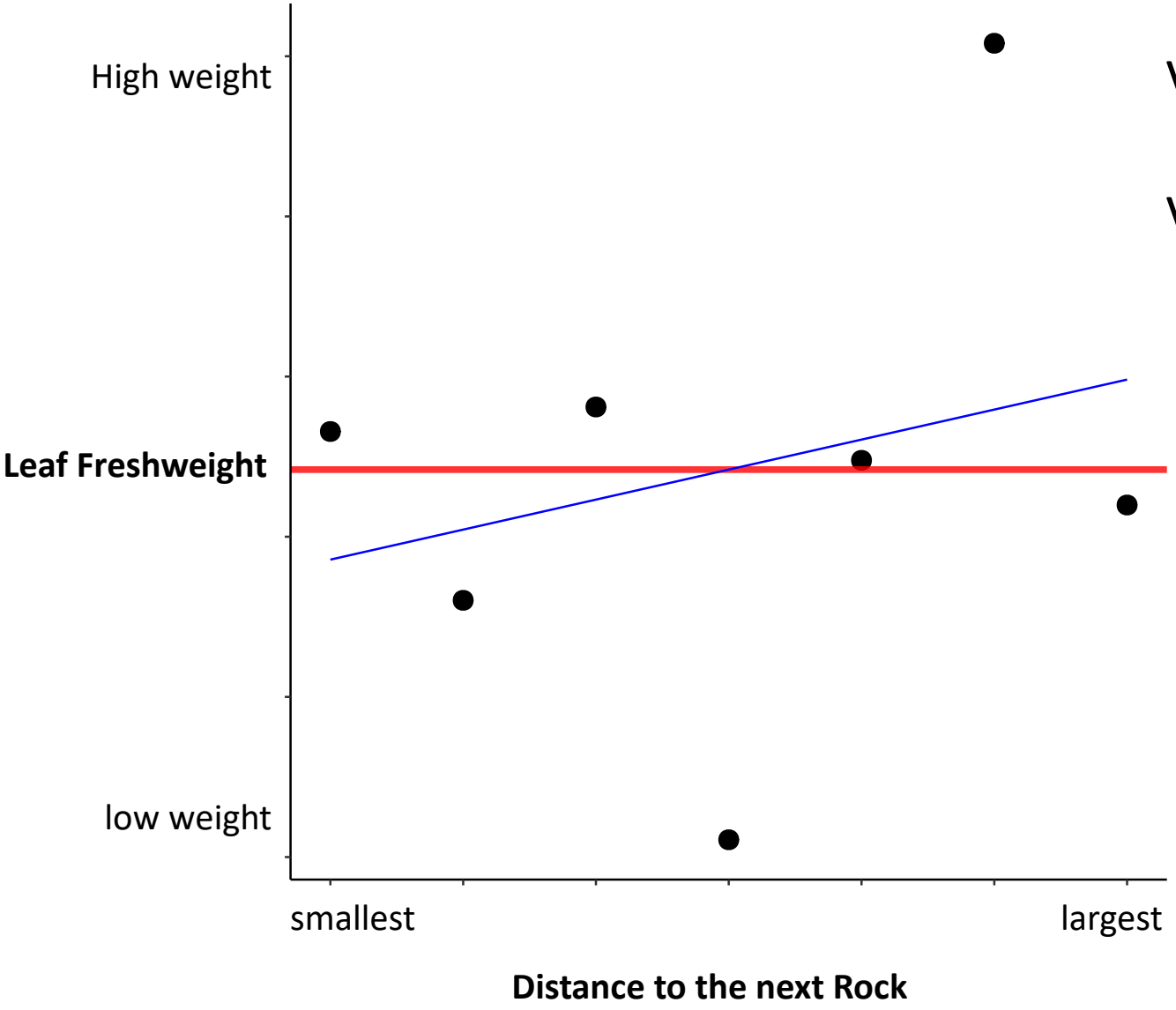
There is 86% less variation around the **line** than the **mean**.

...or...

The size/weight relationship accounts for 81% of the variation

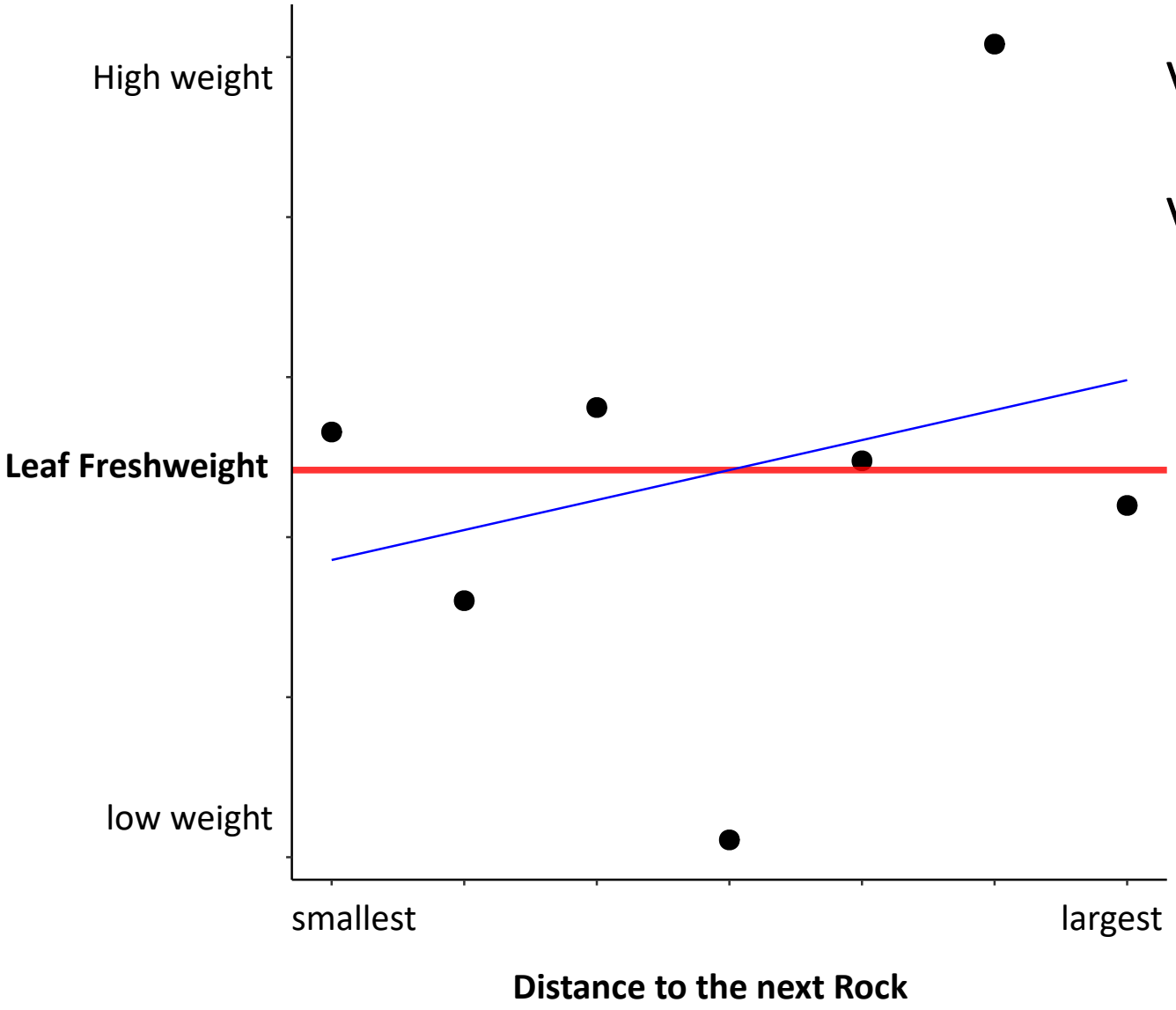

High weight
Leaf Freshweight
low weight
smallest
largest
Distance to the next Rock

Var(**mean**) = 59



Var(**mean**) = 59

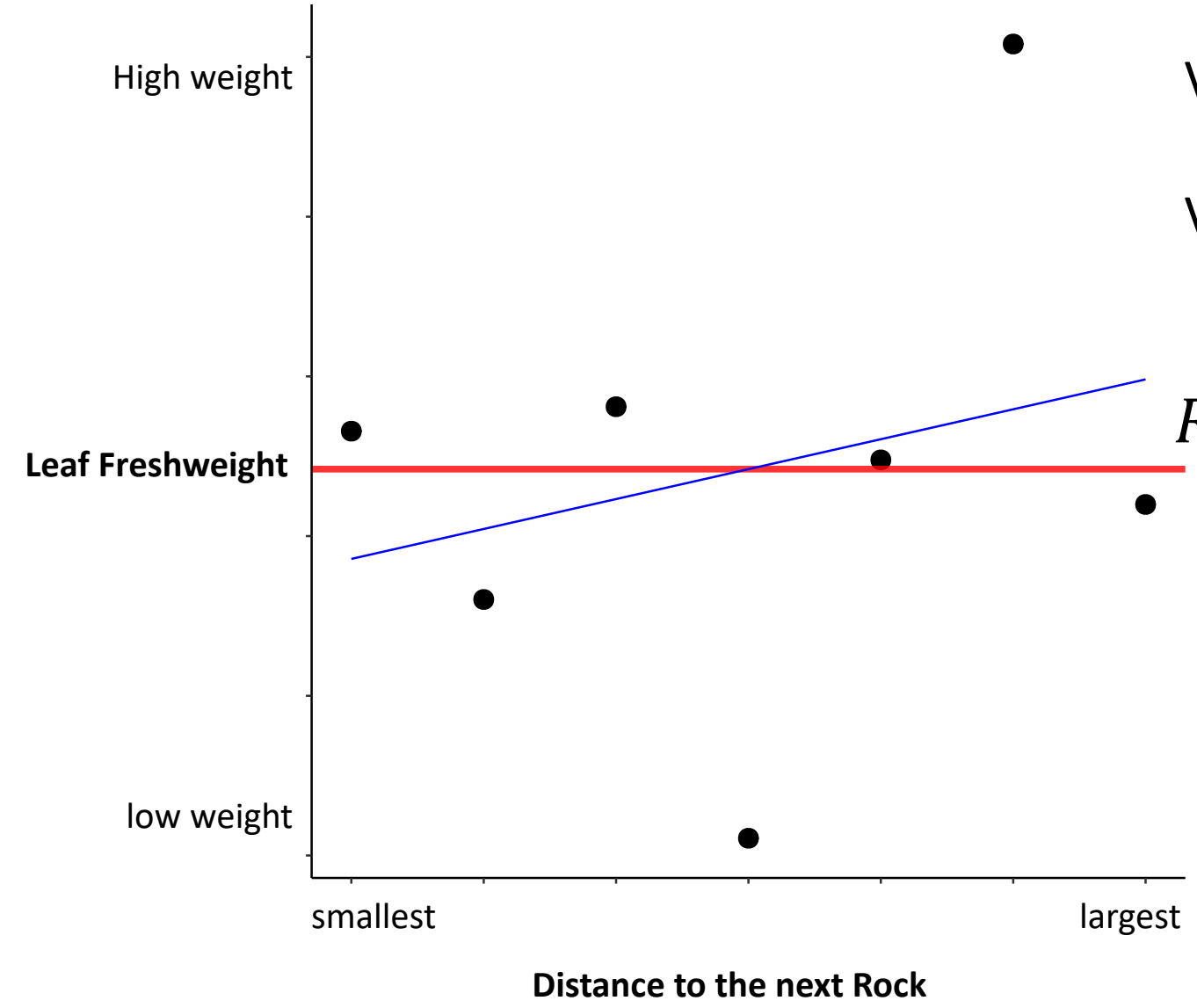
Var(**line**) = 50



Var(**mean**) = 59

Var(**line**) = 50

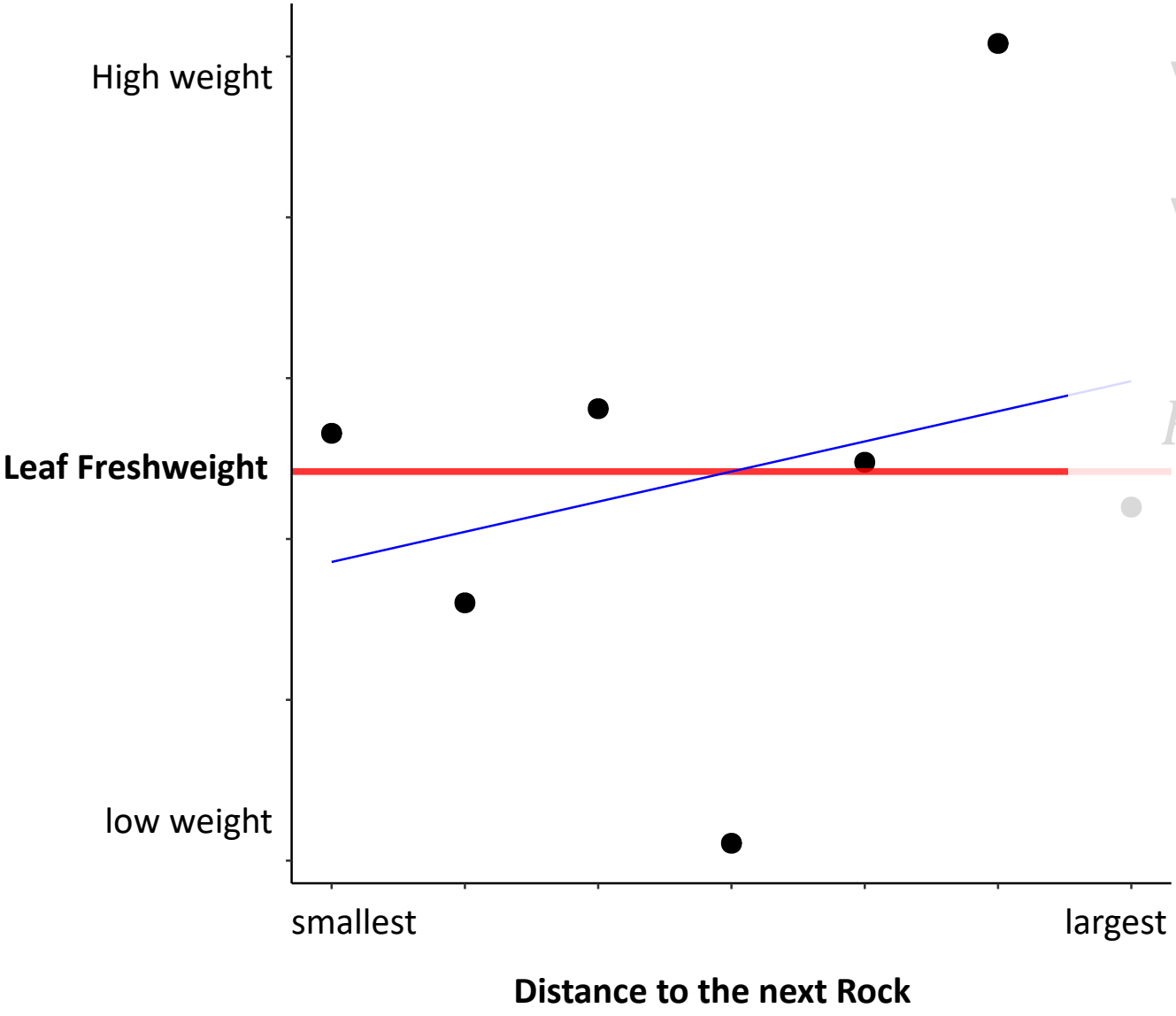
$$R^2 = \frac{Var(\textbf{mean}) - Var(\textbf{line})}{Var(\textbf{mean})}$$



Var(**mean**) = 59

Var(**line**) = 50

$$R^2 = \frac{\mathbf{59} - \mathbf{50}}{\mathbf{59}}$$



Var(**mean**) = 59

Var(**line**) = 50

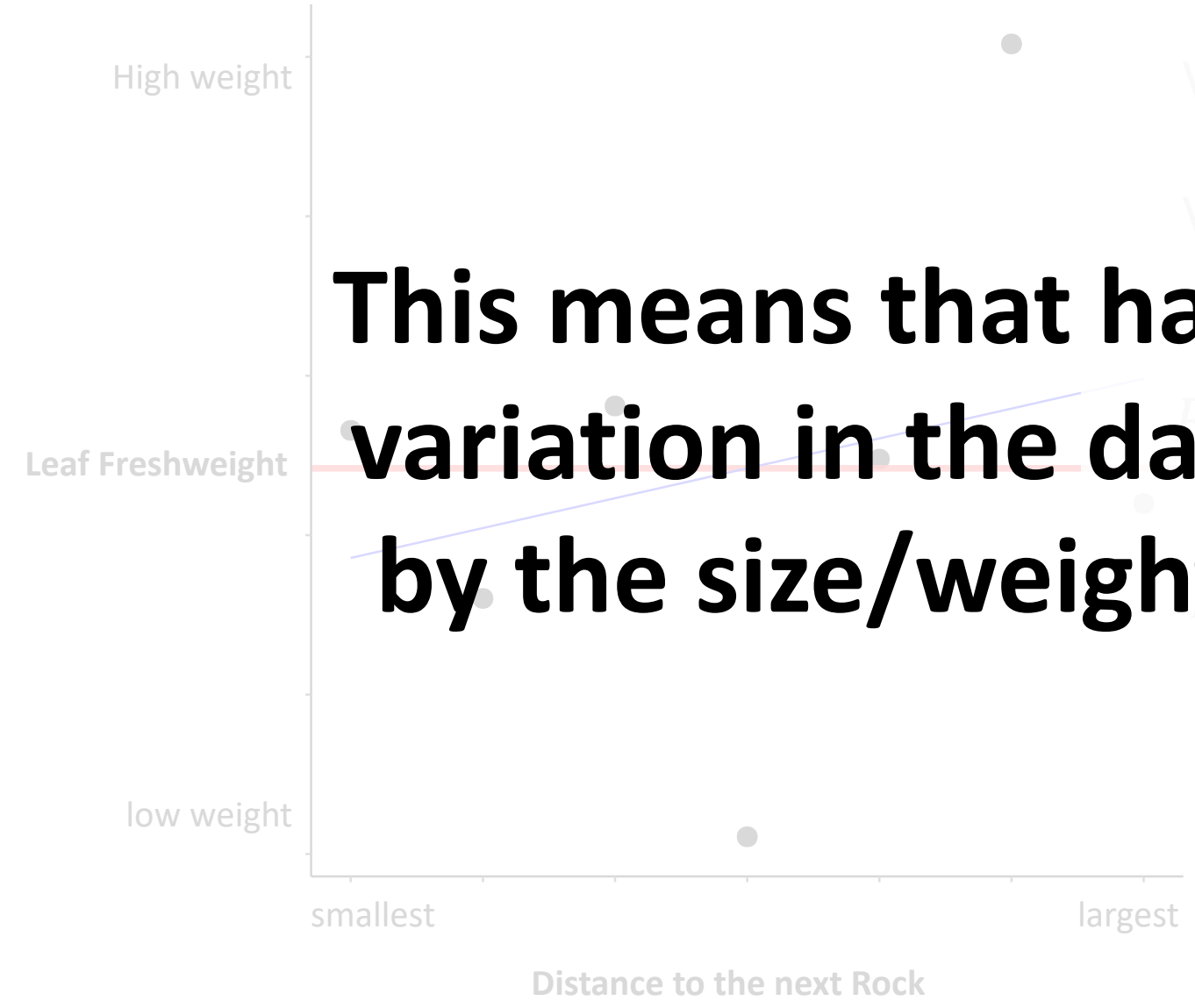
$$R^2 = \frac{59 - 50}{59}$$

$$R^2 = \frac{9}{59} = 0.15 = 15\%$$

There is 15% less variation around the **line** than the **mean**.

…or…

The size/weight relationship accounts for 15% of the variation



This means that hardly any of the variation in the data is explained by the size/weight relationship.

Var(**mean**) = 59

Var(**line**) = 50

$$R^2 = \frac{59 - 50}{59}$$

There is 15% less variation around the **line** than the **mean**.

…or…

The size/weight relationship accounts for 15% of the variation

- Now, when someone says..
  - “the statistically significant R^2 was 0.9”
- You can think to yourself..
  - “Very good! The relationship between the two variables explains 90% of the variation in the data!”

- And when someone else says…
  - “The statistically significant R^2 was 0.01”
- You can think to yourself..
  - “Who cares if that relathionship is significant, it only accounts for 1% of the variation in the data.”
  - ”Something else must explain the remaining 99%”

- Now, when someone says..
  - “the statistically significant R^2 was 0.9”
- You can think to yourself..
  - “Very good! The relationship between the two variables explains 90% of the variation in the data!”

- And when someone else says…
  - “The statistically significant R^2 was 0.01”
- You can think to yourself..
  - “Who cares if that relathionship is significant, it only accounts for 1% of the variation in the data.”
  - ”Something else must explain the remaining 99%”

So, what about R? How is it related to R^2?

- R^2 is just the square of R.

- Now, when someoene says…
  - “The statistically significant R was 0.9”

- You can think to yourself..
  - ”0.9 times 0.9 = 0.81. Very good! The relationship between the two variables explains 81% of the variation in the data!”

I prefer R^2 over plain R, because it is easier to interpret:

How much better is R=0.7 than R=0.5?

I prefer R^2 over plain R, because it is easier to interpret:

How much better is R=0.7 than R=0.5?

If we convert R to R^2, we see that:

$R^2=0.7^2=0.5$	50% of the original variation is explained
$R^2=0.5^2=0.25$	25% of the original variation is explained

That said, R^2 does not indicate the direction of the correlation, because it can never be negative.

If the direction of the correlation isn't obvious, you can say:

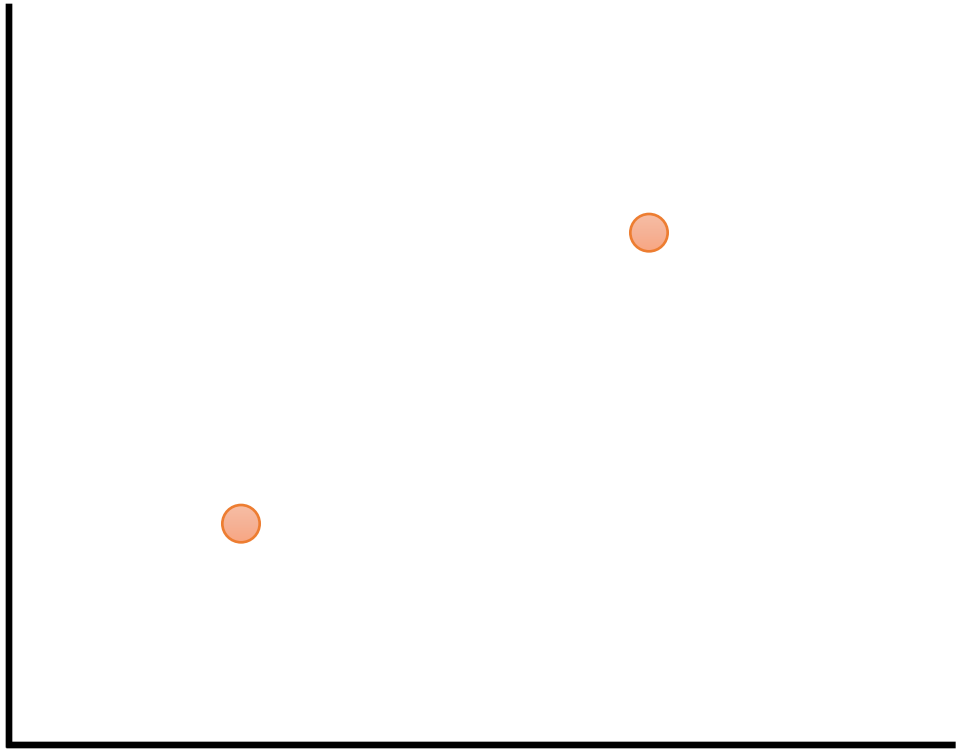
"the two variables were positively (or negatively) correlated with R^2=.."

Main Ideas for R^2

- R^2 is the percentage of variation explained by the relationship between two variables.

- If someone gives you a value for plain old R, square it!

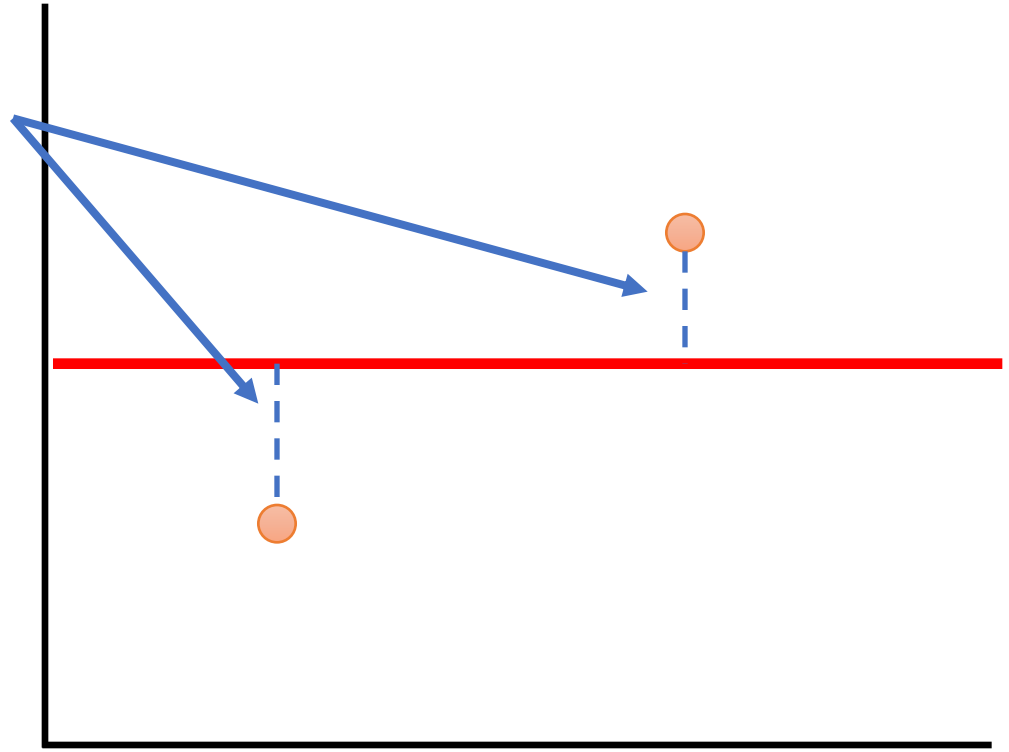
So, this is nice and all, but:



What if we had just 2 measurements?

So, this is nice and all, but:

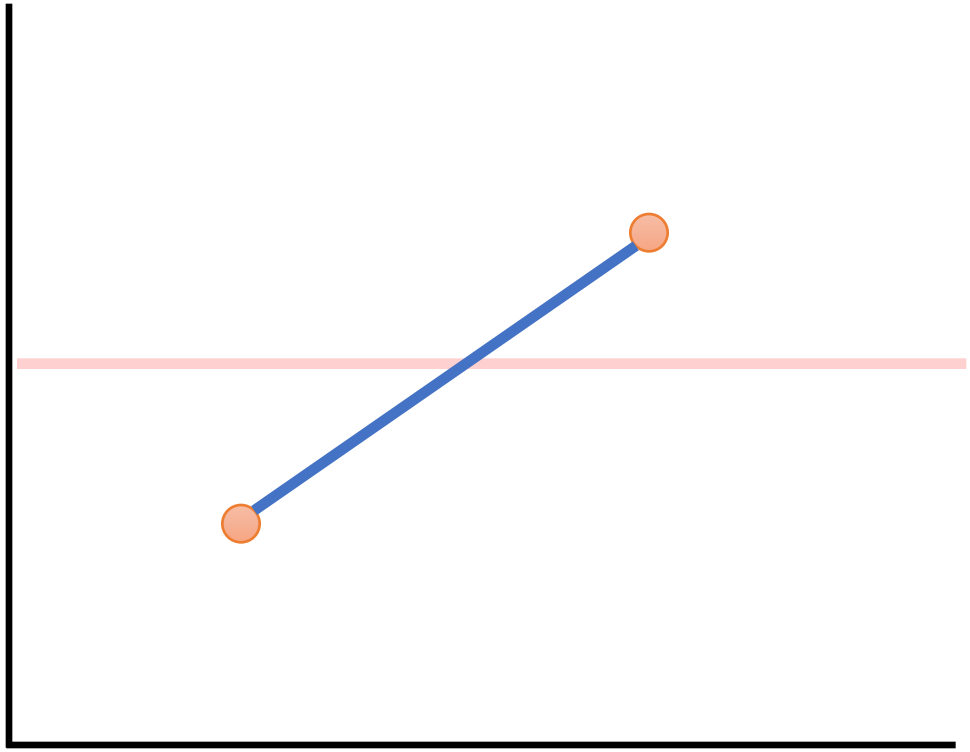
SS(mean) = 10



So, this is nice and all, but:

SS(mean) = 10

SS(fit) = 0

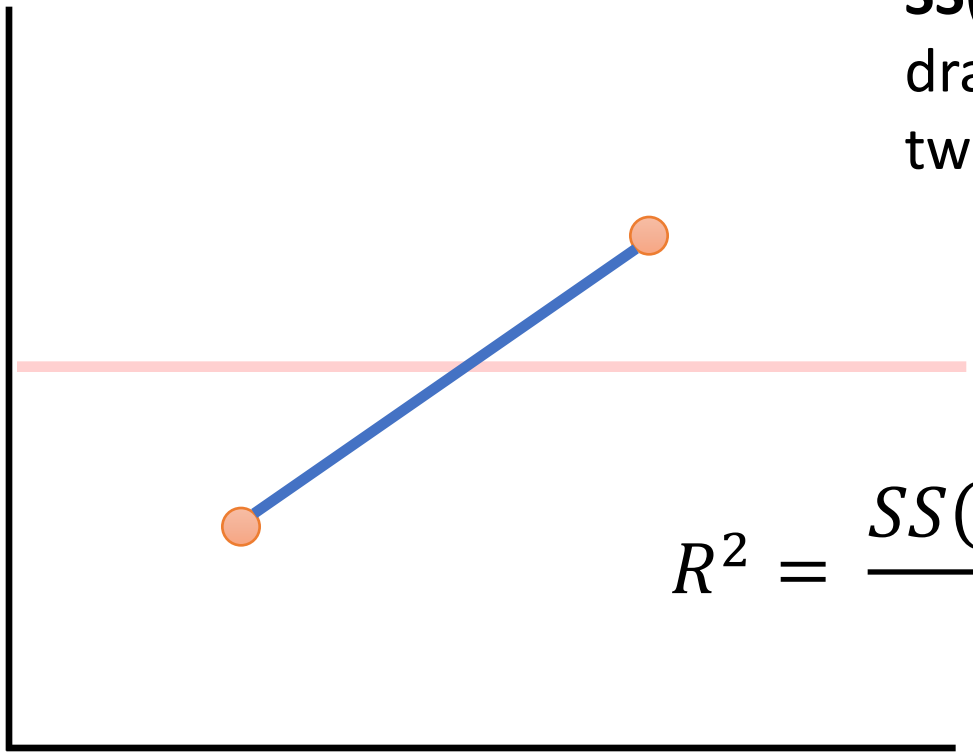


So, this is nice and all, but:

SS(mean) = 10

SS(fit) = 0

SS(fit) = 0, because you can always draw a straight line to connect **any** two dots

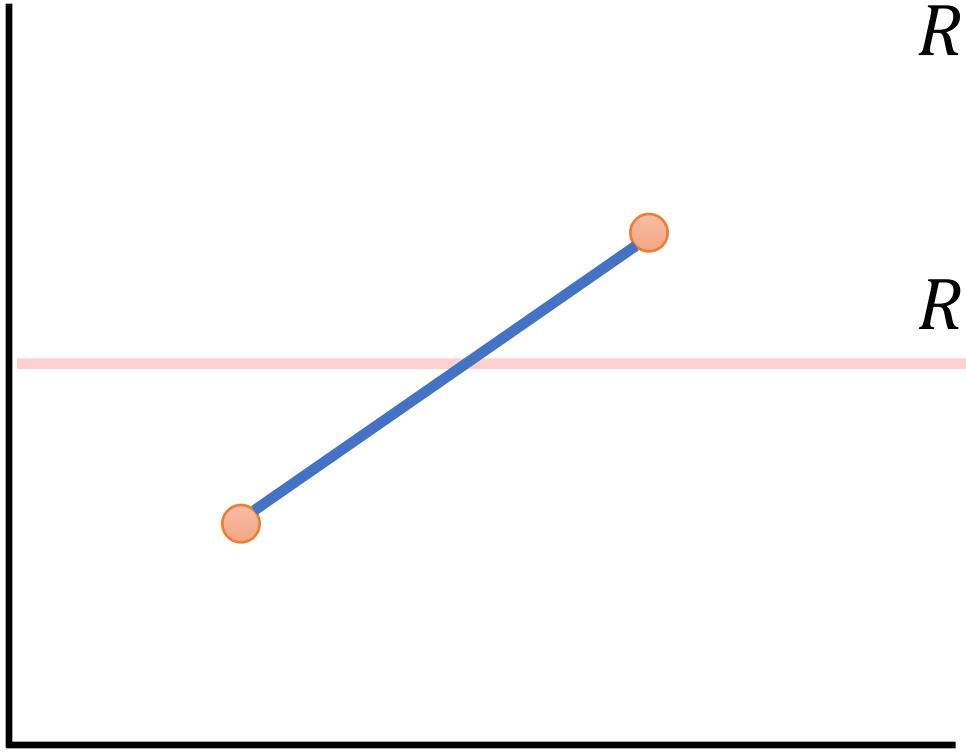


$$R^2 = \frac{SS(\mathbf{mean}) - SS(\mathbf{line})}{SS(\mathbf{mean})}$$

So, this is nice and all, but:

SS(mean) = 10

SS(fit) = 0



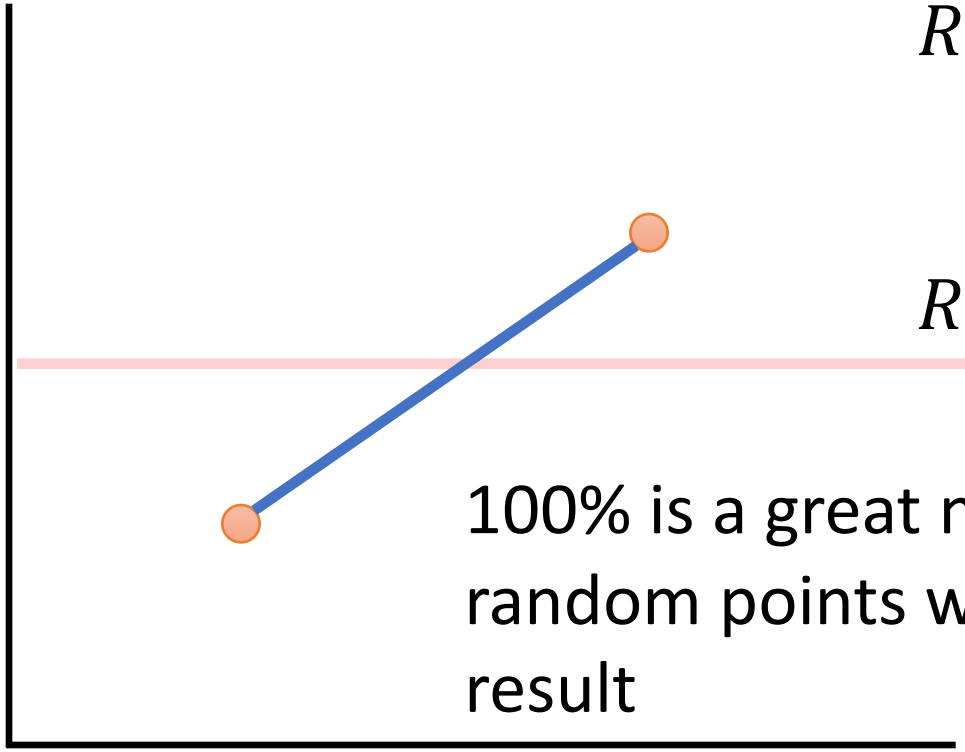
$$R^2 = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})}{SS(\mathbf{mean})}$$

$$R^2 = \frac{100 - 0}{100} = 1 = 100\%$$

So, this is nice and all, but:

SS(mean) = 10

SS(fit) = 0

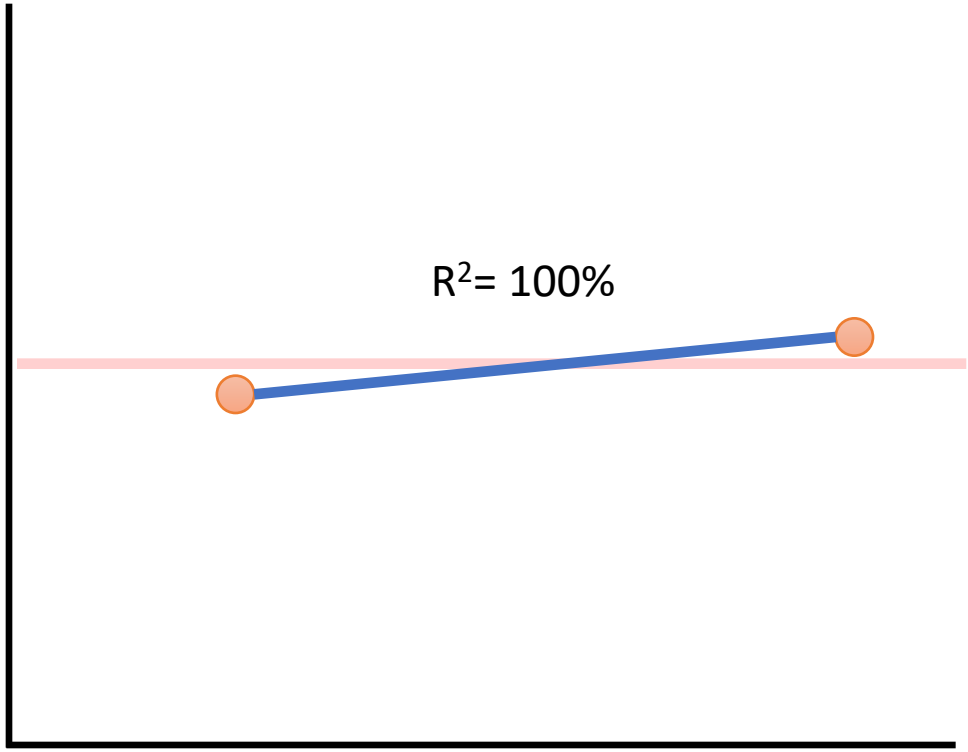


$$R^2 = \frac{SS(\textbf{mean}) - SS(\textbf{line})}{SS(\textbf{mean})}$$

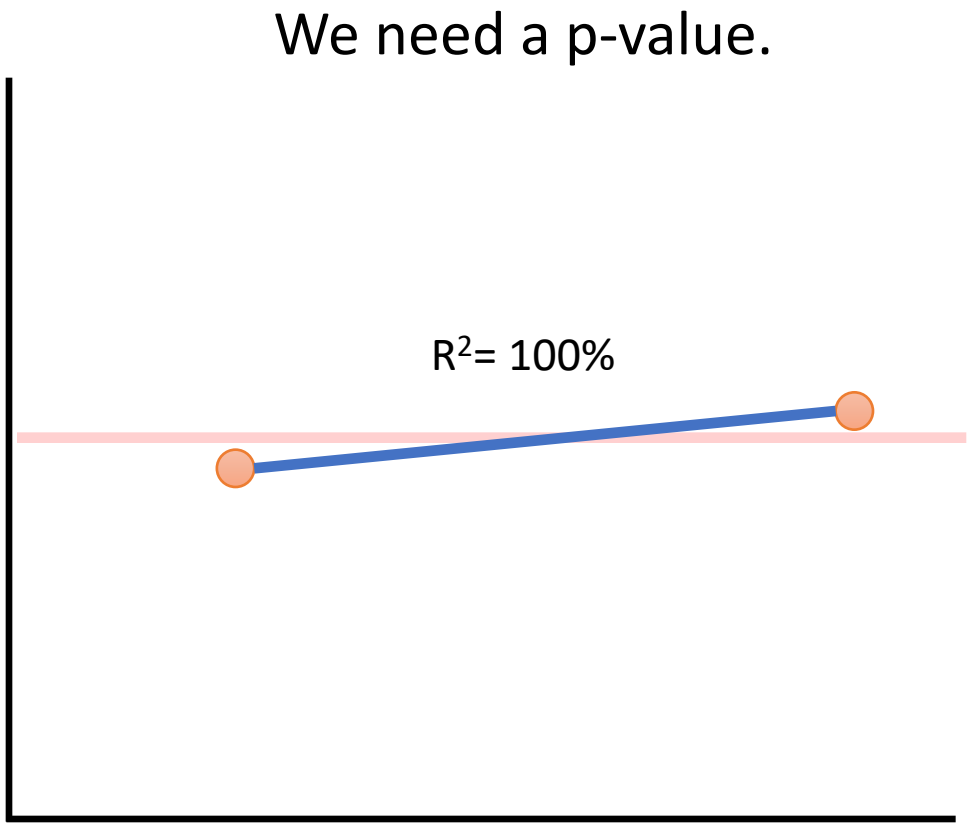
$$R^2 = \frac{100 - 0}{100} = 1 = 100\%$$

100% is a great number, but any two random points will yield the same result

We need a way to determine if the R^2 is statistically significant.



We need a way to determine if the R^2 is statistically significant.



$$R^2 = \frac{SS(\textbf{mean}) - SS(\textbf{line})}{SS(\textbf{mean})}$$

$$\frac{SS(\textbf{leaf weight}) - SS(\textbf{after taking size into account})}{SS(\textbf{leaf weight})}$$

$$\frac{Variation\ in\ leaf\ weight\ explained\ by\ size}{Variation\ in\ leaf\ weight\ without\ taking\ size\ into\ account}$$

$$R^2 = \frac{Variation\ in\ leaf\ weight\ explained\ by\ size}{Variation\ in\ leaf\ weight\ without\ taking\ size\ into\ account}$$

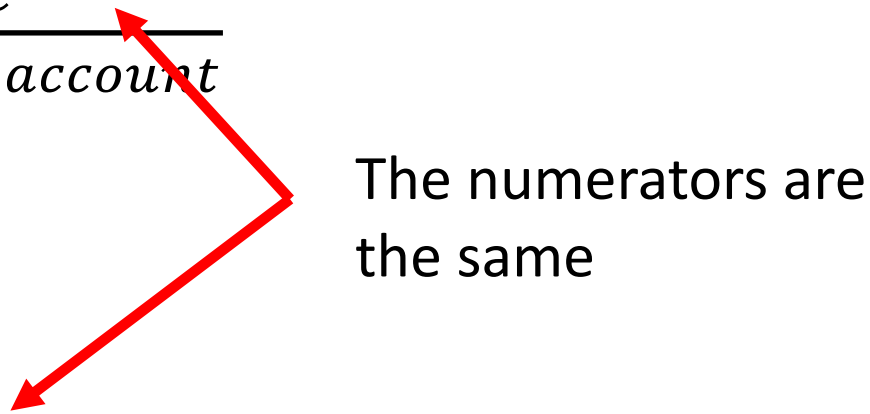
The p-value for R^2 comes from something called “F”

$$R^2 = \frac{Variation\ in\ leaf\ weight\ explained\ by\ size}{Variation\ in\ leaf\ weight\ without\ taking\ size\ into\ account}$$

$$F = \frac{Variation\ in\ leaf\ weight\ explained\ by\ size}{Variation\ in\ leaf\ weight\ not\ explained\ by\ weight}$$

The p-value for R^2 comes from something called “F”

$$R^2 = \frac{Variation\ in\ leaf\ weight\ explained\ by\ size}{Variation\ in\ leaf\ weight\ without\ taking\ size\ into\ account}$$



The numerators are the same

$$F = \frac{Variation\ in\ leaf\ weight\ explained\ by\ size}{Variation\ in\ leaf\ weight\ not\ explained\ by\ weight}$$

The p-value for R^2 comes from something called “F”

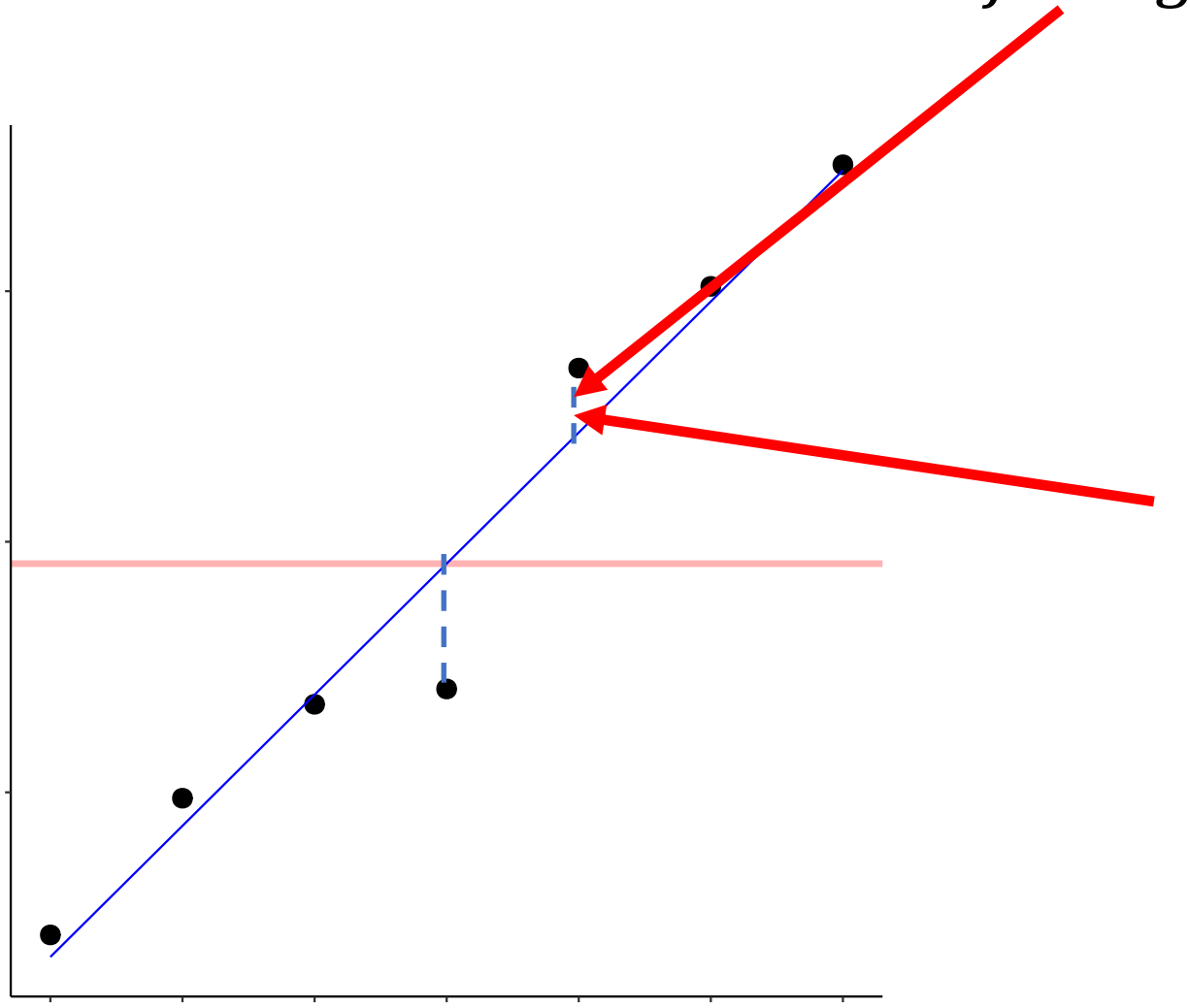
$$R^2 = \frac{Variation\ in\ leaf\ weight\ explained\ by\ size}{Variation\ in\ leaf\ weight\ without\ taking\ size\ into\ account}$$

The denominator is slightly different

$$F = \frac{Variation\ in\ leaf\ weight\ explained\ by\ size}{Variation\ in\ leaf\ weight\ not\ explained\ by\ weight}$$

The p-value for R^2 comes from something called “F”

$$F = \frac{Variation\ in\ leaf\ weight\ explained\ by\ size}{Variation\ in\ leaf\ weight\ not\ explained\ by\ weight}$$



These dotted lines (residuals) represent the variation that remains after fitting the line. **This is the variation that is not explained by size.**

Now let’s look at the underlying mathematics

Now let's look at the underlying mathematics

$$R^2 = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})}{SS(\mathbf{mean})}$$

Now let's look at the underlying mathematics

$$R^2 = \frac{SS(\textbf{mean}) - SS(\textbf{fit})}{SS(\textbf{mean})} \qquad F = \frac{SS(\textbf{mean}) - SS(\textbf{fit})/(p_{fit} - p_{mean})}{SS(\textbf{fit})/(n - p_{fit})}$$

$$R^2 = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})}{SS(\mathbf{mean})} \qquad F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

The ”meat” of these equations are very similar and rely on the same “sum of squares”

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

These numbers over here are the “degrees of freedom”

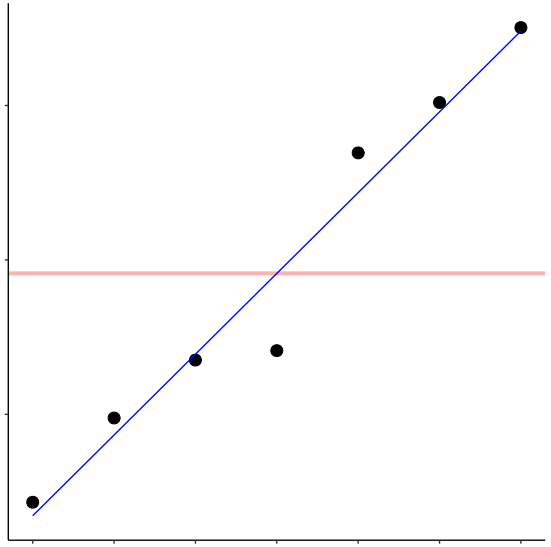
$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

These numbers over here are the “degrees of freedom”

They turn the sums of squares into variances

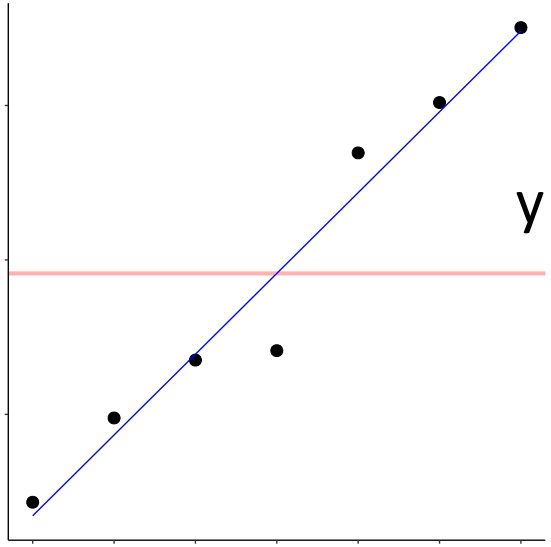
$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

Maybe we can dedicate a whole Seminar to degrees of freedom. For now, let's see if we can get an intuitive feel for what they are doing here.



$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

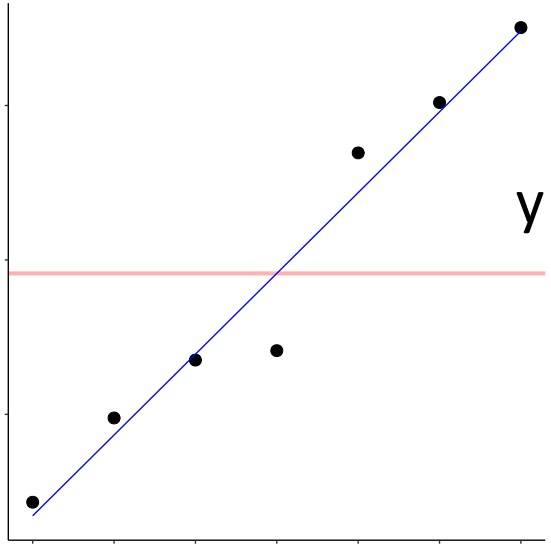
P_{fit} is the number of paramters in the fit line



y = y-intercept + slope x

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

P_{fit} is the number of paramters in the fit line



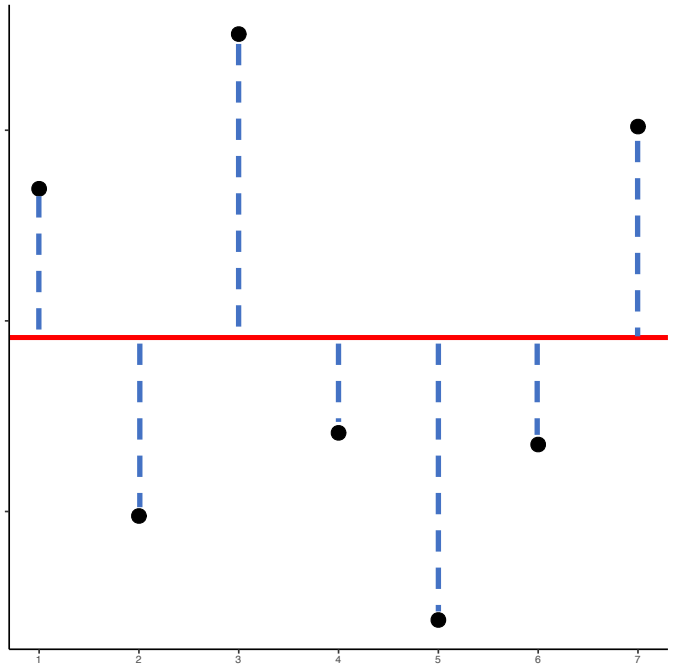
y = y-intercept + slope x

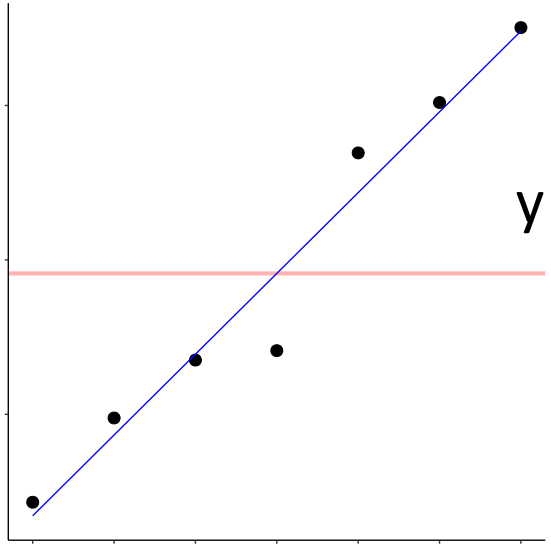
2 parameters

$p_{fit} = 2$

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

p_{fit} is the number of paramters in the fit line





y = y-intercept + slope x

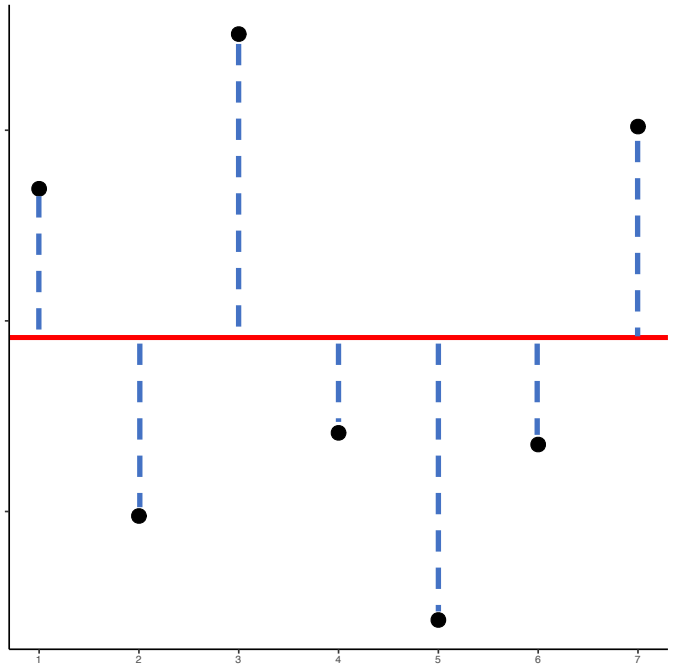
2 parameters

$p_{fit} = 2$

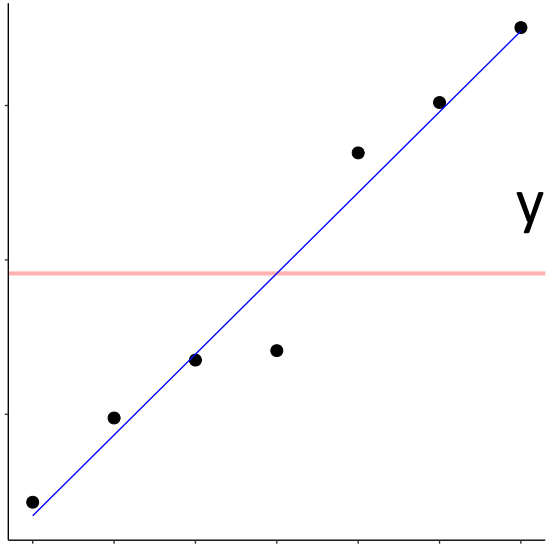
$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

p_{fit} is the number of paramters in the **fit** line

p_{mean} is the number of paramters in the **mean** line



y = y-intercept



y = y-intercept + slope x

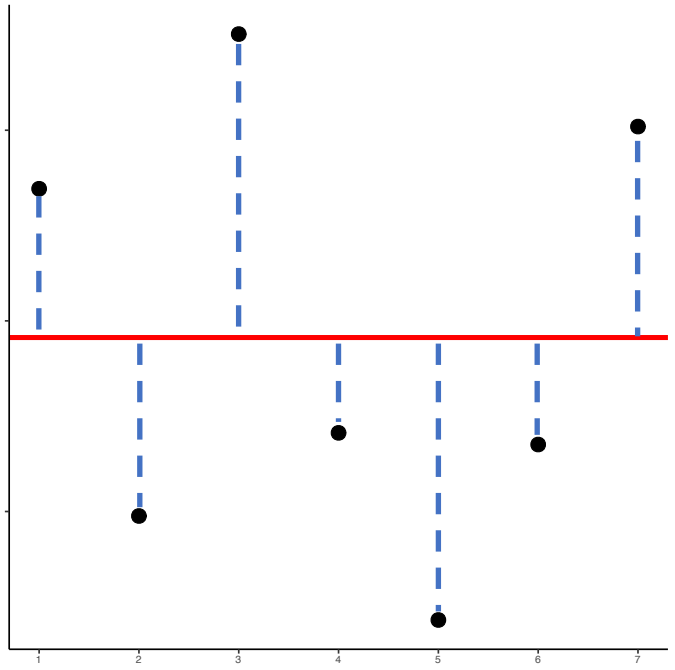
2 parameters

$p_{fit} = 2$

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

p_{fit} is the number of paramters in the **fit** line

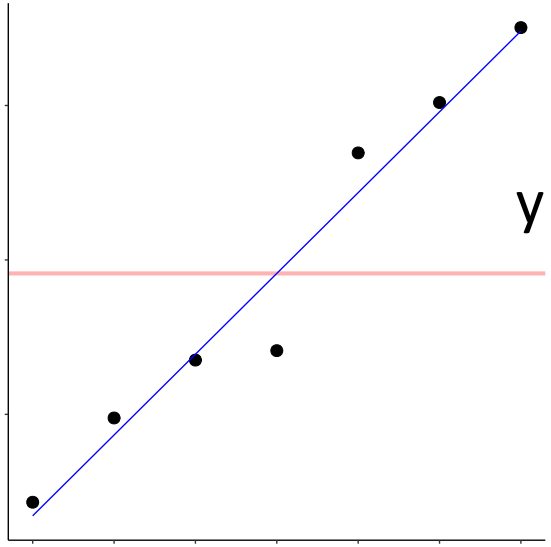
p_{mean} is the number of paramters in the **mean** line



y = y-intercept

1 parameters

$p_{mean} = 1$



y = y-intercept + slope x

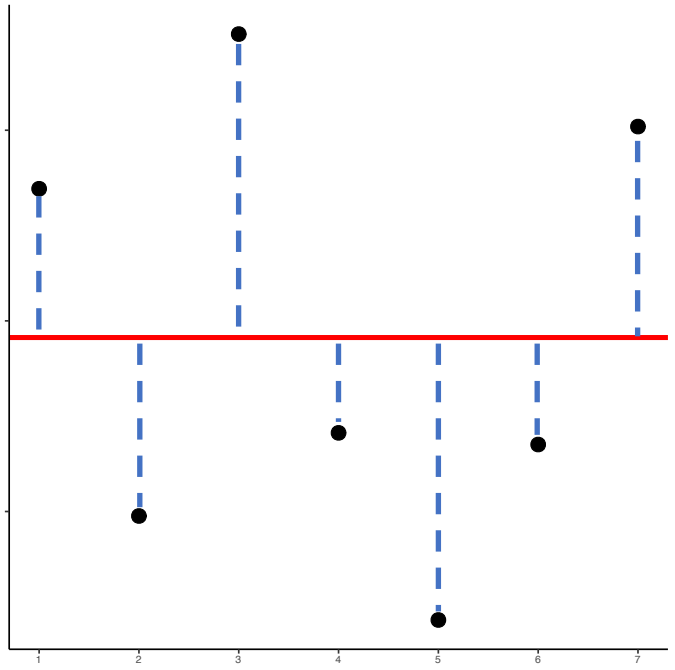
2 parameters

$p_{fit} = 2$

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

Both equations have a parameter for the y-intercept.

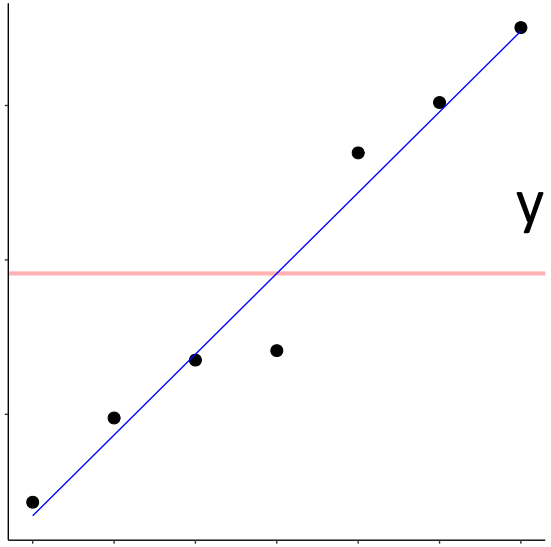
However, the **fit** line has one extra parameter, the slope.
In our example, this slope is the relationship between size and weight.



y = y-intercept

1 parameters

$p_{mean} = 1$



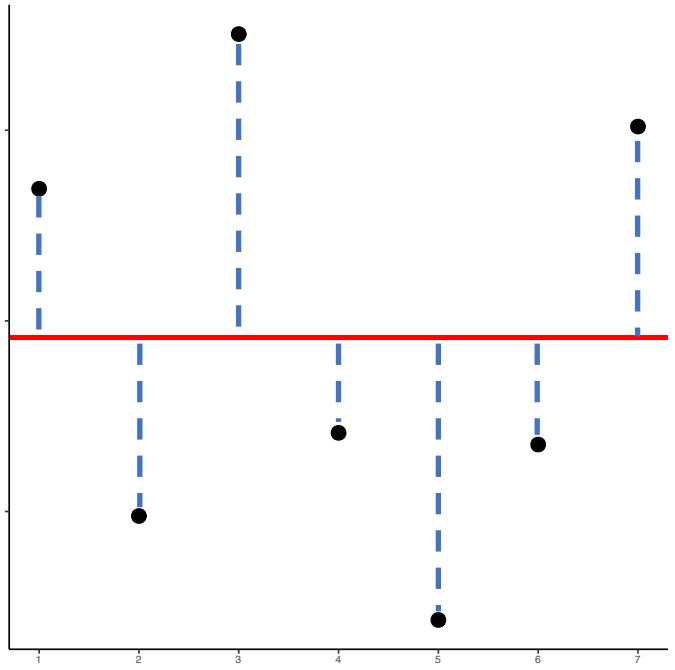
y = y-intercept + slope x

2 parameters

$p_{fit} = 2$

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

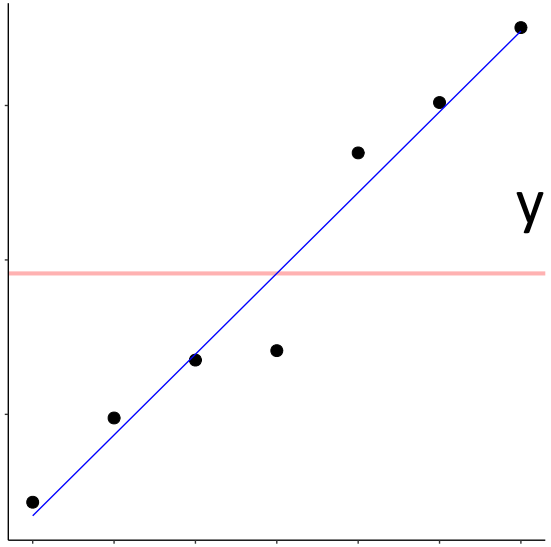
$(p_{fit} - p_{mean}) = (2\text{-}1) = 1$



y = y-intercept

1 parameters

$p_{mean} = 1$



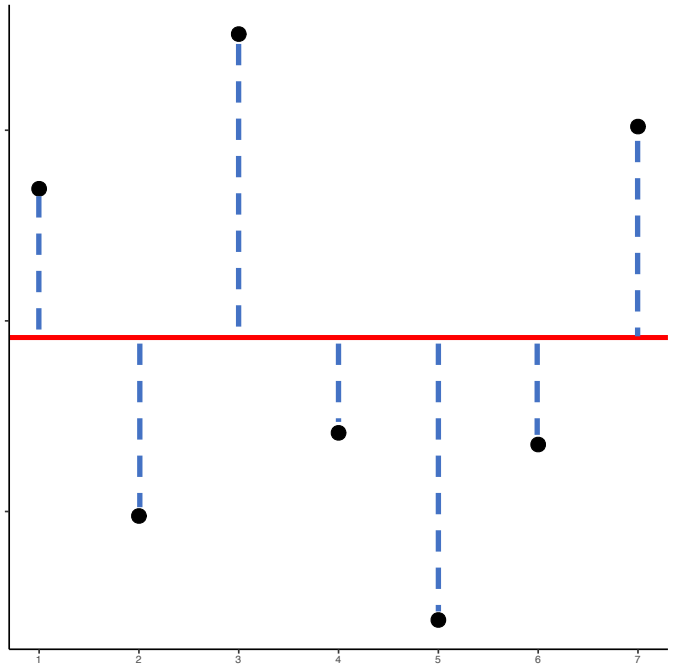
y = y-intercept + slope x

2 parameters

$p_{fit} = 2$

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

Thus, the numerator is the variance explained by the extra parameter. In our example, that's the variance in leaf weight explained by leaf size.



y = y-intercept + slope x + slope z

3 parameters

y = y-intercept

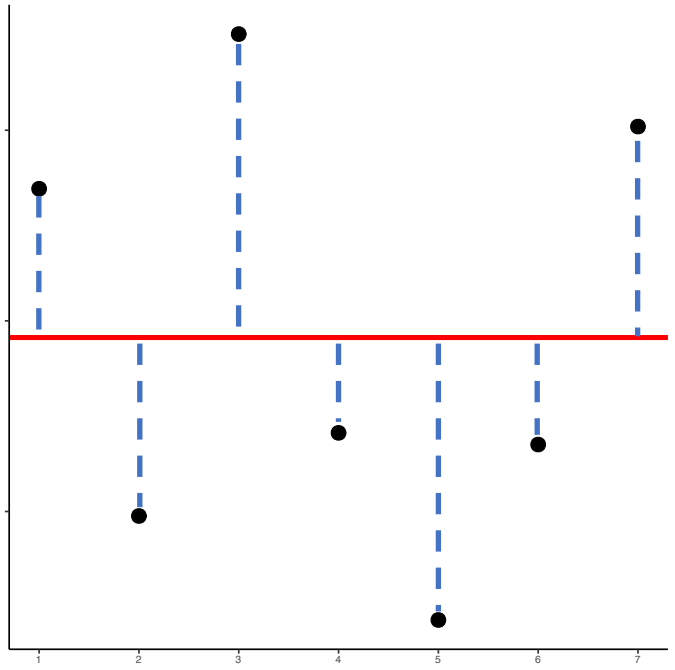
1 parameters

$p_{mean} = 1$

$p_{fit} = 3$

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

If we had a third variable, the amount of water in the leaf, we'd end up with this equation.



y = y-intercept

1 parameters

$p_{mean} = 1$

y = y-intercept + slope x + slope z

3 parameters

$p_{fit} = 3$

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

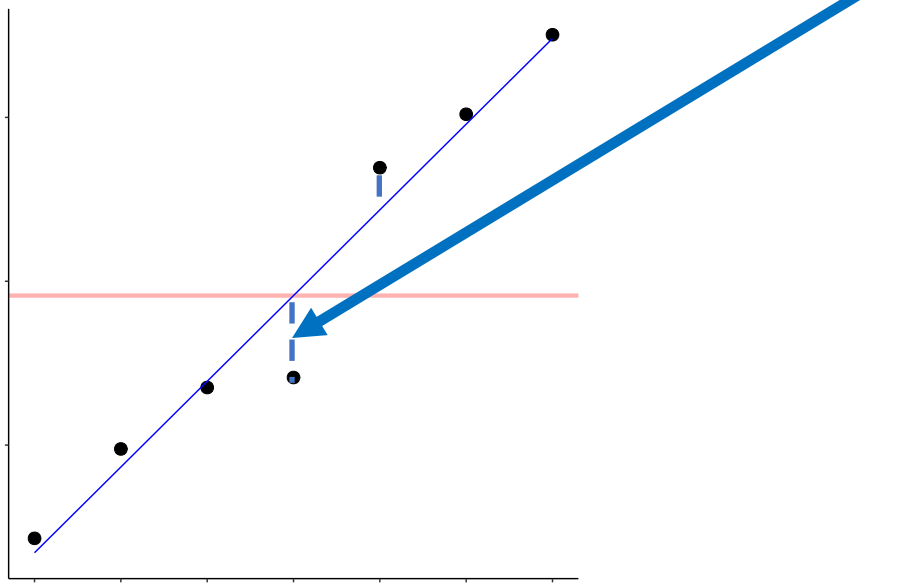
With the fancier **fit**, the numerator is the variance in leaf weight explained by leaf size and water content.

Let’s talk about the denuminator

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

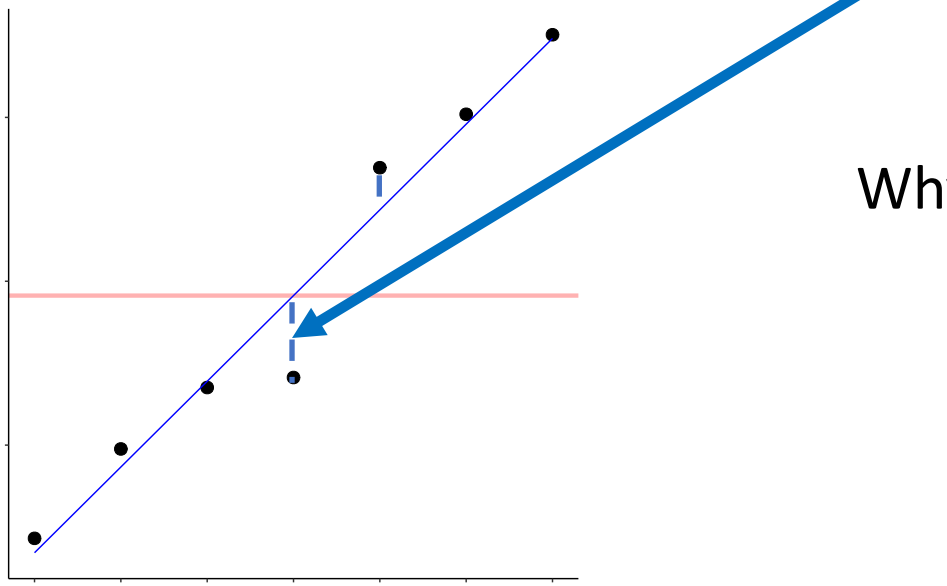
The variation in leaf weight not explained by the fit.

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$



The variation in leaf weight not explained by the fit.

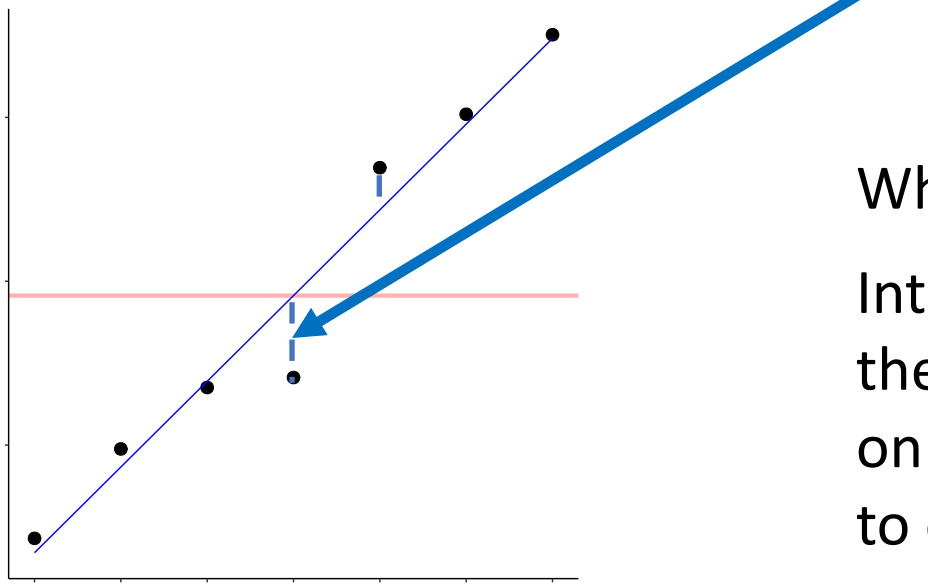
$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$



Why divide SS(**fit**) by n– p_{fit} instead of just n?

The variation in leaf weight not explained by the fit.

$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$



Why divide SS(**fit**) by n– p_{fit} instead of just n?

Intuitively, the more parameters you have in your equation, the more data you need to estimate them. For example, you only need two points to estimate a line, but you need 3 points to estimate a plane.

If the **fit** is good, then …

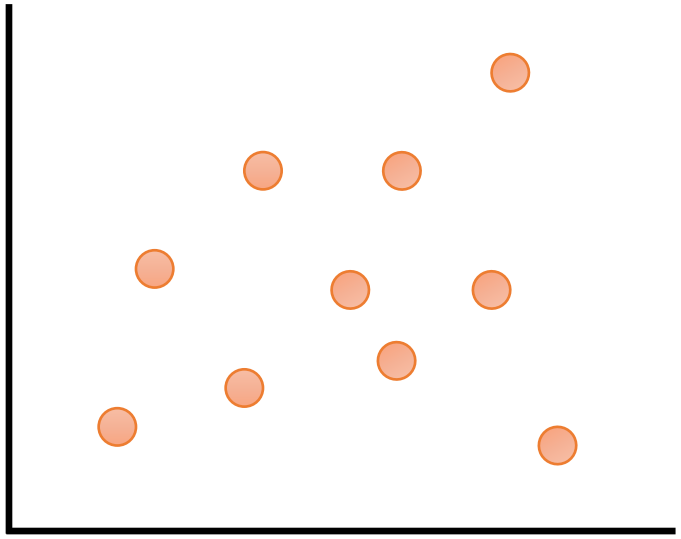
$$F = \frac{Variation\ in\ leaf\ weight\ explained\ by\ size}{Variation\ in\ leaf\ weight\ not\ explained\ by\ weight} \rightarrow \frac{large\ number}{small\ number}$$

If the **fit** is good, then …

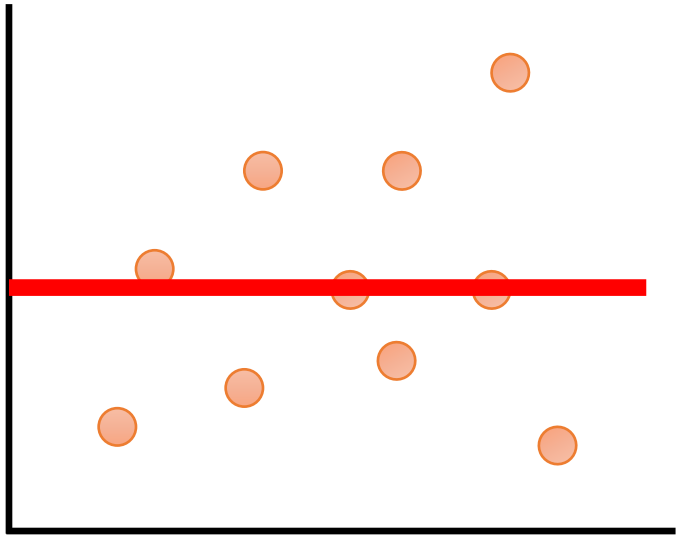
$$F = \frac{Variation\ in\ leaf\ weight\ explained\ by\ size}{Variation\ in\ leaf\ weight\ not\ explained\ by\ weight} \rightarrow \frac{large\ number}{small\ number}$$

F = really large number

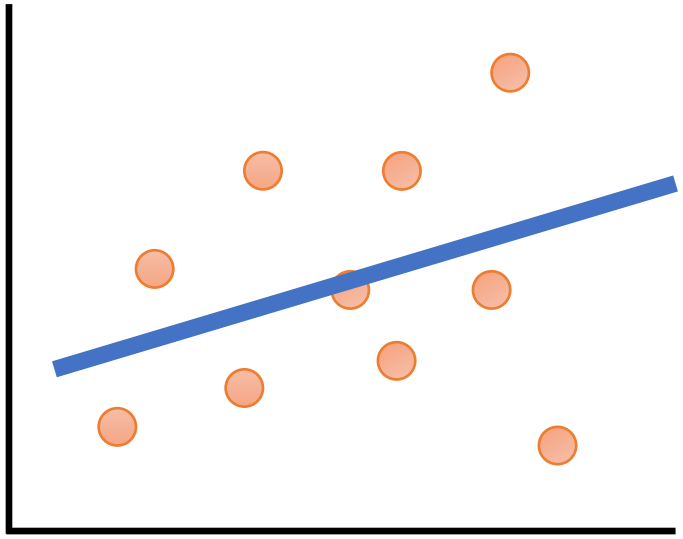
How do we turn this into a p-value?



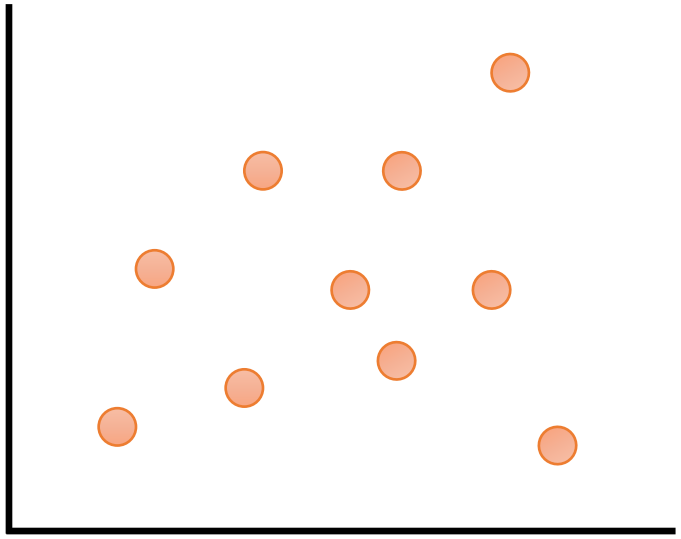
Generate a set of random data…



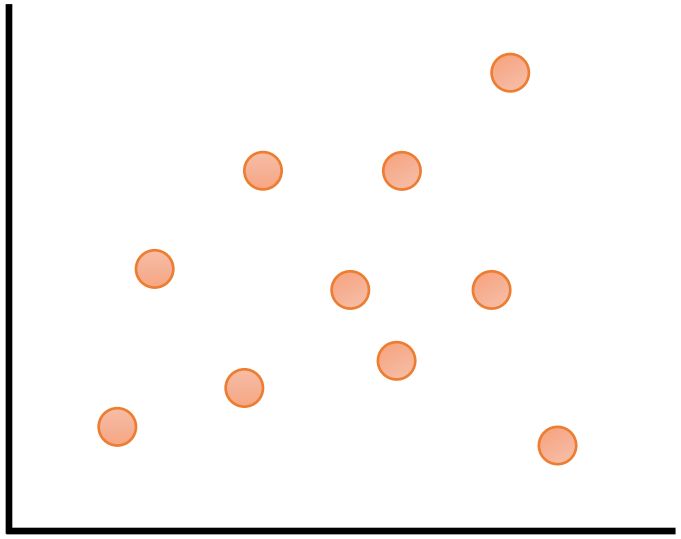
…calculate the **mean** and SS(**mean**)



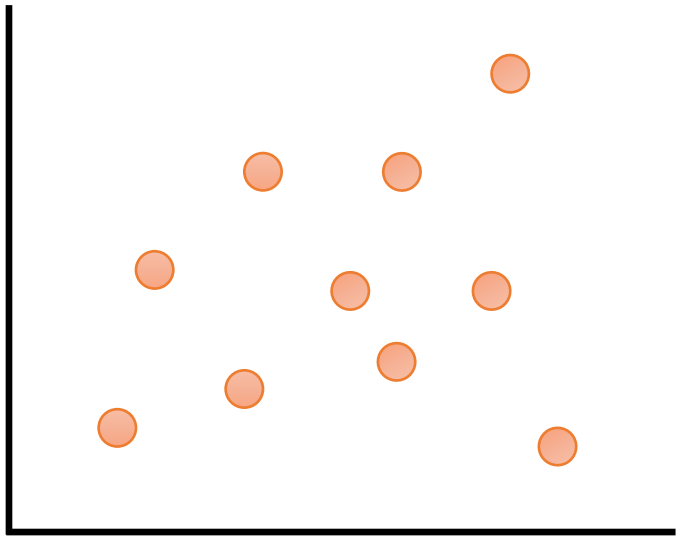
…calculate the **fit** and SS(**fit**)…



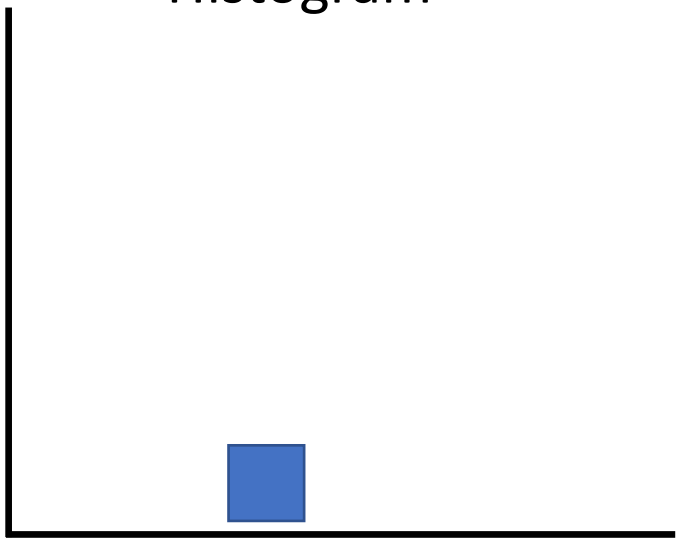
$$F = \frac{SS(\textbf{mean}) - SS(\textbf{fit})/(p_{fit} - p_{mean})}{SS(\textbf{fit})/(n - p_{fit})}$$



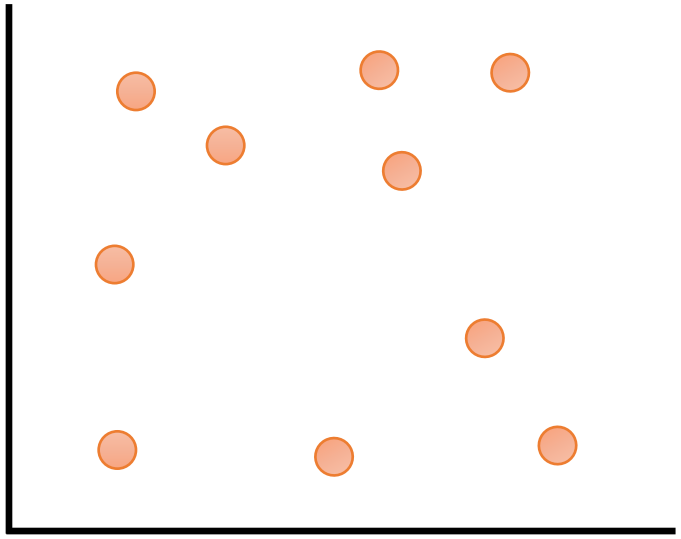
$$F = \frac{SS(\textbf{mean}) - SS(\textbf{fit})/(p_{fit} - p_{mean})}{SS(\textbf{fit})/(n - p_{fit})} \rightarrow F = 2$$



Histogram

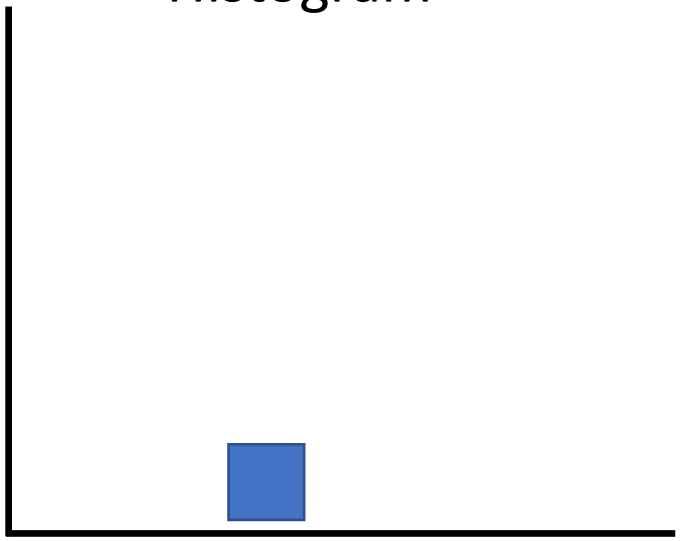


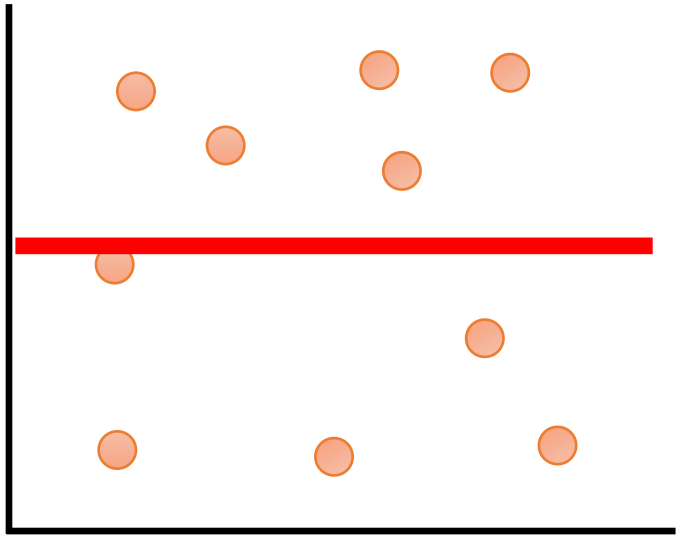
$$F = \frac{SS(\textbf{mean}) - SS(\textbf{fit})/(p_{fit} - p_{mean})}{SS(\textbf{fit})/(n - p_{fit})} \rightarrow F = 2$$



Generate another set of random data…

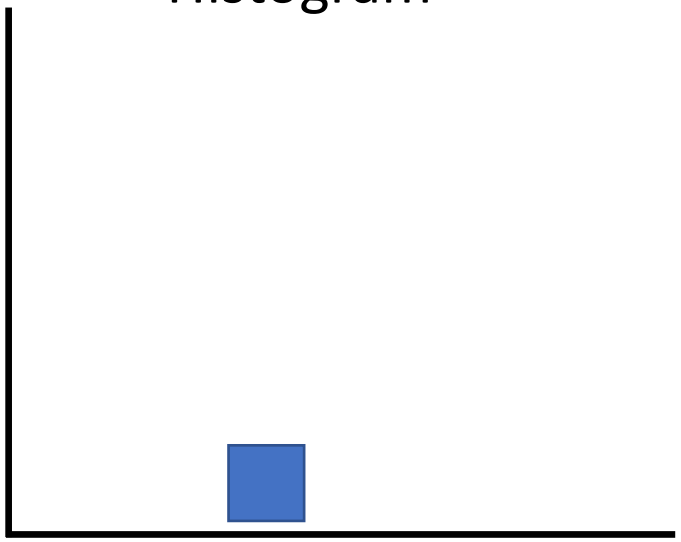
Histogram

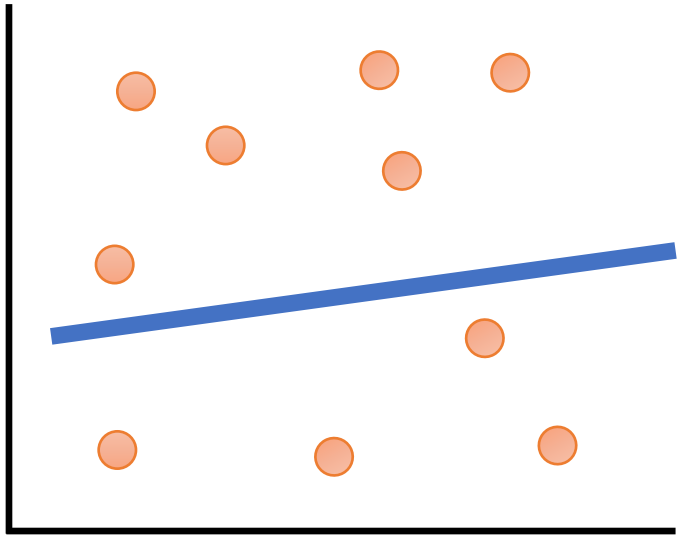




…calculate the **mean** and SS(**mean**)…

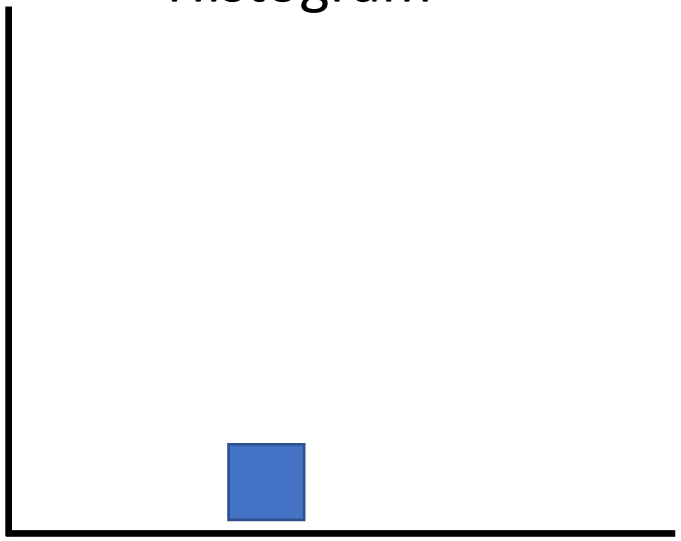
Histogram

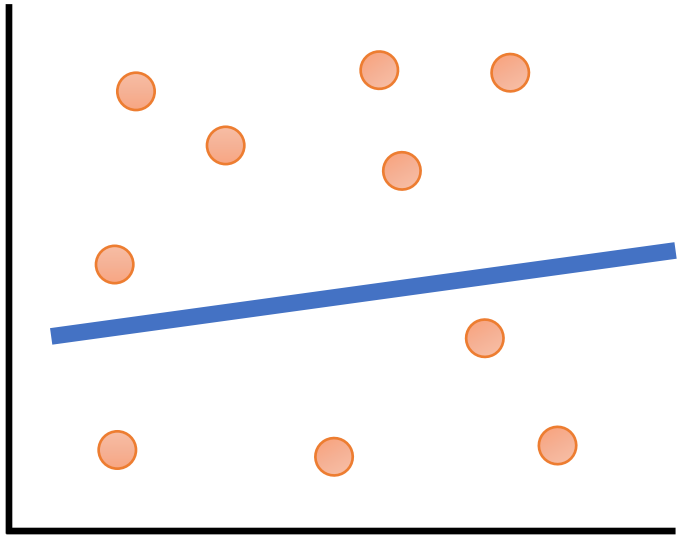




…calculate the **fit** and SS(**fit**)…

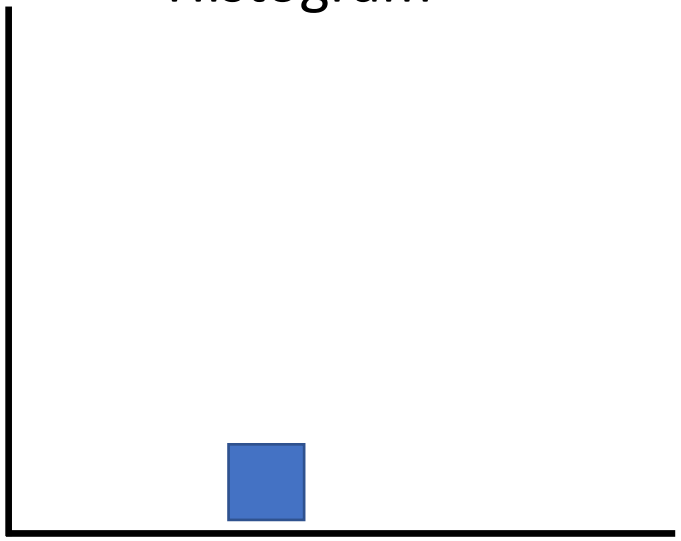
Histogram



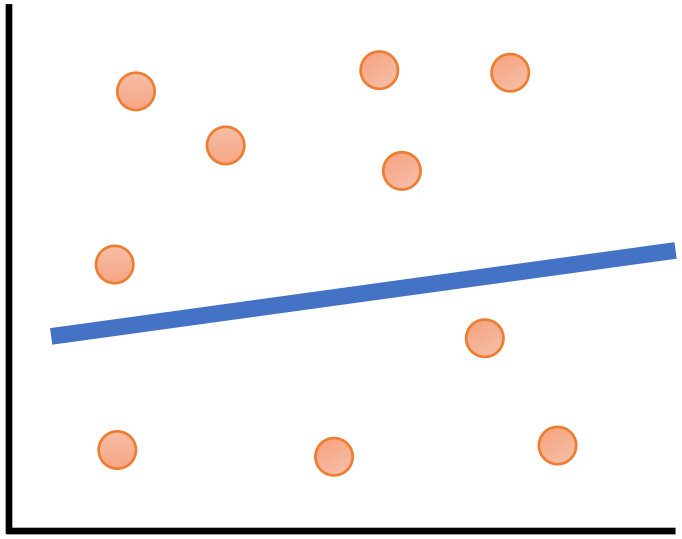


…calculate the **fit** and SS(**fit**)…

Histogram

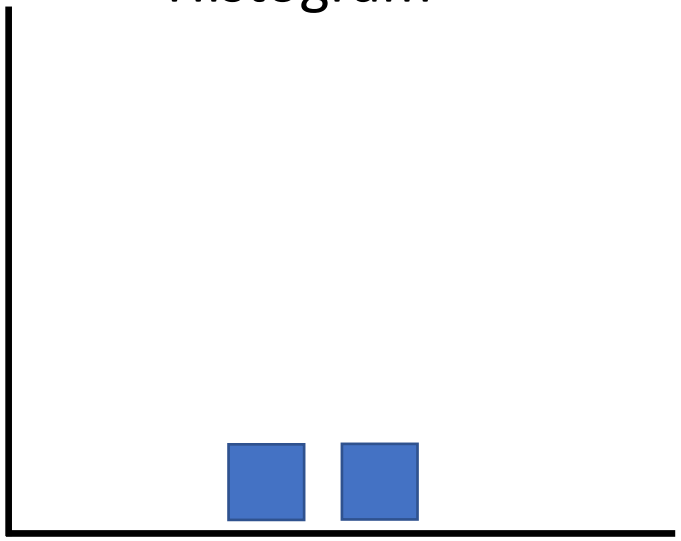


$$F = \frac{SS(\mathbf{mean}) - SS(\mathbf{fit})/(p_{fit} - p_{mean})}{SS(\mathbf{fit})/(n - p_{fit})}$$

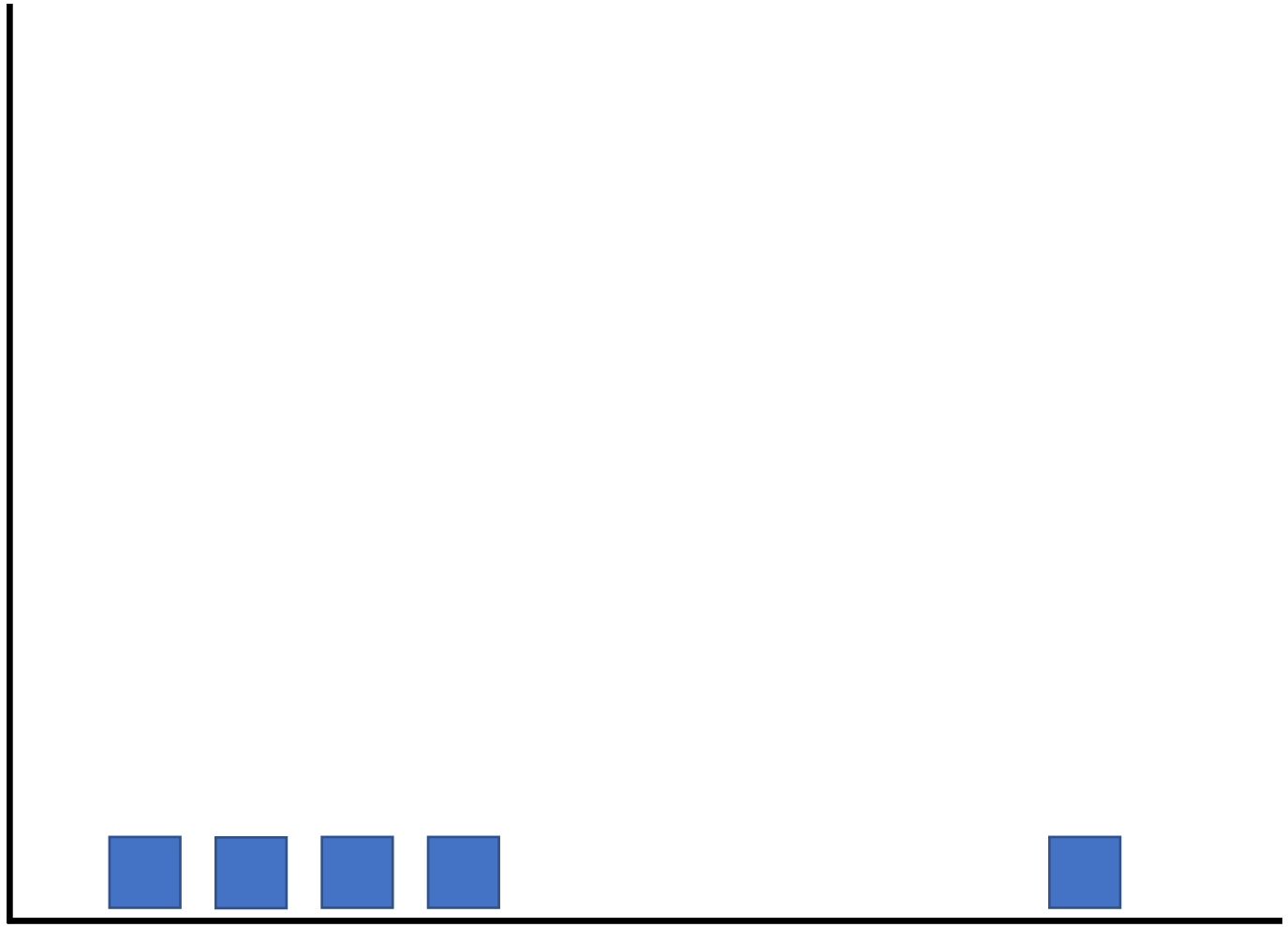


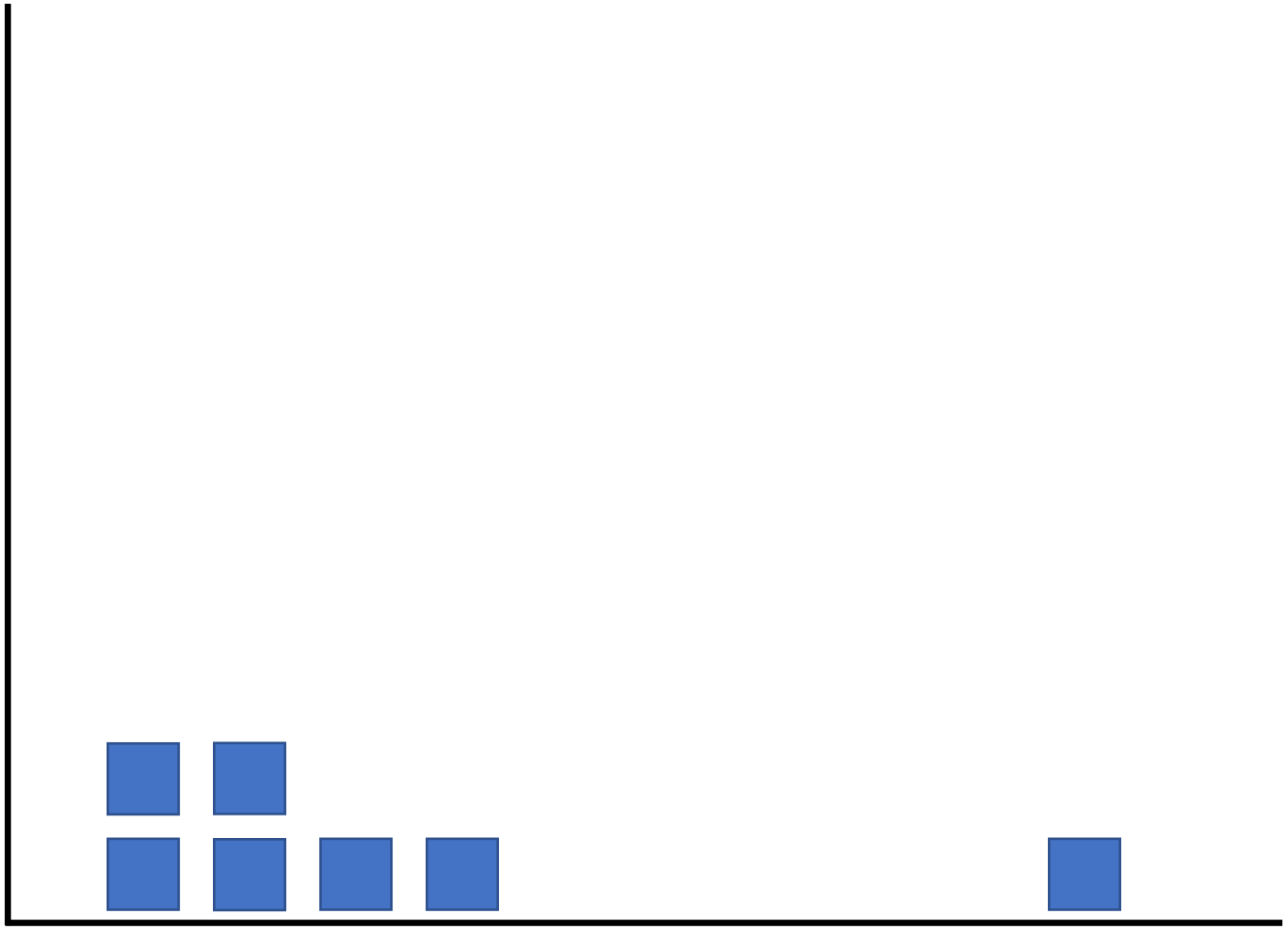
…calculate the **fit** and SS(**fit**)…

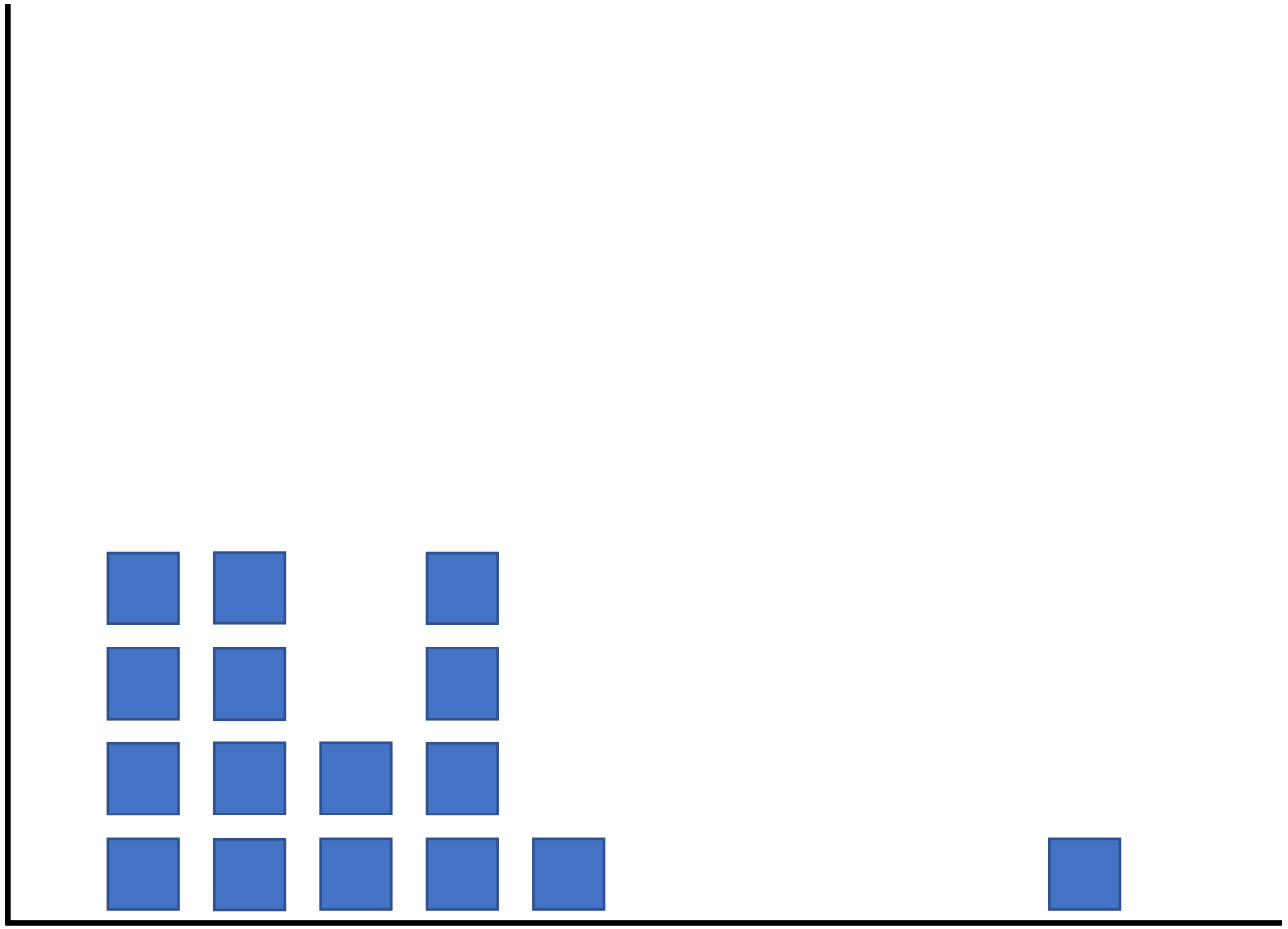
Histogram

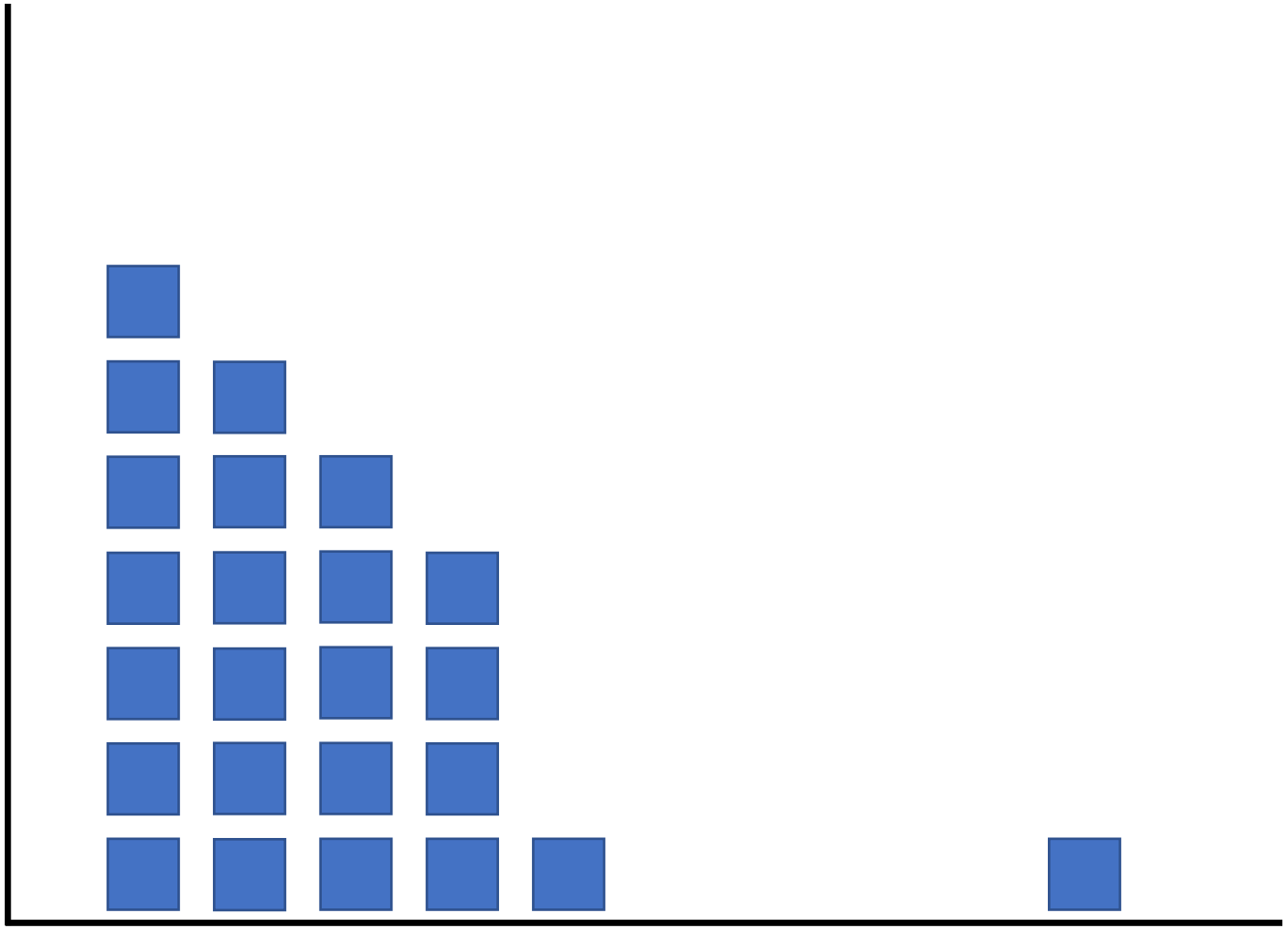


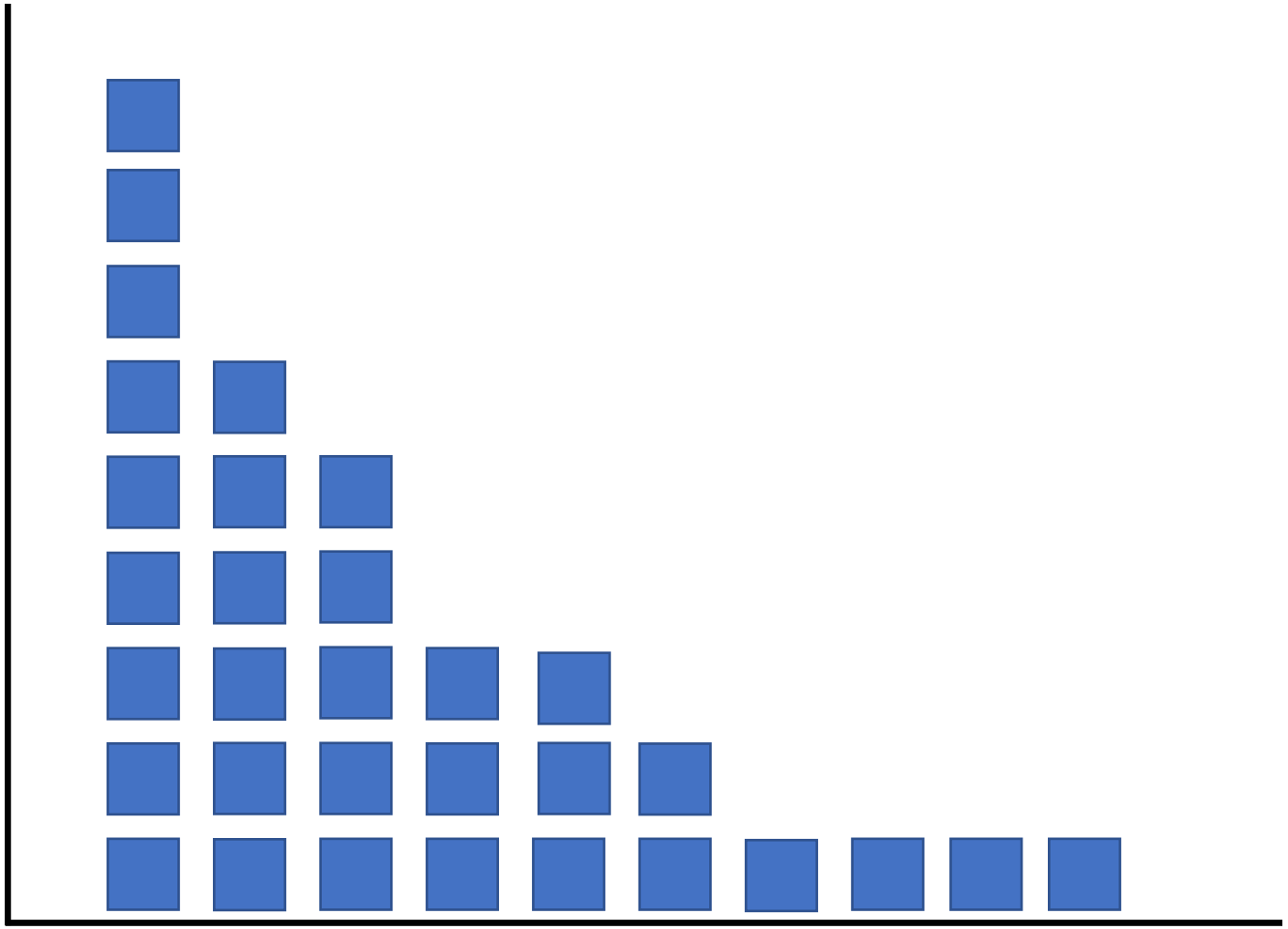
$$F = \frac{SS(\textbf{mean}) - SS(\textbf{fit})/(p_{fit} - p_{mean})}{SS(\textbf{fit})/(n - p_{fit})} \rightarrow F = 3$$



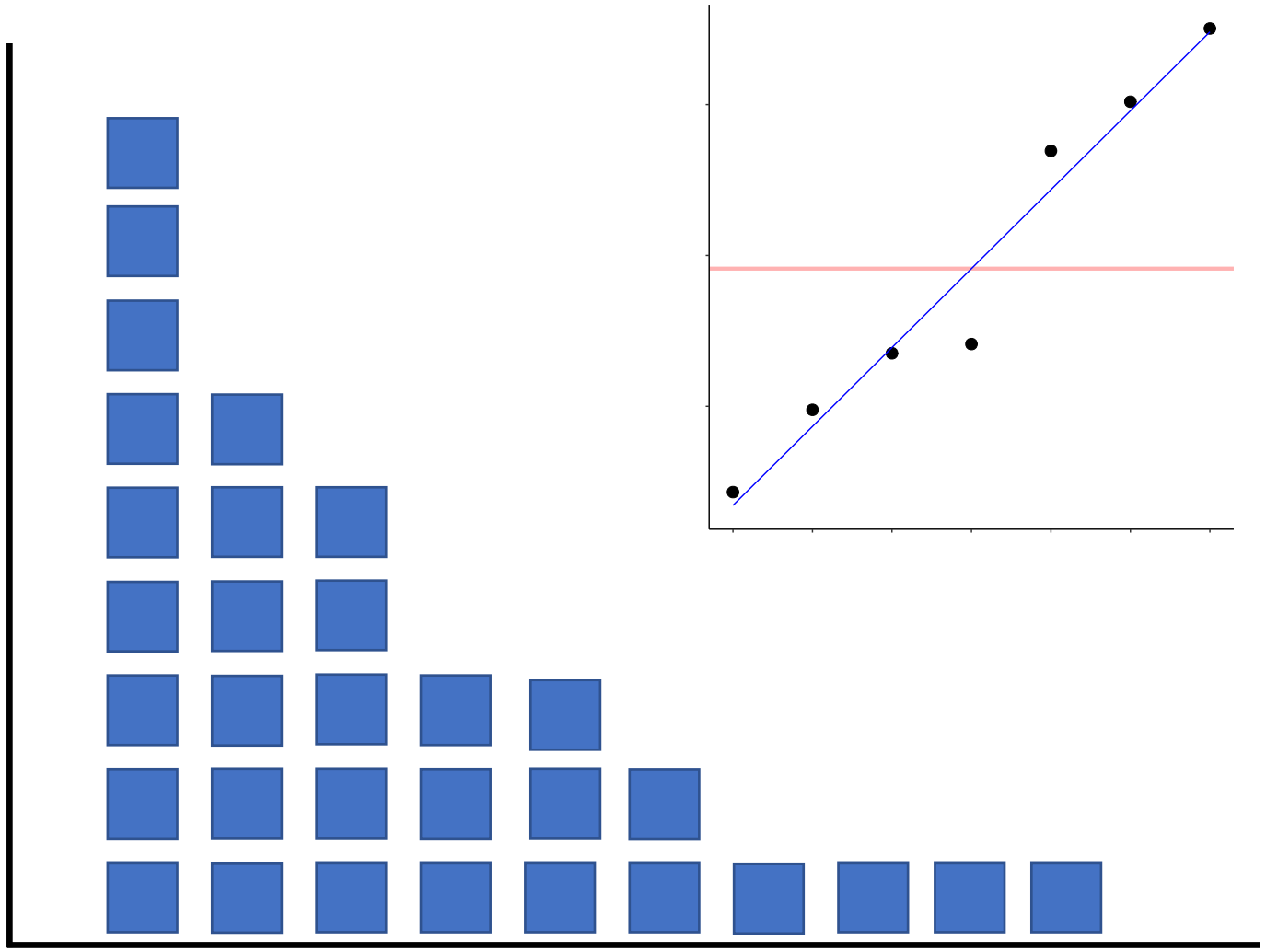


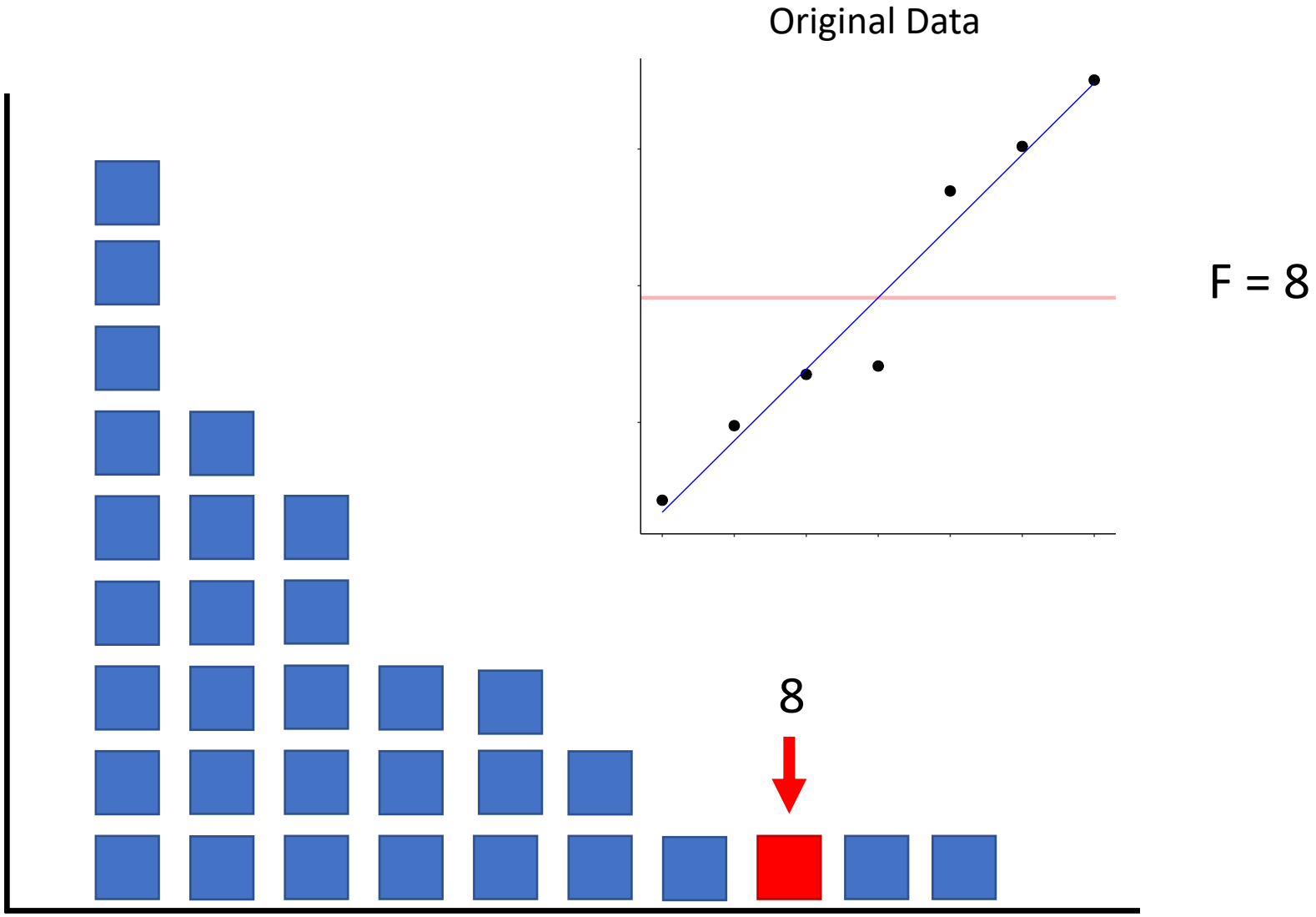




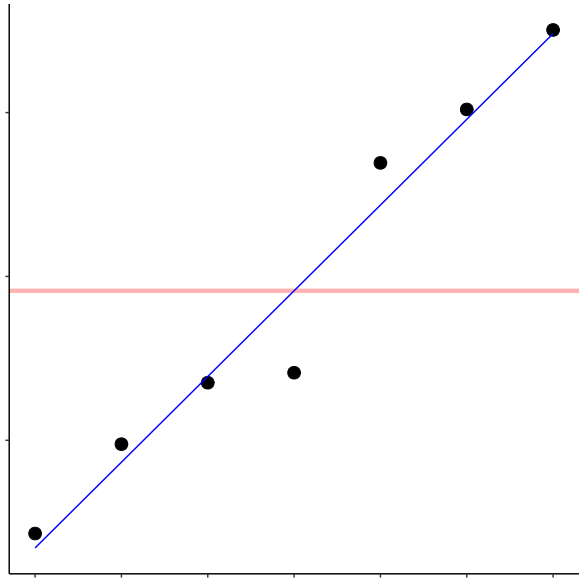


Original Data



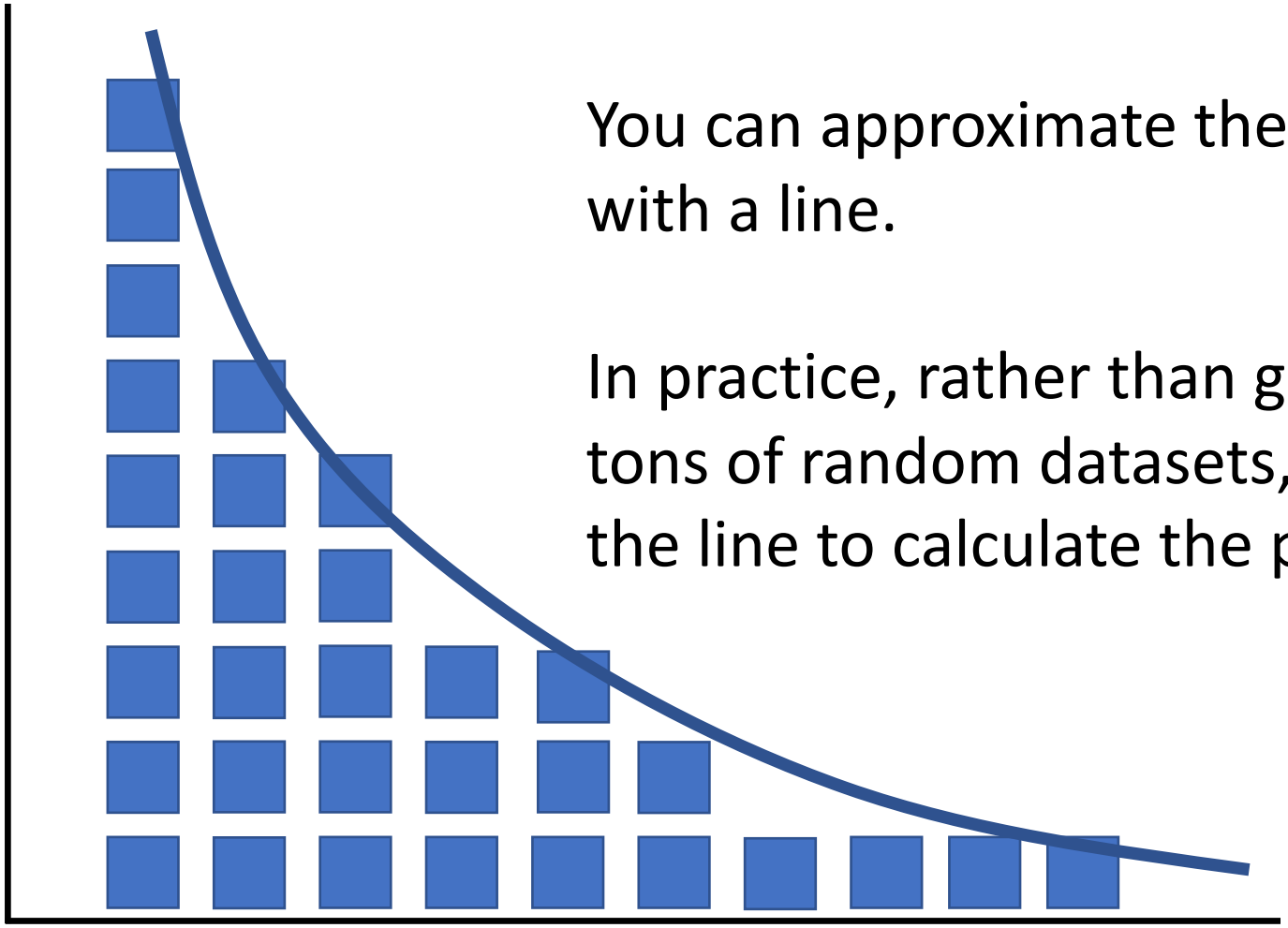

Original Data
8
F = 8

Original Data



F = 8

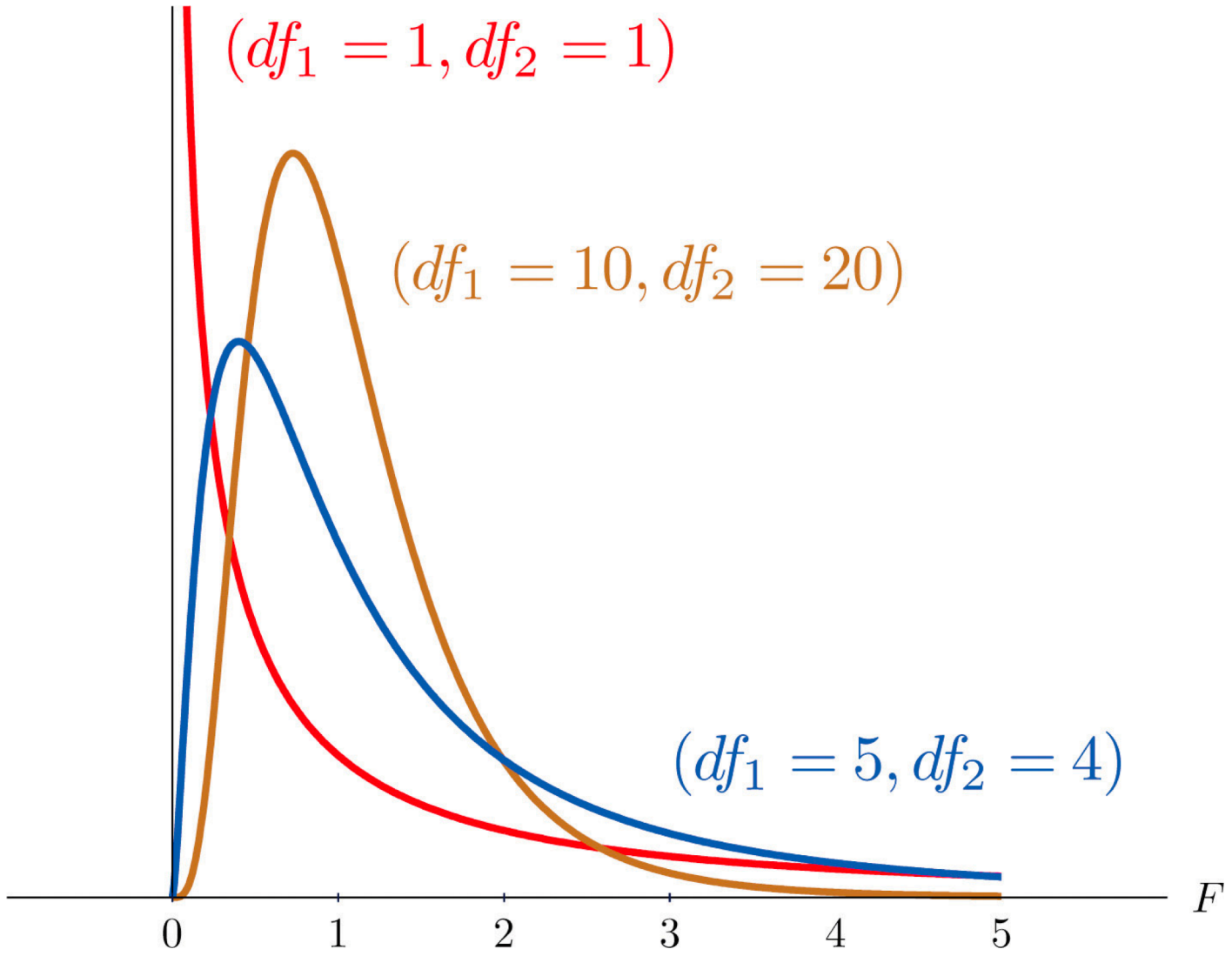
The p-value is the number of equally or more extreme values divided by all the values



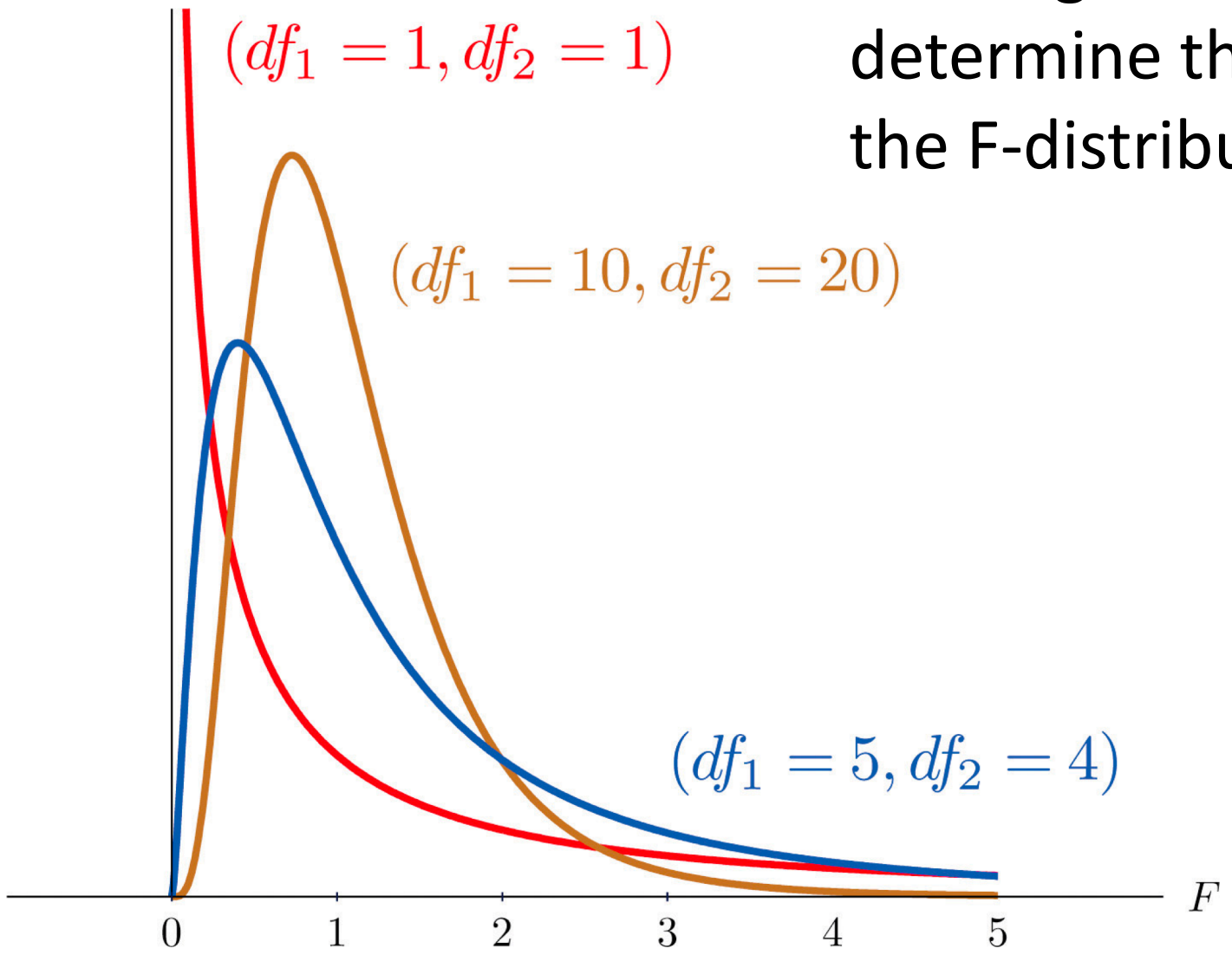
You can approximate the histogram with a line.

In practice, rather than generating tons of random datasets, people use the line to calculate the p-value

Here are some F-distributions

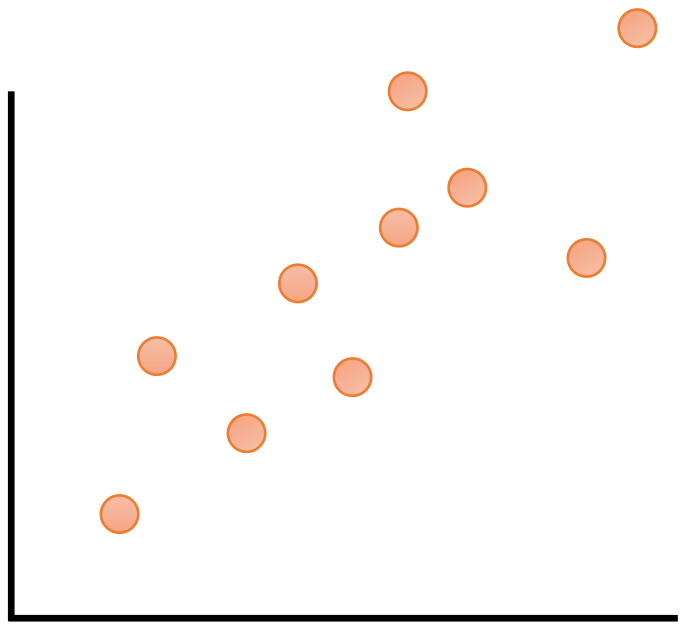


The degrees of freedom determine the shape of the F-distribution



Main ideas

Given some data that you think are related…



Linear regression:

1) Quantifies the relationship in the data (R^2)
2) Determines how reliable that relationship is (this is the p-value that we calculate with F)

You need both to have an interesting result!